# WORKLOAD AND TASK PERFORMANCE IN HUMAN-ROBOT PEER-BASED TEAMS

By

Caroline E. Harriott

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

May, 2015

Nashville, Tennessee

Approved:

Julie A. Adams, Ph.D., Chair

Robert E. Bodenheimer, Ph.D.

Douglas H. Fisher, Ph.D.

John D. Lee, Ph.D.

Matthew Weinger, M.D.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## ACKNOWLEDGMENTS

This work was supported by grants from the Air Force Office of Scientific Research, the National Science Foundation, and an Office of Naval Research MURI award. I was also supported by an Office of Naval Research grant in 2014 that did not directly support this research, but allowed me to continue working on the Time-Structured evaluation and the analysis of the Guided and Collaborative evaluation data.

My advisor, Dr. Julie A. Adams, has inspired me throughout my time at Vanderbilt University. She is a tenacious leader who cares deeply for her students, and I am thankful for the time I have spent in her company. Each of the members of my committee has provided insight necessary for this research. I also thank each committee member for individual feedback regarding the experimental design of the Time-Structured evaluation.

I also thank Electa Baker, Dr. Travis Service, Glenna Buford, and Dr. Ryan Datteri for sharing lunch with me so often over the years. I am grateful to Electa Baker, Matt Manning, Dr. Tao Zhang, Glenna Buford, Dr. Eli R. Hooten, Dr. Sayan Sen, and Dr. Sean T. Hayes for their help preparing for and running the Guided, Collaborative, and Time-Structured evaluations.

Finally, this work was supported by my family; I am forever thankful for their tireless encouragement. My partner, Tim, has been endlessly giving. My parents have inspired me to be persistent, and my siblings and friends were always there to make me laugh. I would not have been able to complete my degree without them.

# Chapter I

## Introduction

Robotic technology is developing rapidly, and humans are beginning to be partnered with robots (Scholtz, 2003; Goodrich and Schultz, 2007; Arkin and Moshkina, 2014). Human-robot interaction is an interdisciplinary field encompassing research questions regarding the extent of the impact of robotic design, presence, task assignments, and behaviors on a human (Goodrich and Schultz, 2007). Aspects of the human's performance and behaviors that can be affected by the robot include: communication styles, trust, safety, companionship, task precision, and task accuracy, among many others. Humans and robots can work through indirect (i.e., the robot executes operator-issued commands) or direct (i.e., working in the same space with bi-directional information flow) interaction (Thrun, 2004). This dissertation focuses on modeling and assessing human task performance and workload, while humans directly interact with robots to achieve tasks.

The goal of this dissertation was building a usable database of workload and task performance metrics that are appropriate for investigating the differences between human-human and human-robot interaction, because these workload and task performance differences influence the design of robotic peers. Specific differences in workload and task performance were assessed via human performance modeling techniques and the analysis of appropriate metrics in three evaluations of peer-based human-robot teams in the first response domain. The first response domain has constraints that impact the measurement of task performance and workload, such as an unpredictable environment, mobility of tasks, specialized protective gear, and time-critical task goals; thus, potential metrics were investigated to address these constraints and included a variety of appropriate and collectable methods (e.g., commercially available physiological sensors, observational data, subjective responses).

Specific metrics are necessary for mobile human-robot peer-based teams because environmental characteristics place limitations on the means of collecting the metrics. The research categorizes general metrics for all aspects of human-robot interaction (e.g., Steinfeld et al., 2006; Murphy and Schreckenghost, 2013). This dissertation categorized and assessed task performance and workload metrics for human-robot peer-based teams. Workload and task performance have been specifically identified in the literature as important aspects of human-robot interaction that can be measured using subjective, physiological, and observational methods (e.g., Murphy and Schreckenghost, 2013; Tiberio et al., 2013; Arkin and Moshkina, 2014; Colin et al., 2014; Novak et al., 2014). This dissertation provides guidance for using workload and task performance metrics in human-robot peer-based teams.

Human-robot teams are of interest due to the development of robotic technology as human partners for

peer-based tasks (Scholtz, 2003). Peer-based interaction involves teammates (e.g., a human and a robot), physically collocated and working on shared goals. Peer-based tasks often incorporate verbal interaction, decision-making between teammates, and adaptation to teammates' behavior (Scholtz, 2003). It is known that individual human performance can impact human team performance (Katzenbach and Smith, 1993). This fact has also been posited as true of humans in human-robot peer-based teams (Goodrich and Schultz, 2007). As human-robot team capabilities improve, it is necessary for the robotic team members to understand how the human's performance capabilities affect the task at hand. Future robots will be able to adapt their behavior, as humans do in human teams, in order to mitigate and accommodate performance changes in their human partners. This robot behavior adaptation will be based upon current and predicted human state. Thus, the focus of this dissertation is the human's performance when working with a robot, rather than on adapting the robot's behavior.

The domain focus is primarily first response, but can cross over to military domains. For example, robots are being developed to work as peers to aid humans in tasks, such as scene surveillance, object tracking, medical triage assessment, victim transport, hazard identification, and resource hauling (Humphrey and Adams, 2009). The human and robot each have strengths and weaknesses that can affect interaction and task performance. For example, the human partner has a limited time in the field and wears protective gear that limits mobility and visibility, while the robot has a limited number of sensors that it can carry and may have limited mobility or vision in certain environmental conditions. Due to the development of robots as peers and their potential to augment human teams, comparisons designed to understand potential differences between human-human and human-robot teams performing the same sets of tasks was the focus of three evaluations for this dissertation.

Many factors influence human performance and appropriate measurement techniques must be chosen to capture these influences in the mobile, dynamic environment common in human-robot interaction. Additional constraints include the human's protective gear, the lack of consistent power sources or communication networks, the need for an uninterrupted primary task, and systems, such as overhead cameras may not be feasible in outdoor environments, unknown areas, or unpredictable lighting conditions. Adapting metrics from other domains, such as human-computer interaction, human factors, and aviation is possible, but selection must consider the special requirements of working with a mobile robot peer in unpredictable environmental conditions.

A significant amount of human-robot interaction research has focused on measuring human performance (e.g., Chen et al., 2007; Murphy, 2004). Research has also focused on modeling the workload and performance of remote robot operators in the first response domain (e.g., Colin et al., 2014), but little research has focused on using human performance modeling techniques to predict interaction between peer-based human-

robot teams. Human performance modeling seeks to capture human behaviors via analytical or computational means (Pew and Mavor, 2007). Typically, the model attempts to replicate the human(s) for a range of tasks and can be used in human-robot interaction (Harriott and Adams, 2013). As human-robot team capabilities improve, it is necessary for the robotic team members to understand how human performance capabilities affect the task at hand. Future robots will use human performance modeling predictions to adjust behavior based on current performance. Additionally, team structure can be planned based on model results. This dissertation analyzed human performance modeling in the context of understanding how well the model results matched, or predicted, the corresponding quantifiable human-robot team results.

Chapter II presents background information regarding workload, task performance, human performance modeling, and human-robot interaction. Chapter III presents the Guided and Collaborative evaluations. Chapter IV presents the entire set of metrics evaluated in the Guided and Collaborative evaluations and discusses their advantages and limitations. Chapter V presents the Time-Structured evaluation, the third evaluation, which combines the assessment of a subset of the metrics in Chapter IV. The Time-Structured evaluation Method is presented in V.1. Results of the Time-Structured evaluation are presented in V.3 and a comparison of the results to the Guided and Collaborative evaluations is provided in V.4. A general discussion is presented in V.5. Finally, the Contributions and Future Work are presented in Chapter VI.

# Chapter II

# Background

This chapter presents relevant background information regarding collaborative human-robot teams, workload, task performance, human performance modeling, and human-robot interaction. Collaboration in human-robot teams is discussed in Chapter II.1 in order to provide background for analysis in Chapter III. Workload and task performance are discussed in more detail in Chapter II.2. Chapter II.2.1 focuses on Workload and its measurement, with a subsection focusing specifically on physical workload (Chapter II.2.1.1). Chapter II.3 discusses human performance modeling concepts, background on modeling workload (Chapter II.3.1) and task performance (i.e., specifically modeling reaction time in Chapter II.3.2), related research in human performance modeling for robotics (Chapter II.3.3), and model verification techniques (Chapter II.3.4).

## II.1 Collaborative Human-Robot Teams

Team collaboration occurs when team members share joint tasks and environmental knowledge (Bruemmer et al., 2005). According to Bratman (1992), collaboration assumes three requirements: 1) mutual responsiveness, 2) commitment to a joint activity, and 3) commitment to the support of the team. Hoffman and Breazeal (2004) indicate that human-robot collaboration requires a shared activity, which incorporates Bratman's assumptions of mutual responsiveness and teammate support. Hoffman and Breazeal specify that humans and robots working collaboratively share joint intention (commitment to the shared goal), common ground (shared knowledge), and goals. A collaborative relationship between teammates requires a peer-based relationship (Scholtz, 2003), where team members work in a shared workspace toward the same goals, while using a shared pool of knowledge. There may be separate tasks for the individual teammates, but no one teammate can complete the task without the other(s).

Collaboration between humans and robots has been investigated in a variety of contexts, including varying levels of autonomy (Brookshire et al., 2004; Stubbs et al., 2007) and modeling robot behavior after human collaborative behaviors (Briggs and Scheutz, 2011; Chernova and Breazeal, 2010; St. Clair and Matarić, 2011). Research has not yet focused on understanding the implications of a collaborative relationship between a human and robot on the human's mental workload.

Brookshire et al. (2004) determined that a task can be successfully completed by a fully autonomous robot or a teleoperator driving a robot, but that a collaborative relationship provided a significant increase in performance and efficiency. However, increases in robot autonomy do not strictly increase collaboration between a human and a robot (Stubbs et al., 2007). Higher autonomy levels prevented humans from fully

understanding the reasoning behind robots' actions and prevented the formation of a common ground, one of the elements of a collaborative relationship.

Similar to Brookshire et al. (2004), Bruemmer et al. (2005) investigated the intersection of robot autonomy and human-robot collaboration. The robot and human had a shared understanding of the task and environment via a collaborative workspace. As robot autonomy and team collaboration increased, team performance increased. These prior results focused on the effect of changing autonomy on the human-robot team's performance, whereas this dissertation also examines the collaborative relationship of human's and robot's effect on mental workload of the human.

Researchers developed collaborative robots for human-robot teams and have improved the quality of a robot's ability to collaborate as people do (Briggs and Scheutz, 2011; Chernova and Breazeal, 2010; St. Clair and Matarić, 2011). These developments have improved human-robot collaboration, but they do not offer insights into the impact of collaboration on the human's performance or internal state.

Hinds et al. (2004) assessed human feelings of personal responsibility, blame, and credit in human-robot collaboration. The authors teamed participants with either a human, a human-like robot, or a machine-like robot in a subordinate, peer, or supervisor role. Results showed that the participants relied on human teammates more than robot teammates. Participants also relied more on robot peers than supervisors or subordinates. The difference in task performance or workload levels between the human-human and human-robot teams, given the different relationship types, were not investigated; however, this research question is addressed in this dissertation.

### II.2 Workload and Task Performance

Workload and task performance were chosen as the focus for this dissertation, because they measure the amount of tasks completed in a given amount of time and the quality of the performance of those tasks. Workload and task performance metrics do not encompass all aspects of human performance, but they do capture how much a person is doing and how well it is accomplished.

Specifically, workload is of interest for direct human-robot interaction, because there are typically multiple concurrent tasks to manage in a given scenario. Traditional workload evaluations assessed pilots, who typically have many active tasks to track and accomplish in a limited amount of time (e.g., Wierwille et al., 1985; Raby and Wickens, 1994). Workload has been of interest in driving tasks as well, because drivers face multi-tasking and distraction while accomplishing their primary task (e.g., Wierwille et al., 1977; Lee et al., 2001). Workload and task performance evaluations in similarly high-pressure healthcare environments has incorporated subjective, physiological, secondary task, timing analysis, and observational measures (e.g., Weinger et al., 1994; Leedal and Smith, 2005; Gawron, 2008). Expectations of and measurement techniques

for workload in peer-based human-robot interaction have not yet become clearly established.

Additionally, task performance is a focus, because it is necessary to evaluate whether human-human and human-robot teams accomplish tasks at similar levels. For example, deployed human-robot teams will not be useful if they accomplish tasks with a significantly worse level of task performance than human-human teams, and it is necessary to determine whether differing levels of task performance are expected.

Workload is discussed in more detail in Chapter II.2.1 with a specific definition of, and measurement techniques for, overall workload, mental workload, and physical workload. Task performance is discussed in Chapter II.2.2, and reaction time and response time are discussed in detail in Chapter II.2.2.1 as specific measures of task performance.

### II.2.1 Workload

Workload can be defined as a measurement that encompasses the amount of work a person performs in relation to the amount of time available to complete it (Wickens et al., 2003). Additionally, workload is considered the differential between the amount of available processing resources and the amount required by a task (Hart and Staveland, 1988). Workload overload occurs when there are too few resources available to allocate to the required tasks, while workload underload occurs when there are too few tasks using available resources. Both of these conditions can hinder overall performance (Nachreiner, 1995).

Workload levels can be affected by other factors specific to the human who is performing the task, such as individual differences including training, amount of skill, or fatigue (Wickens et al., 2003). Additionally, external and environmental conditions can affect workload. Workload has been investigated in the healthcare (Ryu et al., 2003; Medvec, 1994), aviation (Wilson, 1993, 2002) and military domains (Dixon and Wickens, 2003, among others). Workload has also been assessed in the human-robot interaction field (Draper and Blair, 1996; Kaber et al., 2000; Ruff et al., 2002; Steinfeld et al., 2006; Goodrich et al., 2007). These domains rely heavily on human performance in stressful, time-critical situations without allowance for failure. Analysis of workload in any of these domains can offer important information about expected task assignments and system configurations. Assessing workload in human-robot teams is a focus of this dissertation.

Workload accumulates along multiple channels of demand. McCracken and Aldrich (1984) define a processing system involving visual (i.e., external stimuli), cognitive (i.e., information processing), auditory (i.e., external stimuli) and psychomotor (i.e., required physical action) components of every task. Wickens' Multiple resource theory does not define workload, but rather is a four-dimensional model describing multiple-task performance (Wickens, 2002). Tasks interfere more with one another if they are experienced with similar processing stages (i.e., perceptual/cognitive or response execution), perceptual modalities (i.e., auditory or visual), visual channels (focal - detail/pattern recognition or ambient - peripheral/motion), and processing

codes (visual or spatial). These theories determine when tasks interfere with one another.

Workload can be measured using a variety of subjective and objective metrics, and workload is typically measured in terms of overall workload, mental workload, or physical workload. Overall workload refers to an overall measurement of demand experienced during a given amount of time. Humans have limited mental resources; thus, mental workload can be defined as the difference between the amount of available mental processing resources and the amount required by a task (Hart and Staveland, 1988). For instance, an easy, routine task may require only 10 percent of a participant's available resources, whereas a very difficult task may require 90 percent of the participant's mental resources. The relationship between mental and physical workload presents a challenge to human-robot interaction and it is important to understand how to accurately measure physical workload. Physical workload, as defined for this dissertation, refers to the combination of physical demands placed on participants while completing the tasks and incorporates demands from fine motor (e.g., typing), gross motor (e.g., walking), and tactile (e.g., feeling for a pulse) tasks. Physical workload and its measurement techniques are discussed specifically in Chapter II.2.1.1.

Subjective workload metrics include self-report surveys. Hart and Staveland (1988) developed a widely adapted subjective workload measurement tool, the NASA Task Load Index. This tool defines workload as a weighted mean of subjective ratings along six channels of workload demand: mental demands, physical demands, temporal demands, own performance, effort and frustration. Participants first score each of the six channels on a continuous scale from 0 to 100, after which, participants select a channel during paired comparisons, which ask participants to select which of two channels contributed most to the task. The paired comparisons are completed in order to determine the weight for each channel. The six channel scores are weighted and summed to determine a final score, ranging from 0 (no demand) to 100 (high demand). Subjective workload surveys, such as the NASA Task Load Index, have been applied to human-robot interaction evaluations (Draper and Blair, 1996; Kaber et al., 2000; Ruff et al., 2002; Goodrich et al., 2007).

The Multiple Resource Questionnaire collects scores using seventeen channels of mental demand, defined by Wickens' Multiple Resource Theory. Example channels of demand include the auditory emotional processing channel, the facial figural processing channel, and the spatial attentive processing channel. Participants rate each channel on a Likert scale from 0 to 100 in increments of 25 points, where each score corresponds to a usage level of the demand (e.g., a score of 25 implies light usage, a score of 100 corresponds to extreme usage). It has been suggested to reduce the seventeen channels to only the top three that represent the demands of the task (Boles and Adair, 2001). This method ensures less "noise" from channels with little or no demand.

The NASA Task Load Index and the Multiple Resource Questionnaire have been compared in terms of sensitivity to workload changes (Finomore et al., 2006, 2009; Klein et al., 2009) . Each survey offers benefits

and limitations. The NASA Task Load Index provides a sensitive measure of workload and provides information as to the most important (i.e., more highly weighted) channels for the task. The Multiple Resource Questionnaire offers the ability to customize the data collection to the most relevant demand questions for the task.

Subjective workload measurements can provide insight into how participants perceive demand, but typically data collection occurs after completing a task or subtask, which can limit the granularity of knowledge about how workload changes over time. Additionally, many subjective measures of workload are not absolute measures and cannot be directly compared between studies. The individually rated channels of demand can provide insight specifically into perceived mental workload or physical workload. Additionally, measuring workload using self-report surveys in human-robot peer-based teams in domains with active, dynamic tasks (e.g., first response) is not ideal because of the need to pause the primary task in order to collect data. Task interruptions may not always be possible; thus, other methods of collecting workload must be explored.

Objective metrics include observational computations, such as task density calculations (Weinger et al., 1994). Task density assesses how many tasks are completed by a participant in a measured amount of time. The tasks must be known and identifiable, and the amount of time must be measurable. Task density can be compared between task configurations or groups for an assessment of overall workload. Another calculation of objective overall workload is the workload density calculation, which is a time-weighted average of workload factor scores determined from survey values (Weinger et al., 2004). Other objective measures of overall workload include response latency time to a visual cue, which is similar to the response time metric presented in Chapter II.2.2 (Weinger et al., 2004).

Objective metrics also include measures of spare mental capacity and physiological responses. Spare mental capacity can be measured through secondary task questions (Gawron, 2008). Secondary tasks are separate from the primary task and associated metrics, such as correctness and speed of response, which can indicate levels of participant performance and mental workload. Secondary tasks can include activities, such as memorization, simple math, counting, or answering questions.

Physiological metrics, such as heart rate, heart rate variability, and respiration rate have been correlated to mental workload. Heart rate has been demonstrated to increase as mental workload increases (Castor, 2003; Weinger et al., 2004), while respiration rate decreases as mental workload increases (Roscoe, 1992; Keller et al., 2001). Heart rate variability has also been demonstrated to decrease as workload increases (Aasman et al., 1987; Castor, 2003; Roscoe, 1992); however, when the task has requirements beyond the limits of working memory, participants experience an increase in heart rate variability (Aasman et al., 1988). Performance on tasks typically decreases when human information processing resources are limited, and the processes compete for the use of resources (Norman and Bobrow, 1975). There are two general concepts

that categorize the limitation of cognitive processes, resource-limited and data-limited processes. Resource-limited processes are categorized when an increase in available processing resources increases performance, while data-limited processes are those which are independent of processing resources (Norman and Bobrow, 1975). Heart rate variability is sensitive to resource-limited processes, rather than showing sensitivity to data-limited processes (Aasman et al., 1987, 1988).

Other physiological measures of workload include electroencephalography (i.e., measuring changes in voltage within the brain), galvanic skin response (i.e., the electrical conductance of the skin), blink rate of the eyes, pupil diameter, and blood pressure (Kramer, 1990; Mehler et al., 2009; Veltman and Gaillard, 1998; Wilson et al., 1987). For example, an increase of workload results in a longer duration between blinks and an increase in galvanic skin response (Veltman and Gaillard, 1998).

The use of physiological measures of workload in human-robot interaction has not yet been widely adapted, and there are few examples of its use (e.g., Novak et al., 2014). Novak et al. (2014) evaluated electroencephalography signals, respiration rate, skin conductance, skin temperature, and eye tracking information in relation to workload. The findings demonstrated that respiration rate was related to changes in overall workload, even with a task that varied physical workload. Electroencepholography signals from the central and frontal areas related to workload. Physiological measures have also been used to represent some affective states (e.g., Rani et al. (2002); Liu et al. (2006); Tiberio et al. (2013)), but not workload specifically. Subjective workload measures are much more commonly utilized.

Physiological measures include advantages and disadvantages. Most physiological measures require sensors on the body, which can be obtrusive and difficult to place. Some of these sensors can also be difficult to use outside of a stationary laboratory setting. For example, electroencephalography signals are measured via placing approximately 20 to 256 individual leads on the participant's scalp and face (Abdulghani et al., 2009; Brodbeck et al., 2011). Heart rate and heart rate variability have also been measured using a web camera recording a participant's face during a task (Poh et al., 2011b,a), but mounting a web camera in a position to record the necessary view of a participant's face during a mobile task may not be possible. Physiological sensors are subject to noise. Discriminating between signal and noise often requires the use of filters for noise reduction, but ultimately noise cannot be eliminated. Additionally, physiological changes are influenced by many factors that include, but are not limited to, mental workload. Due to these factors, isolating mental workload using physiological signals is not a perfect process (Kramer, 1990); therefore, a combination of subjective and objective measures is needed to assess mental workload.

Steinfeld et al. (2006) proposed that an objective measure of workload for human-robot interaction is the number of interventions (i.e., unplanned robot interactions) per unit of time. Alternatively, workload can be represented by the ratio of operator engaged task time to robot execution time, or the fan out (Goodrich

and Olsen, 2003) number representing the number of robots that can be effectively controlled by the human operator. These workload measures were designed for teleoperation and supervisory interaction.

#### II.2.1.1 Physical Workload

Physical workload is a specific subcomponent of overall workload and can be measured using objective metrics, such as physiological responses and subjective metrics. Heart rate is a widely used physiological metric of energetic load and physical workload, with higher heart rate corresponding to higher physical workload. Reiser and Schlenk (2009) detailed methods of measuring physical activity in order to determine whether medical patients maintained an appropriate level of movement. The presented measurement methods included self-reports (via questionnaires and daily diaries) and objective metrics, specifically heart rate, number of steps (via pedometer), accelerometer data and direct observation. The authors suggested considering recall bias when analyzing self-report data and accounting for an individual's physical fitness when analyzing heart rate, as the hearts of more fit individuals work more efficiently. The authors also specified that pedometers only quantify lower body movement and not upper body motion. Additionally, accelerometer data offers a good correlation to physical movement.

The long-term implications of changes in physical workload include fatigue and maximum possible work time. Wu and Wang (2002) determined that the maximum acceptable work time decreases as physical workload increases. The authors suggest physical workload limits in terms of a percentage of aerobic capacity for 4, 8, 10 or 12 hours of work. For example, during a four-hour shift, a human can be expected to maintain 43.5% of their maximum aerobic capacity, but that percentage drops to 28.5% for a twelve-hour shift. It is important to monitor physical workload and determine maximum work times based on the conditions and duration of deployment.

Disaster response and surgical procedures represent two types of activities in which human-robot teams can function to meet a goal. Both examples involve physical demands that can impact the human and team performance. Human-robot teams deployed to disaster situations may conduct activities over lengthy shifts that occur across multiple days or weeks. For example, after the September 11, 2001 terrorist attacks in New York City, human-robot teams worked to identify victims and inspect areas beneath rubble piles (Casper and Murphy, 2003). The teams were present at the disaster site for twenty-two consecutive days; robots were used during eight deployments in that time period. Physical tasks in human-robot teams included carrying the robot for multiple city blocks, moving robot controls, and traversing rubble with robots. Accurately predicting physical workload in unique human-robot teaming situations can help in analyzing performance capabilities of team members and ideal task assignments over time in order to prevent human team member injuries.

Physical workload can also be alleviated through human-robot interaction. The accurate prediction and measurement of physical workload can be used to optimize the team performance. For example, surgical procedures are often labor intensive, stressful, and can last multiple hours. The use of surgical robotic technology has been shown to significantly improve surgeon posture and relieve discomfort during the procedure (Van der Schatte Olivier et al., 2009).

Garet et al. (2004) measured physical workload using heart rate-estimated energy expenditure (HREEE), which represents a percentage of the cardiac reserve. Energy expenditure is calibrated to heart rate for each individual via whole-body indirect calorimetry. HREEE is calculated using rest and maximum heart rates:

$$HREEE = \frac{HR \times HR_{rest}}{HR_{max} \times HR_{rest}} \times 100, \tag{II.1}$$

where HR represents heart rate, $HR_{rest}$ represents the resting heart rate, and $HR_{max}$ is the maximum heart rate. The corrected HREEE (CHREEE) adds the correction value 15 to the resting heart rate:

$$CHREEE = \frac{HR \times HR_{rest} + 15}{HR_{max} \times HR_{rest} + 15} \times 100. \tag{II.2}$$

The correction value adjusts for blood movement in the participant's awakened state, a technique documented to correlate heart rate to energy expenditure (Garet et al., 2004). The results indicate that corrected HREEE was an accurate method of estimating actual energy expenditure.

Body movement measurements also correspond to physical workload. Hansson et al. (2009, 2010) measured wrist, head, and upper arm positions and movements as well as the muscular activity of the trapezius muscles and forearm extensors to assess physical workload of forty-three different types of work. The electromyography signals increased when muscle load and exhaustion increased.

Li and Buckle (1999) detailed posture observation techniques that indicate physical exertion and exposure to muscoskeletal risk. The metrics included the Ovako Working Posture Analyzing System, which decomposes body segment movements as one of four types: bending, rotation, elevation and position (Karhu et al., 1977). The Rapid Upper Limb Assessment technique was useful for measuring arm postures in mostly sedentary jobs (McAtamney and Corlett, 1993) and the Hand-Arm-Movement Analysis system was employed to identify risk factors for injuries from repetitive movements (Christmansson, 1994). Relying on observational data raised the question of inter-observer reliability and explored how to address intermittent data points gathered from a limited number of observations.

Vector magnitude has been shown to represent gross physical activity, a component of physical workload (Steele et al., 2000). Forty-seven chronic obstructive pulmonary disease patients donned a triaxial movement sensor to measure walking and daily movement activity. Self-reports of movement activity proved to be in-

accurate. Vector magnitude was chosen to measure physical activity, rather than calorie expenditure, because of the patients' ventilatory dysfunction. Walking distance correlated significantly to vector magnitude data and it was concluded that vector magnitude is an appropriate measure of physical activity.

Measuring and assessing body postures throughout a task can indicate physical exertion levels. Cutlip et al. (2000) investigated a variety of postures involved in the dangerous job of scaffold disassembly in order to determine whether a specific set of postures provoked better (or worse) levels of physical exertion. Using a strength-testing apparatus and seven pre-designed postures, individual exertion was evaluated against established acceptable exertion levels. The analysis tested the standard deviation, skewness, and kurtosis of the participants' exerted strength and the required coefficients of friction for a lift in each posture. After evaluating each posture, two were chosen, symmetric front lifting at elbow height and symmetric front lifting at knuckle heights, that caused the least risk and most benefit. The skewness, kurtosis, and standard deviations were found to be in an acceptable range. The coefficients of friction were not found to differ significantly between postures.

Changes to body postures indicate physical movement, and Lasley et al. (1991) measured posture sway to quantify the amount of disorientation in a subject due to visual stimuli. Participants were shown visual stimuli designed to disorient. Participants stood on a suspended steel platform with strain gauges to measure shifts in a participant's posture. Posture sway was calculated by determining the mean squared deviation from the mean postural position, or the variance of the posture moving towards and away from a subject.

Many physical workload metrics combine both physiological and subjective ratings. Paul et al. (1999) employed subjective ratings of perceived load and physiological measures of energetic and postural load. Perceived load is the mean rate of perceived fatigue and energy exertion. Energetic load is calculated using a percentage of the heart reserve, using heart rate. Postural load is defined as the time during which a participant's trunk is flexed more than $45^\circ$ and elevation of the upper arms is more than $60^\circ$. The trunk and arm positions were recorded using the Task Recording and Analysis on Computer system, where an observer noted current positions every 15 seconds. This evaluation found that physical workload can be represented by a combination of subjective ratings, physiological measures, and observations.

### II.2.2 Task Performance

This dissertation is not only concerned with the workload that a participant experiences while working in a human-robot team, but also with the quality of the performance the participant achieves, measured via task performance metrics. Task performance represents how well a participant accomplishes a task and is measured objectively by gauging error, efficiency (e.g., time), and/or accuracy (Gawron, 2008). Examples of task performance measures include how many times a participant accomplishes a task or subtask, how long

it takes, whether or not the task was successful, or how close the method of completion was to the correct method.

Task performance has been extensively examined in relation to mental workload (Gawron, 2008). For example, Wierwille et al. (1985) assessed primary task failure rate, termed meditational error rate, in a simulated flight task. Meditational error rate was defined as the number of incorrectly answered and unanswered task prompts divided by total number of prompts. This measure of task performance was shown to represent changes in mental workload more effectively than respiration rate, pupil diameter, and heart rate. Mental workload and task performance can be related in many circumstances.

Measures of overall workload and task performance can be impacted by the task itself and by the environment. The dissociation theory of workload and performance (Yeh and Wickens, 1988) describes five conditions in which the relationship between subjective workload measures and task performance imply different effects on workload: motivation, underload, resource-limited tasks, dual-task configurations with different competition for common resources, and overload. There is dissociation between the two measures, because while the amount of invested mental resources is the same for a given task in these five conditions, other factors influence the participant and create differing outcomes in subjective workload ratings or task performance. These factors include aspects, such as the type of task, the working memory demand, motivation for completing the task, difficulty of the task, or a demand on central processing resources.

Task performance has been assessed in human-robot interaction evaluations to measure both the human's performance and the robot's performance (Steinfeld et al., 2006). A specific metric, such as intervention response time that measures the length of time a human operator takes to respond to a robot's problem, can encompass task performance in an human-robot interaction scenario (Sheridan, 1992). Task performance metrics can be used to reflect the nature of the relationship between the human and robot.

#### II.2.2.1 Reaction Time and Response Time

Reaction time (RT) and response time are considered task performance metrics in this dissertation, because these measurements represent the efficiency of information perception and processing. RT is defined as the time between the onset of a stimulus and a human's response to the stimulus (Sternberg, 1969; Wickens et al., 2003). Several distinct cognitive and physical processes comprise RT and it is possible to dissect reaction time into individual components of processing (Donders, 1868). This method of dissecting the individual RT components (e.g., stimulus perception, stimulus recognition, memory retrieval, response organization) is referred to as the subtraction method. Donders performed the first evaluation scientifically measuring RT and proposed the subtraction method by comparing a simple RT to both recognition RT and choice RT. Simple RT was shorter than recognition RT and both times are shorter than Choice RT. Simple RT refers to the time to

make a response to a stimulus without requiring stimulus recognition or choosing a response. Recognition RT requires stimulus recognition (e.g., pressing a button for one stimulus type but not for another). Choice RT involves different responses based on types of stimulus (Luce, 1986). Response time is distinct from reaction time and represents the time between a stimulus onset and the moment of an appropriate response. Response time encompasses both the reaction time to a stimulus and the decision-making involved with choosing a response (Gawron, 2008).

The choice component of RT increases as the number of possible responses increases; therefore, reaction time increases given more choices. The Hick-Hyman law represents the relationship between time and choice of response (Hick, 1952; Hyman, 1953). Both authors determined that the presence of additional choices increased decision time predictably.

Many experiments determined the average length of RT components (e.g., Hohle, 1967; Jastrow, 1890; Donders, 1868). Specifically, RT has been utilized as an objective metric to understand the visual perception system and cognitive processing (Nagler and Nagler, 1973; Macflynn et al., 1984). These experiments constructed carefully designed tasks with controlled stimuli and limited response types. RT is the sum of the time it takes to perform all relevant sub-components of the reaction task.

RT has been shown to be affected by factors, such as age (i.e., increased age results in longer and more variable RT, (Welford, 1977; Jevas and Yan, 2001; Deary and Der, 2005)), gender (i.e., males typically react faster than females; (Noble et al., 1964)), and fatigue (i.e., fatigue lengthens RT; (Welford, 1980)). More recently, RT has been investigated in correlation to brain injury (i.e., white matter damage correlates to longer RT; (Niogi et al., 2008)) and attention deficit hyperactive disorder (i.e., children with this condition have faster normal RT, but a higher proportion of abnormally long attentional lapses when responding than the unaffected population; (Hervey et al., 2006)). Additionally, a large portion of RT research focuses on the brain functions involved in RT (e.g., (Eagle et al., 2008; Lo and Wang, 2006)).

Reaction or response time metrics on-board a robot can facilitate real-time task performance assessments reflective of human perception (Sternberg, 1969), multitasking and interruptions (Trafton and Monk, 2007). Human performance modeling can compare real-time measurements to model values and provide perspective on expected task performance (Silverman et al., 2006).

RT measurement reflects signal processing speed, training, and innate ability (Sternberg, 1969). This section presents traditional reaction time measurement techniques typically used within a laboratory environment (e.g., Nagler and Nagler, 1973; Sjorgren and Banning, 1989), potential methods of measuring RT outside of the lab (Johansson and Rumar, 1971), and examples of physiological measures corresponding to reaction time (Dustman and Beck, 1965; Gramann et al., 2011).

Traditional RT evaluations present a visual stimulus and measure the time for the participant to respond

to the stimulus (Sternberg, 1969; Nagler and Nagler, 1973; Macflynn et al., 1984; Sjorgren and Banning, 1989). These laboratory-based experiments seek to determine the exact time required to perceive and respond to stimuli. Participants typically look at a screen presenting visual stimuli and press a button when the appropriate stimulus appears. Other examples include auditory stimuli with a button press response procedure (Sjorgren and Banning, 1989). Participants did not move their heads or bodies and the button-press responses were typically achieved by only moving one finger. The time between the controlled onset of the stimulus and the button press was measured. This method of providing a clear stimulus onset, or method of clearly determining the response time, is difficult to replicate in real-world situations (e.g., not in the laboratory).

RT has been measured in situations involving participants and motor vehicles (Johansson and Rumar, 1971; Dewar et al., 1976; Maltz and Shinar, 1999; Liang et al., 2007; Lee, 2008). RT for pressing a vehicle's brakes is a sufficient real-world measurement of RT with dynamic and diverse visual and auditory stimuli, but physical movement is limited to pressing a button with the foot (Johansson and Rumar, 1971). For example, a driving situation can define RT as the time difference between the controlled onset of a stimulus and the press (or release) of a foot pedal, (Liang et al., 2007). Peer-based human-robot teams will be in mobile environments and likely will not have primary tasks that can consistently be measured via timing of button presses or an interaction with a stationary object. It is necessary to determine methods of measuring reaction time and response time that can adapt to human-robot interaction. These measurements must include the interaction between teammates, rather than exclusive interactions with stationary objects (e.g., brake pedals).

Eye-tracking devices have also been used to assess RT in motor vehicle driving applications, while drivers searched a visual field (Maltz and Shinar, 1999), and experienced fatigue (Ji et al., 2004). Some eye-tracking devices record where participants look on a specific screen with signal receptors, which may not be relevant to applications in real-world scenarios (Topal et al., 2008). Additionally, webcam applications can be used to perform successful eye-tracking (YouEye, 2013; San Agustin et al., 2010). However, a webcam will need to be mounted at a fixed distance from the head, which can obscure vision. This type of device is usable, but may not be ideal for human-robot teams in which the human may become fatigued, or is using personal protective equipment. A portable eye-tracking is likely necessary for a mobile environment in a peer-based human-robot team. Advanced portable glasses eye-tracking systems exist, such as the Tobii Glasses 2 (Tobii Technology, 2013). Wearable eye-tracking systems are emerging as a way to access potential reaction time information in mobile environments.

Head-mounted cameras can also provide eye movement and RT information. For example, both the gaze of cricket batsmen and the oncoming ball were recorded using a head-mounted camera (Land and McLeod, 2000). This study recorded when the batsmen fixated on the ball and responded with a swing and demonstrated that players with better eye movement strategies ended up being better players. The cameras

used recorded both the movements of the left eye and the view of the player. The film was analyzed to determine when critical eye movements and reactions occurred.

A head-mounted camera, like the EyeSeeCam (Schneider et al., 2009) offers a portable way to record the participant's view. The system offers image stabilization and mobility along with a wide angled view. The EyeSeeCam does not use a view of the actual eyeball, but rather represents the eye's view through advanced motion technology using a video-oculography device. This method seems promising for use in RT studies, but the use of expensive and complicated equipment may be limiting.

Cameras mounted throughout an environment have been used to measure RT in athletics (Neto et al., 2009). Two high-speed cameras were used to measure the RTs of trained Kung Fu fighters to strikes during combat. It is difficult, but not impossible, to expand this method to non-laboratory settings, as ensuring camera coverage will require many cameras and knowledge of where a participant may travel during a task. An outdoor task will require tripods for holding cameras in place, which can be obtrusive or block the natural path of a human. Cameras can offer a good method of measuring RT in the real-world if they are placed appropriately; however, a very large number of real world tasks (e.g., first response) will not permit outfitting the environment with cameras. As a non-laboratory environment can contain many dynamic elements, a head mounted camera may be a better option than stationary environment cameras.

The literature demonstrates that some physiological metrics correlate to RT. RT has been significantly correlated to negative phases in alpha brain waves, as measured by electroencephalography (EEG) measurements (Dustman and Beck, 1965). These EEG measurements indicated that reaction time was faster during negative wave cycles. The brain waves on their own do not measure RT, but correlating RT to aspects of physiology measurable during a task can be used to formulate an accurate model of expected RT. Portable EEG measurement systems currently exist (e.g., De Vos et al. (2014)), and may offer methods of recording EEG data in mobile environments.

Heart rate and heart rate variability have been measured during RT evaluations (Porges, 1972). Heart rate and heart rate variability showed an increase during the time period where participants responded to stimuli. Portable monitors are applicable to real-world situations; however, the cited evaluation included visual warnings of the onset of stimuli and response times, which requires validation for use in real-world situations that do not provide warnings.

RT and response time as specific aspects of task performance in human-robot interaction has been investigated in limited areas. Human RT was comprised of perception, decision-making and motor response time when interacting with robot manipulators (Helander et al., 1987). The robot was controlled via teleoperation, and results showed that faster robotic arm movement did not improve overall performance, as human RT prevented rapid enough responses to stop arm movement before overrunning the desired goals. While

this evaluation did not involve peer-based interaction, the modeling and validation of RT in human-robot interaction is an important step towards understanding and measuring real-world RT.

The established increase in RT post-task-switch has been proposed as applicable to RT in multiple robot teams (Squire et al., 2006). The time associated with task switches varied by interface used to interact with the multiple robots, with the lowest task switching time associated with manual control interfaces.

Rather than using the term "reaction time," some human-robot interaction research has analyzed "response time." No clear definition of response time has been found in literature. However, response time is defined in this dissertation as the time taken to perform a requested task (e.g., answer a question, move a pile of books), while reaction time refers to the measurement of time taken to process and act on a stimulus that is not a direct request at the moment it occurs (e.g., take a photo when a specific stimulus appears).

An analysis with physically and virtually present robots showed that response times are lower when the robot is physically present (Bainbridge et al., 2008). Peer relationships typically require physically present partners and understanding task response time for closely coupled relationships motivates the first step toward developing peer-based team models of RT.

RTs and response times have also been measured in unmanned aerial and ground vehicle control (Donmez et al., 2009; Chadwick, 2006; Prewett et al., 2009). Results showed that RT degraded (Donmez et al., 2009) and response times were longer as more unmanned vehicles were taken under control (Chadwick, 2006). Peer-based teams may also be affected by the impact of the number of robots present in a peer-based team. RT has also been shown to increase as the number of targets to photograph with an unmanned aerial vehicle increases (Consenzo et al., 2006). Likewise, response times increase with dense search environments (Folds and Gerth, 1994). Understanding these trends for peer-based human-robot interaction includes the investigation of RT based on stimulus density, but must be validated in peer-based scenarios.

This dissertation utilized the reaction time and response time measurement techniques discussed in this chapter in order to create generalized metrics of and definitions for reaction time and response time. These metrics will be usable in non-laboratory evaluations and evaluated in a peer-based human-robot teaming evaluation.

### II.3 Human Performance Modeling

Human performance modeling seeks to capture human behaviors via analytical or computational means (Pew and Mavor, 2007). Typically, the model attempts to replicate the human(s) for a range of tasks. Models can be employed to predict the performance attainable between a human and the system or models can used to describe the resulting performance by adjusting model parameters to match existing data (Baron et al., 1990). The purpose of modeling is to provide results that can be used by decision makers to answer specific

questions, but such models must represent both the human and the system. Performance models can be used to understand the implications of decisions on human performance and develop theories about performance. Models can also be used to inform system design and evaluate systems. Fundamentally, the purpose of modeling human performance is to better understand human behavior and performance within a particular system.

Often system prototypes are expensive to develop or simply do not exist. Human performance modeling provides a means of exploring the system implications on human performance prior to the existence of system prototypes. Modeling efforts also permit the exploration of new and alternative human-system interactions for less investment, since real systems or dynamic prototypes are not required (Baron et al., 1990). More specifically, modeling a system provides the capacity to understand the system in ways that observation cannot support (Baines and Benedetti, 2007; Foyle and Hooey, 2008). For example, the modeling effort often allows manipulation of the modeled system in ways that the actual system cannot support.

Another important application of human performance modeling is the development of theories related to human performance that can be further investigated via human-in-the-loop evaluations (Baron et al., 1990; Harriott et al., 2011a,b). Models can be developed to represent the typical conditions under which humans interact with or use systems in order to complete tasks that potentially overlap in time and space (Baron et al., 1990).

Modeling can be used to understand the trade-offs between system changes and human performance changes in extreme conditions. For example, when multiple system designs exist for handling extreme conditions, the analysis of the alternatives can provide insight into the costs necessary to remedy the situation. While this approach may not be preferred for analyzing the alternatives, it is cost effective in times when it is unrealistic to develop actual system prototypes and conduct human-in-the-loop system evaluations. This approach also provides a safer alternative to exposing humans to extreme conditions (Dahn and Belyavin, 1997).

Technology and associated capabilities change rapidly, thus it is not always possible to know exactly how humans will use and interact with future systems. The advent of technological advances can limit the application of knowledge regarding human performance with existing systems to future systems (Booher, 2003). Modeling allows for a generic representation of generic technology and human interaction with the technology, as well as an ability to explore the potential future system capabilities and usage. Modeling activities during the early system planning and concept design stages often lead to the development of measures of effectiveness and performance that can be refined during the design process and applied throughout system development and testing (Booher, 2003). The early modeling activities also provide expected performance with the new system that form a basis for human-in-the-loop evaluations during later design iterations. Per-

formance modeling allows for developing an understanding of the operation of future systems, which can provide information that impacts system design, facilitates cost-benefit analysis, and conceptual interpretations (Dahn and Belyavin, 1997). The modeling of future systems can provide improved understanding of the design alternatives that can lead to reducing the set of design alternatives and tractable human-in-the-loop evaluations of the alternatives.

### II.3.1 Modeling Workload

Workload can be predicted using human performance modeling techniques. Modeling workload allows for the analysis of task allocation and system configuration comparisons. Tools such as the Distributed Operator Model Architecture (D-OMAR) (Deutsch et al., 1999), the Man-Machine Integration Design and Analysis System (MIDAS) (Tyler et al., 1998), and the Integrated Performance Modeling Environment (IPME) (Dahn and Belyavin, 1997) have capabilities to model humans' workload in complex systems. This dissertation focuses on creating models of human workload in IMPRINT Pro (Archer et al., 2005; United States Army Research Laboratory, 2009). Other examples of modeling workload in human-robot interaction are discussed in Chapter II.3.3.

IMPRINT Pro provides the capabilities to set up complex task networks, model workload and incorporate other human performance moderators (e.g., heat, cold, protective gear, sleepless hours, noise, whole body vibration, military rank, and training). Any human performance moderator can be added to the model via the User Stressors module, but the workload models are already integrated into the system (United States Army Research Laboratory, 2009).

Predicting workload levels requires inputting the task timing. IMPRINT Pro provides micromodels of human behavior to help determine task timings using established human factors data sets. For example, if a model contains a task for a human to walk 10 feet, the micromodels calculate the average time a human takes to walk that distance.

The models also require assignment of values of demand. IMPRINT Pro provides guidelines for assigning tasks' demand values, which combines values on seven workload channels: Auditory, Visual, Cognitive, Fine Motor, Gross Motor, Tactile and Speech workload. The values on each channel were assigned based upon channel guidelines. Using the previous example of walking 10 feet, the Gross Motor workload value is based on walking on even terrain and there may be a visual component for looking where one is going or an auditory component for listening for directions, depending on the modeled situation. The composition of each task is determined by the modeler. The probability of success is the input. When the model executes, the task executes successfully based on the expected task accuracy. If the task fails, the modeler specifies what happens (i.e., a different task executes, the model ends or nothing happens). Workload for a given set

of tasks can be computed via a time-weighted average of task demand values.

### II.3.2 Modeling Reaction Time

Reaction Time (RT) was modeled in IMPRINT Pro and serves as an example of a modeled aspect of task performance. Other human performance modeling tools are also capable of representing the components of RT. Modeling systems that are able to incorporate the short tasks composing RT can be utilized in modeling RT. For example, the Goals Operators Methods and Selection Rules (GOMS) modeling system can account for each step when interacting with a computer, including RT (Card et al., 1980). Cognitive architectures such as the Atomic Components of Thought - Rational (ACT-R) (Anderson and Lebiere, 1998) offer the ability to pinpoint each mental task and the associated time taken to complete the task.

IMPRINT Pro incorporates modeling RT by utilizing the theory of the subtraction method (Donders, 1868). IMPRINT Pro's micromodels of human behavior can be used to create a complete RT model with the necessary sub-components. The provided micromodels include, for example, recognition time, decision time, and pattern-matching time (United States Army Research Laboratory, 2009).

RT measurements are typically made in controlled environments with responses required from a finite set of stimuli. However, the modeled reaction time results provide an estimation of a RT from a task with many possible stimuli unknown to participants. The modeled RT estimates how long it will take participants to visually recognize and respond to an item in their field of view (more details will be presented in Chapter III). This model's aim was to compare this representation of reaction time within the investigation scenario to evaluation results. IMPRINT Pro's micromodels and the subtractive method of calculating RT based upon atomic actions were used to estimate RT for the model.

### II.3.3 Human Performance Modeling for Robotics

The relationship between humans and robots is unique and must be accounted for when applying human performance modeling techniques to situations involving robotics and automation. The purpose of this section is to present a review of related work and the modeling paradigms that have been utilized and modified for these types of situations involving humans interacting with robotic systems. Table II.1 provides a summary of the presented research. The table details the model's focus (e.g., physical behavior, cognition), the purpose of using human performance modeling (e.g., predicting performance, system design), the validation metrics used and the chosen modeling tool.

Table II.1: Summary of human performance modeling relevant to human-robot systems

| **Authors** | **Model Focus** | **Purpose of Modeling** | **Validation Metrics** | **Modeling Tool** |
|---|---|---|---|---|
| Goodrich and Boer, 2003 | Cognitive | Identify human behaviors | car velocities, pedal state, cruise control state | Mental models |
| Olsen and Goodrich, 2003 | Cognitive | Predict performance | task effectiveness, neglect tolerance, robot attention demand, free time, fanout, interaction effort | Mathematical models |
| Crandall, Nielsen and Goodrich, 2003 | Timing | Predict performance | secondary task performance, neglect tolerance, interface efficiency, time to completion | Mathematical models |
| Jung et al., 2007 | Cognitive | System design | - | EM-ONE |
| Best and Lebiere, 2006 | Cognitive | Versatile agent development | - | ACT-R |
| Haddadin, Albu Schäffer and Hirzinger, 2007a, 2007b | Physical responses | Predict performance | robot movement speeds, crash test dummy injuries | Safety tree chart |
| Harriott, Zhang and Adams, 2011a, 2011b; Adams, Harriott, Zhuang, and DeLoach, 2012 | Workload and reaction time | Predict performance | subjective workload ratings, physiological measures, secondary task performance, item reaction time | IMPRINT Pro |
| Howard, 2006, 2007 | Workload | Predict performance | execution time, composite task scores | Fuzzy logic model, HumAnS-3D |
| Trafton et al., 2005 | Cognitive | System design | task success | Polyscheme |
| Gluck et al., 2005 | Cognitive | System design, personnel training | degree from correctness, spatial interference effect in visual working memory | ACT-R, VERBOSE |
| Hunn and Heuckeroth, 2006 | Workload, timing | Predict performance | surveys of subject matter experts | IMPRINT |
| Petkosek, Warfield and Carretta, 2005 | Workload | Predict performance | - | CART |
| Deutsch, 2006 | Workload, timing | Predict performance | - | D-OMAR |
| Ritter et al., 2006 | Cognitive, physical | System design | lane deviation, total time before crashing | ACT-R, Segman (DUMAS) |
| Kaber, Wang and Kim, 2006 | Cognitive | Predict performance | execution time, path error | GOMS |
| Liu, Rani and Sarkar, 2006 | Affect | Predict performance | anxiety levels, satisfaction with game, game scores, perceived challenge of game | Affective model |
| Trafton et al., 2008 | Verbal | System design | subjective ratings of conversation naturalness | ACT-R, ACT-R/E |
| Crandall, Cummings and Nehme, 2009; Crandall et al., 2011 | Attention allocation | Predict performance | number of objects found, number of robots lost in explosion | RESCU |
| Reitter and Lebiere, 2010 | Cognitive | Predict performance | normalized area between itineraries | ACT-R |

#### II.3.3.1 General Models

There are four factors to consider that affect the human's use of automation: 1) the Limitations of automation and the human's knowledge of the system, 2) the Responsibility transfer between human and automation, 3) the acceptability and predictability of the automation's executed behavior Dynamics and 4) the effect of using automation on system Efficiency. The effect that an automated system has on a human's driving was analyzed with respect to these four factors and the authors developed mental models of the involved processes in order to gain insight regarding human behaviors during driving and interacting with a cruise control (Goodrich and Boer, 2003).

The car driving task was divided into three categories of subtasks: speed regulation, time-headway regulation and active braking. Speed regulation involved the choice to accelerate, time-headway regulation determined following distance, and active braking occurred when the driver or system perceived the need to decelerate and press the brake pedal. Determining the benefits of added automation involved emulating drivers switching between each of the three subtasks, with switches precedent by perceptual cues. A divide between the braking and regulation tasks was found when evaluating these metrics between the models and participant behavior. The models helped inform why cruise control affecting speed regulation was acceptable, but active braking automation is not a good choice. These four principles of automation interaction and the notion that modeling human responses can inform design choices can be applied to other work.

Metrics, such as the amount of cognitive effort expended and the amount of free time available for a secondary task are relevant during human-robot system assessment and design. Using this knowledge, operators can compute how much time robot operators have to complete tasks in addition to robot control, which increases overall productivity. Goodrich and Olsen (2003) devised six interrelated metrics for analyzing human-robot systems: 1) task effectiveness, 2) neglect tolerance, 3) robot attention demand, 4) free time, 5) fanout and 6) interaction effort. Task effectiveness measures, relative to each task, how well a human-robot team achieves the task goal. Neglect tolerance reflects the level of a robot's effectiveness over time, since the human teammate last interacted with the robot and is a mechanism for measuring the robot's autonomy. Robot attention demand is the proportion of deployment time spent attending to the robot. Free time is the proportion of time spent not attending to the robot, and can be measured using performance on a secondary task. Fanout is a measure of the approximate number of robots a person can effectively operate and is the reciprocal of the robot attention demand. Finally, interaction effort is related to the time spent interacting or engaged with a robot. Interaction effort is not directly measurable, as recording the exact time when the human is cognitively focused on the problem is difficult. Instead, interaction effort can be determined using the relationship between the other metrics, such as neglect tolerance, secondary task performance and fanout.

Overall, the system goals may be to increase task effectiveness, neglect tolerance and free time, while reducing interaction effort. Team organization and task assignments can be modeled by considering these metrics and may prevent the deployment of an ineffective team. The metrics predict the performance of a potential team. These metrics have been applied in multiple robotics system evaluations (e.g., Crandall et al., 2003; Olsen Jr. and Wood, 2004; Elara et al., 2010).

Crandall et al. (2003) expanded upon Olsen and Goodrich's idea by creating an algorithm to predict the performance of a single-human, multiple-robot team and experimentally validated the results. The neglect tolerance and interface efficiency (the effectiveness of the interaction between a robot and a human) of human-robot teams were measured in multiple user evaluations. This information was supplied to an algorithm that determined time to completion for various system configurations of one human and three robots. These completion times were experimentally validated and shown to be good predictions of actual completion times, with the exception of configurations that were very difficult. Teams performed faster than the predictions in difficult configurations. Data from previous user evaluations can be used to create algorithms that successfully predict team performance in subsequent tasks.

While the previous topics focused on more general principles of modeling in human-system interaction, specific systems have been deployed that utilized modeling techniques for a single human interacting with a single robot, either directly or through remote operation. Additionally, modeling techniques have been applied to humans interacting with multiple-robot systems in order to glean information about such systems. The following subsections focus on these systems.

#### II.3.3.2 Single-Robot Interaction

Jung et al. (2007) sought to improve single-human, single-robot interaction. The authors believed that robots lack common sense and wanted robots to use prior experiences to evolve the way they respond to situations through cognitive human-robot interaction. The EM-ONE cognitive architecture (Singh, 2005) includes the human reactive, reflective and retrospective layers of thinking. This architecture was modified and integrated into a semi-autonomous script generating software system utilized by the robot path-planning. The software maintained task-related information, such as pre- and post-conditions for possible actions within the current scenario. Utilizing models based upon human cognition and task-related information, the authors were able to modify robot behavior and system design to become more human-like. The system is able to script robot behaviors and interactions with humans based upon its model of human information processing and learning.

Direct and immediate contact between the humans and robots or automation has been examined, evaluated and incorporated into human performance models. Domains where modeling has been tested include urban military scenarios (Best and Lebiere, 2006), industrial robotics (Haddadin et al., 2007a,b), and space

exploration (Howard, 2007). The following cases offer examples of modeling teamwork, injury, workload, completion time, affect, and perspective-taking.

Urban military scenarios can involve human-robot teams acting as teammates to achieve goals according to structured protocols. Best and Lebiere (2006) used the ACT-R cognitive architecture system to develop human and robotic agents interacting as a team in an urban military operation with clear rules and structured goals. The objective was to create agents that can be used in simulations of team tasks in real and virtual environments. Robotic agents were developed based upon the Pioneer 3-DX general-use robot incorporating its visual and navigation systems. Teamwork between the agents involved sharing plans, communication of the steps involved to complete a plan, and a method of sharing a representation of the surrounding space. The authors proposed the possibility of extending the use of cognitive models, which are not tied to lower-level environment details, to other domains. The created agents were used to assist in training military personnel for urban terrain operation.

Some models of human-robot systems can quantify possible outcomes of physical interaction. Haddadin et al. (2007a,b) investigated safety considerations for physical human-robot interaction, specifically with industrial robots. The authors first evaluated the safety of robot systems colliding with the human head, neck and chest. These tests cannot be performed using actual humans; therefore, crash-test dummies were used. Equations outlining the predicted severity of an injury to the head, neck and chest were adapted from the automobile industry, but were shown not to extend to the much slower speeds in physical human-robot interaction. New models of physical human-robot interaction were necessary and the authors later developed tests and simulations to investigate the impact of robot mass and velocity on robot-induced human injuries. A safety tree model was developed to classify possible injuries and predict worst-case scenarios. This model is a chart that provides an outline of worst-case scenarios and possible injuries that can occur during the human-robot interaction. A user of this model can analyze the scenario conditions, find the matching conditions in the model and determine levels of risk during the physical human-robot interaction. It was concluded that blunt impacts to the head or chest were not life threatening no matter the robot's size, as long as the human is not pinned against a hard surface. These findings can impact the design of human-robot systems by allowing for more accurate predictions of injury. Using this model, system designers can evaluate the tradeoff between increasing robot speed and size with possible injury to those working in close proximity with the robots.

Space exploration involves human-robot teams with missions that involve detailed planning before execution. Howard (2007) worked to optimize role allocation for these human-robot team systems by incorporating the task switching process of alternating attention between tasks and assessing the workload levels of human teammates. Team performance can be optimized by intelligently choosing whether the human or robot is assigned each role. Increasing the time interval between the task completion and a subsequent stimulus for

the next task was shown to reduce the task switching cost. Switching between similar tasks has a lower cost, but switching from a set of similar tasks to an unrelated task has a higher cost. A fuzzy logic model of these observations was created and helped develop a set of algorithms to determine optimal system performance based upon human workload and expected performance. Using the virtual test environment HumAnS-3D (Howard and Paul, 2005), the authors compared a fitness function based solely on individual performance values to the new methodology that incorporates switching costs. The new task-switching model resulted in faster completion times, showing that using customized models to aid in role allocation for the team scenario can improve system performance.

Robot behavior can be modeled after human behavior to provide a more natural interaction within human-robot systems. Trafton et al. (2005) argue that modeling human-robot systems based upon human-human interaction is best. Three conceptual guidelines to consider in human-robot systems are: (1) robot perception, reasoning, and representation need to be similar to human systems, (2) cognitive systems need to integrate cognitive architectures, and (3) heuristics similar to those which humans use are effective. The authors designed a robot that was able to create a representation of the visual perspective of the human with whom it was communicating to make the human-robot conversation more like one held between humans. After analyzing videos of human-human interaction during astronaut training, a model using the cognitive architecture Polyscheme was created to represent the human behavior of simulating another visual perspective (Cassimatis, 2002). Using Polyscheme to create a representation of the environment from the human's perspective aided the robots by helping to disambiguate commands from and actions taken by the human. The model was placed on a robot, which was given ambiguous commands from a human to select objects in the room. The robot was unable to solve the ambiguities before the perspective-taking model was added. Integrating a skill based on human-human interaction aided the human-robot conversation.

Human performance models and predictions are useful for situations where humans directly interact with robots, such as role allocation, robot behavior design and workload estimation. Similarly, interacting with a robot in a remote location can also be improved by modeling human performance patterns and limitations.

A common interaction with a remote robot is with unmanned aerial vehicles. Often these robots perform surveillance or reconnaissance tasks. Spatial awareness and orientation are important when working with a remotely located robot. Gluck et al. (2005) developed cognitive models using ACT-R to simulate a human's accuracy in three-dimensional spatial orientation tasks, with the eventual purpose of using these models to aid system design and provide additional training opportunities. Experimental validation showed that these models provided good predictions of human behavior. Models simulated control of Predator unmanned aerial vehicles in the Predator synthetic task environment (Schreiber et al., 2002). The models were shown to be good predictors of expert pilot performance for the basic maneuvering tasks. Interfacing the synthetic

task environment with ACT-R was achieved by re-programming the visual display for the task environment. Extending this research is the VERbalization Between Operators and Synthetic Entities (VERBOSE) project, which aimed to create simulated agents that can use verbal communication as a computational cognitive linguistic system (Ball, 2004). Simulation language skills and task performance will provide a valuable training tool for pilots of unmanned aerial vehicles.

Some unmanned aerial vehicles require multiple people to support their operation. Specifically, the Shadow 200 System was determined to require approximately 22 to 28 personnel by an IMPRINT model (Hunn and Heuckeroth, 2006). The model's goal was to explore workload of and task assignment to the human operators of the Shadow 200. The model inputs of time required, number of people required, and frequency of each subtask were determined by surveying military personnel who possessed training and experience with these unmanned aerial vehicle systems. Workload values for each subtask were assigned according to IMPRINT's built-in guidelines. The model was used to provide estimates of the crew's workload during the mission, individual crew member workload, baseline levels of workload for each crew member, how long the mission will take, and how many personnel it takes to complete the task. Using data collected from field-experienced personnel and IMPRINT's workload level assignment guidelines, the model was able to answer all of these questions. The model was not validated with an empirical evaluation, but the use of subject matter expert data to create the model provided useful information for the authors.

Petkosek et al. (2005) executed a similar investigation into unmanned aerial vehicle operation scenarios by creating models using the Combat Automation Requirements Testbed (CART) to model task completion time and workload levels for unmanned aerial vehicle operators. CART is based on the IMPRINT tool. The scenario involved the surveillance of a potentially dangerous situation, including refueling a number of grounded passenger planes. The models served as an investigation tool for task decomposition and workload estimation, but were not experimentally validated.

Deutsch (2006) used a variant of the same scenario and created three human performance models of each of three unmanned vehicle operators. Using the Distributed Operator Model Architecture (D-OMAR), the authors utilized a test-bed for the unmanned aerial vehicles and were able to test the impact of the tasks on each of the three operator roles: sensor operator, aerial vehicle operator and multi-function operator. The model's goals included analyzing workplace design and operating procedures, improving model robustness and reducing training time. Possible team tasks featuring these three operators can be simulated using this model to test potential assignments and to identify minimal staffing combinations before spending the time and money for testing using actual unmanned aerial vehicle equipment and personnel.

Urban search and rescue is a burgeoning avenue of research for remotely located robots. Teleoperation in an urban search and rescue environment is difficult and requires multiple tasks often performed under time

pressure, (e.g., steering the robot and monitoring sensor feedback). Ritter et al. (2006) aimed to create a model of an operator that can improve human-robot systems in these domains by using a simulated operator to test potential designs. A vehicle-driving user model was created using the ACT-R cognitive architecture and the SegMan eye and hand simulator (St. Amant et al., 2004). This model was called the Driving User Model in ACT-R and SegMan (DUMAS). After the experimental validation, two limitations of the model were determined: the lack of a complex interface in the test and the model's visual system. The model was extended to process visual information in real-time, acting more similarly to the human visual system, and was tasked with teleoperation (Ritter et al., 2007). An evaluation compared human performance to the model, using metrics of lane deviation and total time before crashing. The model was shown to be a reasonable predictor of human performance. This work demonstrated the ability for a model of human behavior to interact with a real-world robotic system in real-time, while performing complex visual processing. Using similar models of human users can aid in the development of increased efficiency for interfaces with remotely located robots.

Kaber et al. (2006) extended a GOMS Language model for a human-robot system to predict human performance while teleoperating a robot. The GOMS Language model code was fed into the Error-extended GOMS Language Evaluation and Analysis (EGLEAN) compiler. The model emulated human control of a single small ground vehicle traveling on a simple path at a fixed speed. Data gathered from a single operator executing the task guided the model's creation. The model predicted an execution time that was approximately ten times the actual human performance time, but with one tenth of the human error on path navigation. While the time and accuracy predictions were not perfect, this work demonstrates the ability to use GOMS models to represent human behavior in a human-robot system. However, this work also suggests that GOMS may not be the best way to provide accurate timing or error-rate predictions.

The above systems are traditional examples of human performance modeling. A broader definition of human performance modeling can include affective models. Models of performance can be developed during interaction with a system, rather than created as a prediction of the interaction with a system. Models created during interaction can help to adjust the simulation for aiding the individual human working with the system. Along with workload levels and task completion time, a person's affect can offer insight into performance. Rani et al. (2004) developed a means of monitoring a human's affective state through physiological responses while interacting with a robot and used signals from the human to modify the robot‘s behavior. Liu et al. (2006) expanded upon this methodology and altered a robotic basketball hoop's movements according to participant anxiety levels. Participants threw basketballs into a moving robotic basketball hoop and performed at low, medium, or high levels with low, medium, or high anxiety. The robotic basketball hoop was able to produce lower levels of anxiety and higher levels of performance when responding to the participants' affect.

These affective models of human anxiety levels map how changes to the experimental situation affect the state of the human playing the game. An individual model of affect was required for each participant. Liu et al. (2009) also used physiological measures to create a model of human affect usable during human-automation interaction. Participants wore a physiological monitoring system that recorded anxiety levels when interacting with a video game. The authors compared interactions using information regarding the user's affective state to modify the game's difficulty and compared it to games where difficulty was modified only by participants' task performance levels. Each participant went through a training period and the data was used to create an affective model. Alternate methods of creating an affective model were tested and the regression tree-based affective model provided the most accurate predictions (Rani et al., 2007). The affective model-based game modification resulted in higher performance levels for most participants. Participants also rated the modified game as more challenging and satisfying, and participants perceived lower levels of anxiety. While a limitation of this technique includes the extensive training period to build an affective model for each participant, the benefit to performance of developing these systems is apparent. Modeling the internal state of each human benefited the system's overall performance. Performance can be affected by internal factors, such as workload, interaction effort, and anxiety.

Additionally, human performance modeling can be considered when using a cognitive architecture to emulate human behaviors by a robot. Trafton et al. (2008) modeled a robot's behavior during conversation with a human, based upon behaviors present in human-human interaction. People speaking in group conversations typically look at whoever is talking, but wait approximately 500ms to switch to a new speaker. ACT-R (Anderson and Lebiere, 1998) was adapted into ACT-R/E for an Embodied robotic conversation member by connecting the audition and vision modules to a real robot and adding pedal and spatial modules for moving about and representing the environment. The robot was tasked to look at whichever person was currently speaking, even in the presence of challenges, such as ambient noise and speaker interruptions. An evaluation compared the perceived naturalness of the robot looking at each speaker with and without the 500ms pause before switching gaze. The addition of the very human-like pausing period was perceived as much more natural compared to the system that did not wait, indicating that a more human-like behavior is seen as more natural, showing that the system design was improved by including a model of human behavior.

#### II.3.3.3 Multiple-Robot Interaction

Interactions with multiple robots have also been analyzed through human performance modeling. General models of the interactions will first be presented followed by specific systems that have implemented human performance modeling for multiple-robot systems.

Determining how many robots, robot autonomy levels, and appropriate system design are not obvious

choices. Crandall and Cummings (2007) proposed a set of metrics for the control of multiple robots performing independent tasks in remote locations. Metric classes are categories of measurements that represent the effectiveness of a system and contain key performance parameters of the team, identify the limits of agents in the team and are able to generalize to other situations. The supervisory control of multiple robots differs from a single-robot case in that an important part of the system is the division of the human's attention between each of the robots. A metric class specific to multiple robot interaction is attention allocation efficiency, encompassing global situational awareness, switching times, selection strategies, the percent of time spent following an optimal policy, and wait times due to loss of situation awareness. The interaction efficiency and neglect efficiency are two metric classes that are relevant to both single robot and multiple robot control. Interaction efficiency measures the effectiveness of the team and neglect efficiency represents the efficiency of a single robot when neglected by the human. The important factor in the supervisory control of multiple robots is the addition of the necessary division of the human's attention between robots.

The metric classes for multiple-robot control were extended to create a model of the interaction based on human behavior and robot performance (Crandall et al., 2009). The need for attentional division amongst multiple robots was modeled in an emergency response scenario using the Research Environment for Supervisory Control of Unmanned Vehicles (RESCU). The scenario involved a search and rescue mission where an operator controlled multiple robots and was tasked to find as many objects from a building during the eight minutes before the building explodes. The model predicted both the number of objects found and the number of robots lost in the explosion. The difference between the number of robots that were left in the building to explode and the number of objects found evaluates to the overall score for the mission. Two interface types and two levels of autonomy for the robots were tested. The results showed that the models successfully predicted the number of objects collected and the number of robots lost in the explosion. The model was a good predictor of team behavior, but has only been shown to apply to the single-human multiple-robot situation where the robots do not collaborate, there is one human operator and the operator cannot interact with groups of robots together. Despite these limitations, the model provides a good prediction of team performance within the specified assumptions.

As stated above, when a human supervises a team of multiple robots, attention must be divided between the robots. Crandall et al. (2009) informed the need to improve attention allocation. Crandall et al. (2011) created a system to determine what human operators can attend to at a given time, based upon a model of optimal attention allocation using the prior scenario. Each robot's state was evaluated and categorized according to the priority of that state given the current time, (e.g., during the last minute it was more important to move idle robots and pick up objects, but in the first minute re-planning paths and assigning tasks took top priority). Model predictions indicated that optimizing attention allocation can increase system performance.

A second user evaluation focused on the impact of attending to all manual choices of a robot, all automated choices, and a guided mode that provided suggestions. Overall scores did not improve with the automated choices and participants preferred the guided mode. This work demonstrates that optimal attention allocation may not be necessary and aiming for satisficing (i.e., strategy to satisfy requirements without aiming for the optimal (Simon, 1956)) choices with flexibility may be ideal for human operators.

An important aspect of moving about any physical environment is path planning. Human-robot systems that include remotely located robots sometimes require that an operator designate a path for the robot to follow. Reitter and Lebiere (2010) created an ACT-R cognitive model of the human path planning process by incorporating information about the environment retrieved from declarative memory and visual search of the environment. The model's memory components defined location memory chunks that represented a system's state and path memory chunks that represented transitions between location chunks. Production rules allow the model to create subgoals for moving between locations, utilizing recorded memory of paths to the goal. The model's visual aspect allows for the choice of paths to subgoals within a short visual range using straight lines and without using memory chunks or production rules. Incorporating the two strategies allowed the model to successfully navigate mazes. The memory-based strategy model and the visual-based strategy model were each validated separately before using a combination of the two in a third evaluation where participants controlled 4-, 8- and 12-member robot teams through a computer interface. The goal was to investigate the layout of an unknown building. Participants created itineraries for robots traveling through the building, which included the path's start and end points. The difference in itineraries between model-created itineraries and the participant data was computed using trajectory area normalization, which determines the area of the space between the two itineraries. The model was able to predict the robot itineraries created by participants very closely. These results validated the two-part cognitive model of human path-planning.

Human performance modeling has been applied to robotics and automation interaction in a variety of ways, (i.e., interacting directly with a single robot, remote interaction with both single and multiple robots, teleoperation concerns and teamwork). Metrics have been designed for both single-human single-robot interaction and a single human supervising multiple remote robots. Human performance can be emulated with robot behavior or analyzed to identify trends resulting from working with robots, as opposed to humans. System design can improve based upon modeling comparisons, and workload levels from common missions can be analyzed using human performance models. Overall, introducing human performance modeling techniques to robotics can occur in vastly different and valuable ways.

### II.3.4 Model Verification

An important component of the performance modeling process is the verification of the resultant model (Allender et al., 1995; Pew and Mavor, 2008). It is necessary to verify that the resultant model provides predictions that are sound and within an expected range of realistic bounds. Often, a model is developed for a base condition for which data exists, thus allowing the resultant model to be compared to the existing data set. Additionally, the model can be verified via user evaluations, either simultaneously or after modeling, that capture a representative data set for verification. Either approach can result in modifications to the developed model.

The verification of human performance models, especially when applied to robotics, is critical. Evaluating the degree to which a model represents real system performance typically involves a user evaluation. Gawron (2008) proposes considerations for performing an experiment including: 1) define the question to be answered with this experiment (e.g., does the model predict accurate task times? Is timing different between Task A and Task B?), 2) check for qualifiers, which restrict the generalizability of results, 3) specify the conditions (e.g., Task A - modeled and Task A - experimental) and whether they are between- or within-subject conditions, 4) select performance measures, data collection equipment and data recording equipment, and 5) match trials with the same length, level of difficulty and environmental conditions.

Once the question is defined and a scenario is outlined, the model aims to include what the participants experienced throughout the evaluation and how the robot or automated system is incorporated in order to match the environmental conditions in the model and evaluation in consideration 5. Consideration 4 is crucial to the verification of models through experimentation. A modeling tool is chosen that can represent and model the metrics to be measured in the evaluation. John and Newell (1989) recommend that a valid model does not deviate more than 20% from human performance. Possible metrics include primary task performance, secondary task performance, physiological measures and subjective responses. Additionally, a test of model validity can include testing whether evaluation results fit within a 95% confidence interval of model results (Roberts and Pashler, 2000). If the evaluation results are within the model's confidence interval, the model is a good fit.

Task performance and task timing measures can be directly compared between the models and experimental results. Steinfeld et al. (2006) offered a set of metrics specifically for human-robot systems. While the robot's performance alone is not the focus of a human performance model, the robot's effect on the human is and metrics, such as fanout and time for the human to notice a robot's request can be modeled and compared to experimental results. The relationship between the human and robot is also an important consideration. During teleoperation, other metrics, such as obstacles avoided or percentage of time on the correct path can

be relevant validation metrics.

It is important to know what question is being asked in the evaluation and derive the correct comparable measurements to evaluate whether the model truly represents human behavior. Overall when choosing to validate a model of human performance, it is important to choose metrics that can be directly compared, whether gathered through physiological measures, subjective ratings, timing data or task performance.

### II.4 Summary

The focus of this dissertation is to investigate how workload and task performance differ between human-human and human-robot interaction. Achieving this goal included determining the appropriate measurement techniques for assessing workload and task performance, including reaction and response time, in human-robot peer-based teams. Additionally, workload and the task performance metrics reaction time and response time were modeled using the human performance modeling tool, IMPRINT Pro. This chapter presented an overview of human-robot collaborative teams (see Chapter II.1) and workload and task performance measurement (see Chapter II.2). This chapter also discussed human performance modeling, examples of modeling workload, RT, and in robotics, and model verification techniques (see Chapter II.3).

# Chapter III

## The Guided and Collaborative Evaluations

This dissertation assesses task performance (including reaction time and response time) and workload levels (including physical workload) of human-human (H-H) and H-R collaborative teams in order to address potential differences between them. Two evaluations were performed as a first step in evaluating workload and task performance in human-robot peer-based teams, the Guided and Collaborative evaluations. Each evaluation centered on a scenario involving first response tasks and captured a variety of workload and task performance metrics. Each evaluation paired participants in H-H or H-R teams to complete the evaluation tasks.

The Guided evaluation was performed in Spring 2010 and involved a triage scenario (see Chapter III.1). Teams were tasked to perform triage steps on non-ambulatory (i.e., unable to walk) victims. The Collaborative evaluation was performed in Summer 2011 and involved human-human and human-robot teams investigating a floor of an academic building for suspicious items (see Chapter III.2). The Collaborative evaluation focused more on building a dynamic and conversational relationship between teammates.

There are three primary questions addressed by this chapter: 1) Is there an effect of the introduction of a collaborative relationship on the human's mental workload and task performance? 2) Are there general trends in workload and task performance that differ between H-H and H-R teams? 3) Is human performance modeling a useful predictive tool for mental workload in H-R interaction?

Details from each evaluation are presented, including a brief description of the primary and secondary tasks, the modeling description, the experimental method, and the evaluation results. The results include overall and mental workload results, physical workload results, and reaction time and response time results. The results of the two evaluations are compared in Chapter III.3).

### III.1 The Guided Evaluation

The Guided evaluation scenario required participants to play the role of an uninjured ambulatory victim during a mass casualty incident involving contaminants (e.g., a chemical). Standard procedures (Humphrey and Adams, 2011) indicate that first responders are not to enter incident sites involving unknown containments until it is known what personal protective equipment is required and a decontamination area is established. The scenario assumed that the incident had occurred recently, and there was insufficient information to identify the contaminant and deploy first responders. The participant was considered "contaminated" and was not permitted to leave the area until a decontamination area was established. The participant was either paired

with a remotely located human or a locally situated robot.

The set of tasks included a rapid triage assessment using the Simple Triage and Rapid Treatment (START) technique (Benson et al., 1996). Trained responders typically complete the START triage steps in 60 seconds (Benson et al., 1996). A responder begins by assessing the victim's breathing rate, pulse, responsiveness and injury level and then assigns a triage level of Minor, Delayed, Immediate or Expectant. The adapted START steps are outlined in Figure III.1. Victims classified as minor are ambulatory and coherently responsive; this classification was not included in the evaluation. Delayed triage victims can survive while waiting up to a few hours for care. Such victims have breathing rates of less than 30 beats per minute, a regular pulse and are mentally responsive. Victims classified as immediate require treatment as soon as possible. Immediate victims are breathing, but are either unresponsive, have an extreme breathing rate (over 30 beats per minute) or do not have a regular pulse. Expectant victims have either passed away or will soon expire. Such victims may not be breathing, even after simple resuscitation attempts.

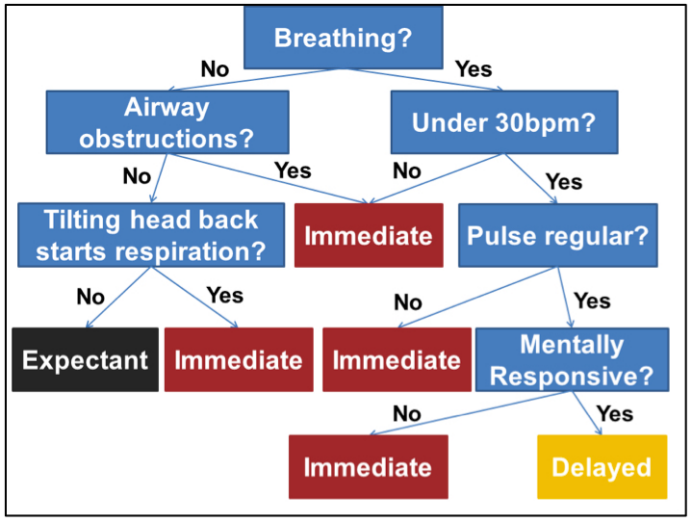
The hypotheses of the Guided evaluation included:

- $H_G^1$ = There is a measurable difference in overall and mental workload between H-H and H-R teams.
- $H_G^2$ = Human performance models, developed using tools not validated for human-robot peer-based teams, can be adapted to represent workload in human-robot peer-based teams.
- $H_G^3$ = Physical workload is higher in human-robot teams.
- $H_G^4$ = Reaction time and response time represent different measurements in human-robot peer-based teams.

### III.1.1 Modeling Description

Human performance models were developed using the IMproved Performance Research INtegration Tool (IMPRINT) Pro, where each teaming condition was modeled (Allender et al., 1995; United States Army Research Laboratory, 2009). IMPRINT Pro was developed by the United States Army Research Laboratory and represents a discrete event simulation modeling system. IMPRINT Pro has been used to model personnel on a United States Navy destroyer bridge, the Land Warrior System and the U.S. Army's Crusader System (Allender, 2000) all for improvements to existing U.S. military systems. IMPRINT has also been used to model pilot performance while flying simulated unmanned air vehicles (United States Army Research Laboratory, 2009). Typically IMPRINT has been used for military simulations and evaluations of possible improvements to existing systems.

Figure III.1: The START-based triage steps.




IMPRINT Pro permits the simulation of human behavior for a variety of conditions through the representation of task and event networks. IMPRINT Pro includes a number of pre-defined human performance moderators (e.g., workload) and permits the incorporation of those performance moderators not already predefined via the User Stressors module (United States Army Research Laboratory, 2009). IMPRINT Pro has been employed in the reported research to model both the human-human and human-robot conditions.

Models built in IMPRINT Pro use atomic task time, task ordering, number of crew members, training, equipment, stressors, and operator mental workload for each task as the model's inputs. Model outputs include values that measure mission success, mission time, and an individual's mental workload per unit of time. The stressors contained in IMPRINT Pro include a variety of human performance moderator functions, such as temperature and humidity, whole body vibration, and noise level. Stressors can affect the timing and accuracy of tasks, which affects the number of tasks that can be accomplished in a certain amount of time by an individual and that individual's overall mental workload level during a mission.

The models in this research were developed for each task scenario and for the two teaming conditions: human-human (H-H) and human-robot (H-R). The resulting models are sequential networks representing subtasks that individuals perform during the scenario.

The Guided scenario models integrated a highly rigid script based on the triage steps. The model assumed a linear sequence of steps for the victim triage and did not account for missed steps or alternate paths; therefore, this model incorporated no uncertainty. Multiple simulations of the models provided identical results (Harriott et al., 2011b).

The triage scenario required the participant to perform the START triage steps on six victims with differ-

ing levels of required triage and then repeat the triage steps on all victims who were not classified as Expired during the initial triage. During the second round of triage, victims were visited based on the triage level determined during the first triage round; those determined to have the most severe triage level were visited first. The second triage round of triage did not require visiting the victims in the same order as the initial triage.

The modeled H-H scenario assumed that the participant contacted 911 to report the incident and volunteered to assist a remote (e.g., located outside of the contaminated incident area) first responder partner with the triage task. The scenario further assumed that the participant communicated with the responder partner via cell phone, and the responder provided step-by-step instructions that lead the participant through the triage steps. The participant provided responses that were recorded by the responder partner in order to assist with incident response planning.

The modeled H-R scenario assumed that the robot partner was deployed into the contaminated incident area and came across the participant who agreed to assist the robot with the triage task. The participant executed the triage instructions provided by the robot partner and reported results to the robot. The robot also provided color-coded triage cards for the participant to place on the victims once a triage level had been determined. The robot reported this information, as well as the participant's location to incident command. The robot communicated with the participant using voice interaction. The H-H and the H-R models differed in the timing of some tasks, because the robot spoke 1.3 times slower and traveled the same distance 1.5 times slower than the human partner.

Both scenarios used the same task, which required performing an initial triage assessment and classification on the injured victims before conducting a follow-up second triage round. The victim order, provided triage instructions, and victim information are identical across the conditions. The human-robot model factored into account the robot's slower speech pace and the extra step of placing a triage card on each victim.

The IMPRINT Pro models iterated through each task involved in the entire triage scenario. Tasks included each START triage step for determining each individual victim's needs. For example, when the participant is asked to count the number of breaths a victim takes in one minute, the model includes each step for the discrete, atomic tasks; the participant must listen to instructions about how to complete the task and begin counting when instructed. Next, he or she must watch the victim's chest rise and fall for one minute, while counting the number of breaths and listening for the responder to say "stop." The participant reports the total number of breaths counted.

All steps involved for each victim can be grouped together into a higher-level function. Figure III.2 is the function for completing triage on victim 3, an 8 year old boy, during the first triage round. The rounded-corner rectangles indicate the individual tasks. The transition lines show the progression from task

to task. The workload questions are a separate task signaled at the end of each victim assessment, but before moving to the next victim. The overall model represents the tasks executed by the uninjured victim and the corresponding modeled team member, the remote human responder or the robot responder.

Each modeled task requires a specified running time, title and workload values. Each mental workload channel has a range of associated values. The Auditory, Cognitive, and Fine Motor channel values range from one to eight, the Visual and Gross Motor channels' ranges are from one to seven, and the Tactile and Speech channels' ranges are from one to five. IMPRINT Pro provides task timing guidelines based on micromodels of human behavior developed from published psychology, human factors, and military user evaluation data (e.g., walking ten feet takes approximately 1.9 seconds) and task demand guidelines based on task type (e.g., walking on level ground is assigned a Gross Motor demand value of 1.0) (United States Army Research Laboratory, 2009). Upon running the model, the assigned workload values for each task are in effect during the entire running time for each atomic task. Calculating the workload for an entire victim or function requires weighting each task's workload values for the portion of time the specific task takes in the function.

#### III.1.1.1 Modeling Results

The Guided evaluation model analyzed timing and workload levels for each of the eleven triage assessment periods that participants completed in the human-human (H-H) or human-robot (H-R) conditions. Two models were created, a H-H model and a H-R model. Once a model completes execution, the model outputs the list of tasks completed by the uninjured victim when triaging the injured victims. Along with each atomic task, the results include the time required to complete the task and the associated workload value for each workload channel and an overall workload value. Figure III.3 displays the total time taken to complete triage tasks for each victim by each of the models, H-H and H-R.

The model was used to specifically represent physical workload by examining the motor and tactile channels of workload. The mean modeled motor workload for the H-H model, taken from the eleven victim assessments, was 5.19 (St. Dev. = 0.95) and was 4.87 (St. Dev. = 0.65) for the H-R model. A t-test found that the difference between conditions was not statistically significant. There is no standard deviation for each victim assessment, because the modeling method produced a single workload prediction.

While the mean modeled motor workload was not significantly different between the conditions, the overall trend demonstrates that the H-R condition has lower motor workload than the H-H condition for ten of the eleven victim assessments. The modeled physical workload was compared to the evaluation measures of physical workload in Chapter III.1.3.2.

It can be seen that overall, the human-robot team took a longer time to complete all tasks. Both teams experienced very similar trends in workload with peaks and valleys in similar spots within each victim as-

Figure III.2: Guided scenario IMPRINT Pro model example of the function to perform triage on 8-year old victim in first triage round.

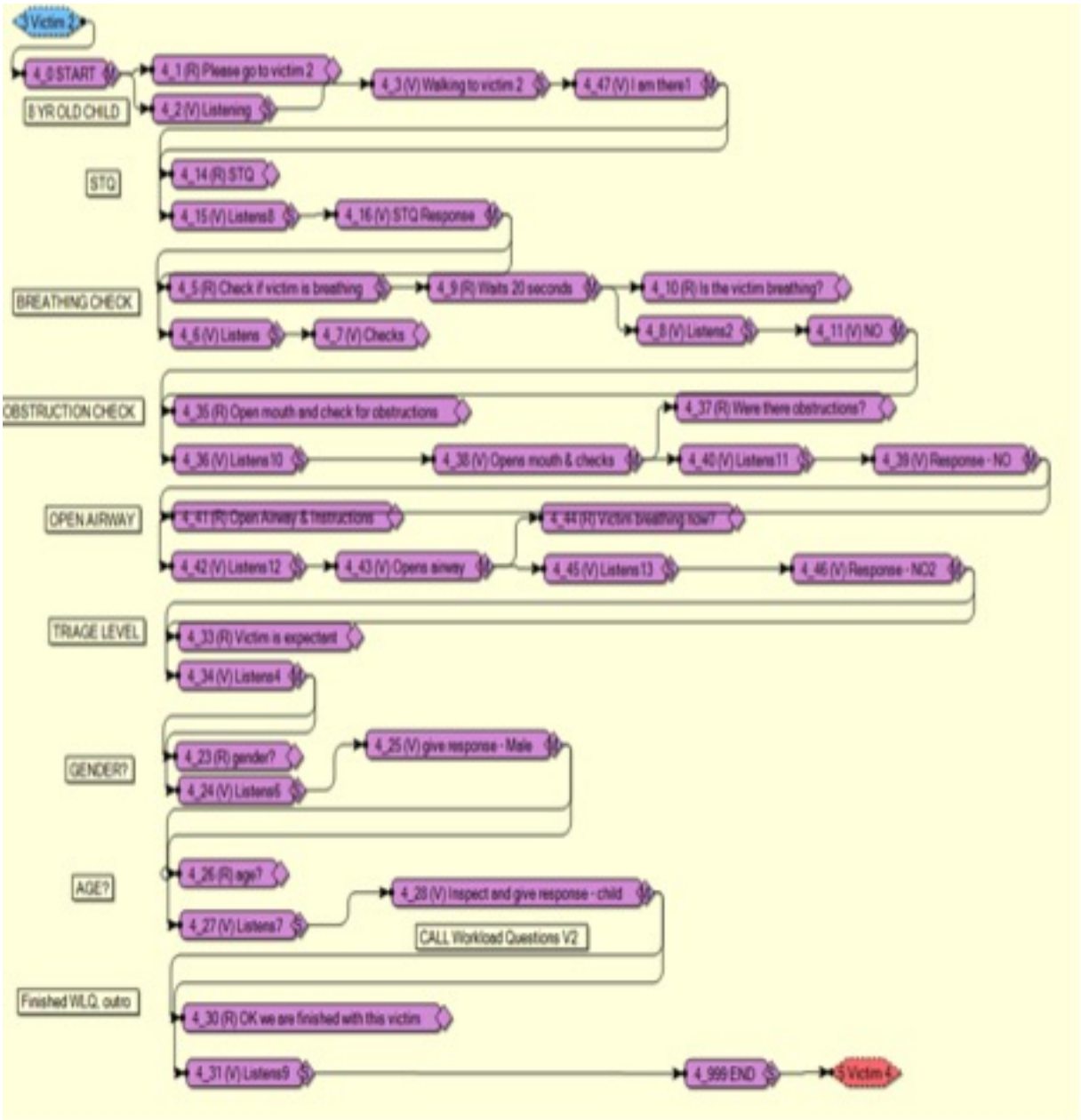


Figure III.3: Predicted triage time by victim for each model, H-H and H-R. The abbreviations along the figure's x-axis represent the victim number followed by round number, for example, V1R1 refers to Victim 1 assessed during Round 1.

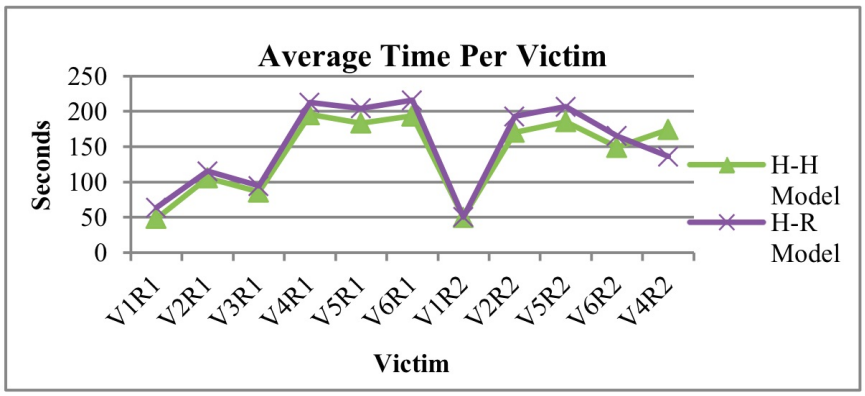



Figure III.4: Overall workload from the H-H and H-R models.

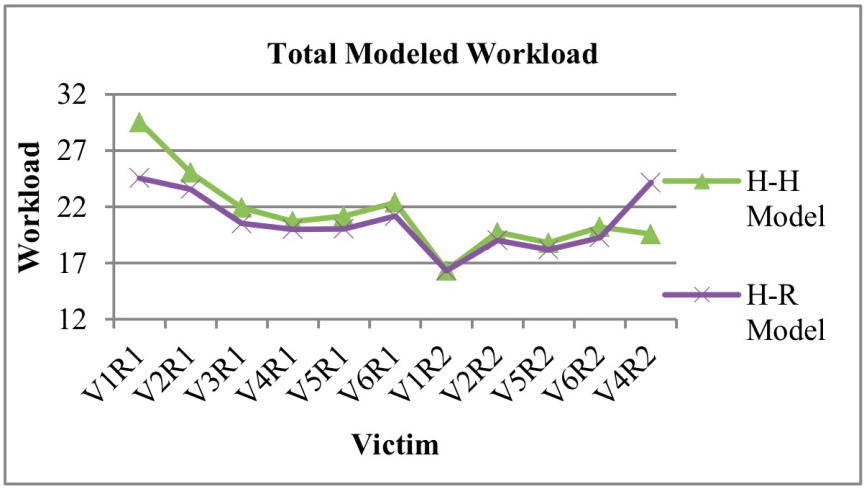



Figure III.5: Individual workload channel values for H-H model.

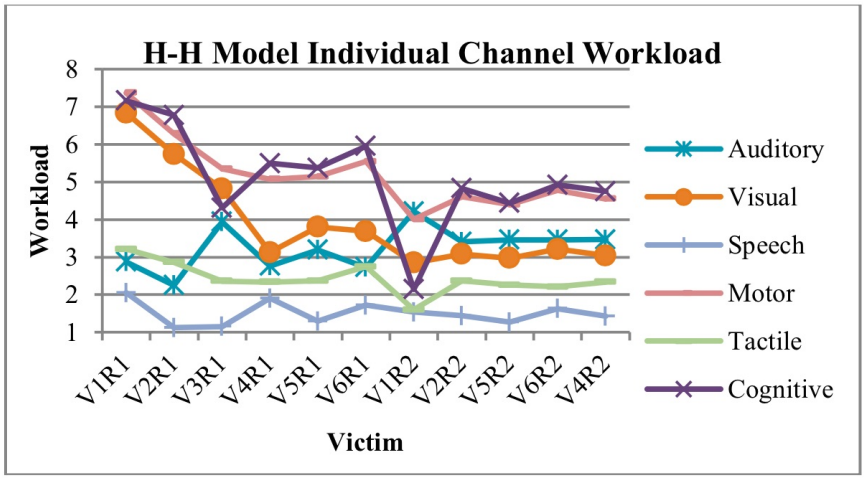
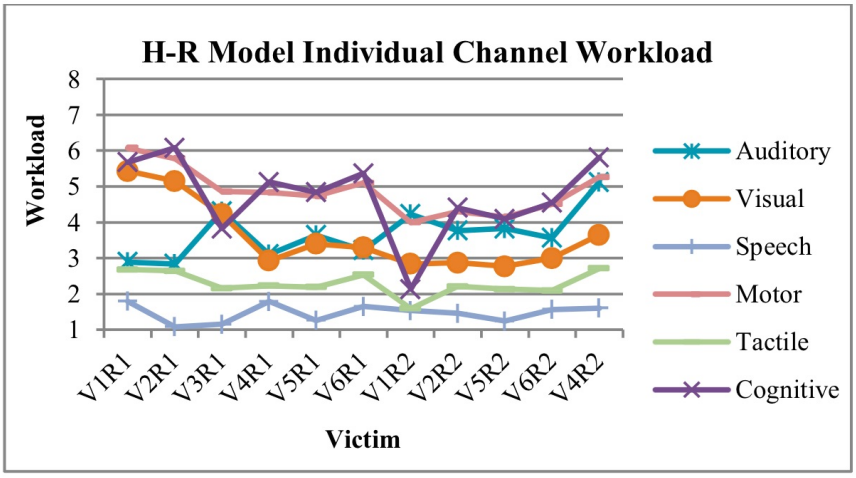


Figure III.6: Individual workload channel values for H-R model.




sessment period, but the overall workload may be different due to the time weight for each task. Figure III.4 shows the overall total workload score for each victim for the two models. Figure III.5 displays the workload values from each individual workload channel for the H-H model, while Figure III.6 provides the individual workload channel values for the H-R model.

The models predict changing workload values based on the specific victim assessed. The models also predict that total workload for both conditions will follow the same general trend independent of teaming partner, either human or robot. The human-robot team was expected to exhibit lower overall workload based upon the modeling results (Figure III.4).

### III.1.2 Experimental Method

#### III.1.2.1 Environment

The Guided evaluation occurred in the Center for Experiential Learning and Assessment within the Vanderbilt University School of Medicine. The pre-trial briefing, evaluation and post-trial debriefing all occurred in different rooms.

The evaluation room was a large open room with no windows, but included two-way mirrors along one wall. The evaluators were located in the control room behind the two-way mirror in order to monitor the evaluation. During the evaluation, the room lights were dimmed and a background noise track played in order to create a more realistic and stressful environment for the triage assessments. The background noise was a compilation of explosion noises, street noise, people coughing and screaming, construction noise and sirens. The noise was kept at a low enough volume so the participant was able to clearly hear his or her teammate. The room was kept at a temperature of $67 - 69^{\circ}$ F.

Figure III.7: Right half of Guided evaluation space. From bottom left, clockwise: Victims 4, 5, 6, and 3.

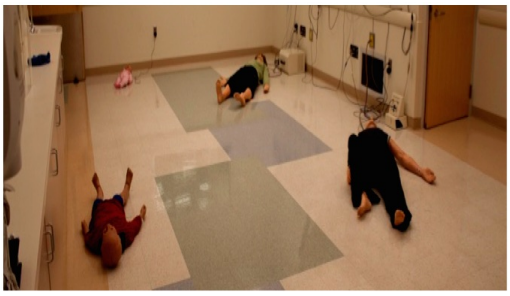


Figure III.8: Left half of Guided evaluation space. From bottom left, clockwise: Victims 3, 1, 2, and 4.

Six medical training mannequins were distributed on the evaluation room floor. All mannequins were dressed as civilians. The mannequins had capabilities that include breathing rate, pulse rate, blinking eyes, speech and injuries. The six mannequins are shown in Figures III.7 and III.8, from different angles of the room. Four mannequins' breathing rate, pulse and responsiveness level were predetermined and controlled by the experimenters, the fifth mannequin was an infant and only able to cry, while the sixth mannequin was a child and had no responsive behaviors. This sixth mannequin was used as an expired victim. The four active mannequins had a predetermined breathing rate and pulse rate. Three of these mannequins were adults and the last was a toddler. Two of the adult mannequins had speakers. These mannequins were used to determine as responsiveness by the participants when the participants asked them questions. These mannequins responded with their names; when asked, the Center's technicians sent a response through the mannequin's speaker as if spoken by the mannequin. Another of the active mannequins was a toddler and while able to have a breathing and pulse rate, it did not have a speaker. The fourth active mannequin had eyes that blink via technician control. When the participant was instructed to ask this mannequin to open and close his eyes, the mannequin responded by blinking. The neo-natal baby mannequin also had a speaker that was used to make the baby cry. The capabilities of the mannequins with their associated victim numbers are summarized in Table III.1.

Both conditions used the same medical mannequins placed in the same locations distributed throughout the space. Table III.2 describes the conditions and triage levels of all victims assessed by participants. Both conditions required the participants to kneel next to the victims to conduct their tasks. The participants were not required to stand up between triage steps and during the H-H condition, most participants remained kneeling when triaging a specific victim. However, during the H-R condition, many participants proceeded to stand up between triage steps.

Table III.1: Capabilities of six medical mannequins used in Guided evaluation.

| Mannequin | Age | Gender | Capabilities |
|---|---|---|---|
| Victim 2 | Adult | Female | Pulse<br>Respiration Rate<br>Voice |
| Victim 3 | Child | Male | None |
| Victim 4 | Adult | Male | Pulse<br>Respiration Rate<br>Voice |
| Victim 5 | Toddler | Female | Pulse<br>Respiration Rate |
| Victim 6 | Adult | Male | Pulse<br>Respiration Rate<br>Blinking Eyes |

Table III.2: Victim settings for each round, in order visited.

| Round | Victim | Triage Level | Details |
|---|---|---|---|
| 1 | 1 - Newborn | Immediate | Cries when mouth is opened |
| 1 | 2 - Adult | Immediate | Breathing at 40 bpm |
| 1 | 3 - Child | Expectant | Not Breathing |
| 1 | 4 - Adult | Delayed | Breathing at 20 bpm; Regular Pulse; Responsive |
| 1 | 5 - Toddler | Immediate | Breathing at 18 bpm; Regular Pulse; Not Responsive |
| 1 | 6 - Adult | Delayed | Breathing at 28 bpm; Regular Pulse; Responsive |
| 2 | 1 - Newborn | Expectant | Not Breathing; Not Responsive |
| 2 | 2 - Adult | Delayed | Breathing at 19 bpm; Responsive |
| 2 | 5 - Toddler | Immediate | Breathing at 18 bpm; Regular Pulse; Not Responsive |
| 2 | 6 - Adult | Immediate | Breathing at 28 bpm; No pulse |
| 2 | 4 - Adult | Immediate | Breathing at 11 bpm; Not Responsive |

#### III.1.2.2 Apparatus

All H-H condition participants completed the evaluation prior to the H-R condition participants. During the H-H condition, an evaluator played the role of the first responder and human teammate. A script dictated verbal interactions between the human participant and human experimenter. The same female experimenter was partnered with all participants.

The robot was a Pioneer 3-DX robot equipped with an autonomous navigation system including a laser range finder and a pre-programmed map of the evaluation area and victim locations (Scheutz et al., 2007). The robot spoke using a digital voice through speakers mounted on its platform. Participants donned a wireless microphone headset to amplify their voices when communicating with the robot. The experimenter heard participant questions and responses, and used this knowledge to advance the robot's pre-programmed speech script. When participants asked questions, the experimenter chose from a set of pre-programmed responses or repeated the robot's last statement.

The Guided evaluation used a mixed design with the participants as a random element. The experimental condition, H-H or H-R, differed in a between subjects comparison. The within-subjects element was the series of eleven victim assessments.

#### III.1.2.3 Participants

Twenty-eight participants completed the evaluation. The 14 male and 14 female participants, 14 in each condition, ranged in age between 18 and 57 years old. The H-H participants' mean age was 24.2 (St. Dev. = 10.3) and 26.2 (St. Dev. = 7.8) for H-R participants. Participants rated their experience with first aid and robots as little or no experience with no significant difference between conditions, based on Likert scales from 1 (little or no experience) to 9 (very experienced).

#### III.1.2.4 Metrics

The objective metrics included: physiological data from a portable BioHarness ECG monitor (Biopac Systems, 2013) (heart rate variability, breathing data, R-R data, heart rate, respiration rate, skin temperature, posture, vector magnitude data and acceleration data), subtask time, correctness of responses to the secondary task questions, secondary task failure rate, primary task failure rate, working memory task, accuracy of triage assessments, primary task reaction time, secondary task reaction time, primary task response time, secondary task response time, peak acceleration data, and movement count. Subjective metrics included in situ workload ratings collected after triaging each victim, post-experimental questionnaire responses, and post-experimental NASA-TLX (Hart and Staveland, 1988) responses. The physiological responses presented in this chapter include heart rate variability, heart rate, respiration rate, posture, vector magnitude, and peak

acceleration.

The primary task reaction time is the time required to act immediately on a primary task without extended thought. The moment of a reaction is the time when a first action following the stimulus onset is executed. Reactions can include head nods or using stalling words (e.g., "Uh"). Primary task reaction time is calculated by subtracting the stimulus onset time from the time of the first subsequent action. Primary task reaction time can capture immediate perception of the task prompt, which can be explicit (e.g., audio cue), or tacit (e.g., entering a room), in relation to the processing of the world (e.g., too much stimuli from the environment results in longer reaction time).

The secondary task reaction time is the immediate reaction to a secondary task prompt. Secondary tasks can be explicitly prompted with a cue (e.g., verbal question), or they can be tacit tasks (e.g., check on variable X when there is spare time). Secondary tasks can interrupt the primary task. Secondary tasks, performed constantly throughout the evaluation (e.g., repeating a phoneme), may not be appropriate for measuring reaction time, as there is not a distinct stimulus onset. Secondary task reaction time is calculated by subtracting the stimulus onset time from the time of the first subsequent action. Secondary task reaction time captures the speed that a person can switch from attending to the primary task to perceiving a secondary task cue.

The primary task response time is the time to execute an appropriate response to a primary task prompt. Primary task response time includes reaction time and the time taken to consider an appropriate response or is calculated by adding the time to select a response to the primary task reaction time. Primary task reaction time and primary task response time can be equivalent if the first reaction to the stimulus onset is an appropriate response. It is important to measure primary task response time, because decision-making time in selecting an action can be measured. Verbal tasks are interesting to investigate because they offer insight into the interaction between teammates; a long decision-making time can indicate a lack of knowledge of appropriate responses (i.e., poor communication of goals), or the use of many stalling words or unrelated conversation (e.g., requests to repeat the question).

The secondary task response time is the time taken before executing an appropriate response to a secondary task prompt, and includes the decision-making time to select an appropriate response to the secondary task reaction time. Secondary task response time can be the same as secondary task reaction time if the first reaction to a stimulus is an appropriate response. Similar to primary task response time, secondary task response time offers insight regarding the decision-making process.

Primary task reaction time was determined by recording the stimulus onset time whenever the responder partner finished uttering a primary task prompt (e.g., "What is the victim's breathing rate?"). The time of the reaction was recorded when the participant reacted following the prompt. Secondary task reaction time recorded the stimulus onset time at the end of the secondary task question. The reaction was any

verbal response after the question, including stall words. Primary task response time used the same stimulus onsets as the primary task reaction time. The response times recorded when the participants performed an appropriate response. Verbal questions required an appropriate answer; stall words were not included. Secondary task response time defined the stimulus onset as the end of the secondary task question. The response time was the moment a participant provided an appropriate answer, not including any stalling words. The video coding was completed by two video coders with a Cohen's kappa value of 0.90.

The subtask time represented the time participants spent completing the triage steps for each victim. Timing began when participants arrived at a victim and completed when the responder partner (human or robot) announced the victim's triage level. The timing data serves to ground the physical workload discussion, since longer victim triage times can result in higher physical workload.

Vector magnitude is a measure of overall physical activity (Steele et al., 2000; Johnstone et al., 2012) and combines acceleration from the three axes of movement. Vector magnitude measured participant physical movement throughout the evaluation area. The unit of measure is Vector Magnitude Units (VMU).

Participant posture, recorded via the heart rate monitor, is measured using the degrees off of vertical in any orientation. A positive value indicates the participant leaning forward, while a negative value indicates leaning backwards. The physical workload results present four uses of the posture data: posture skewness, posture kurtosis, postural load and posture variance. A mean raw posture value was calculated for each participant during each victim assessment by computing the mean raw posture value over the total time dedicated to assessing that victim. Skewness and kurtosis tests were performed on this data. Skewness evaluates the asymmetry of a distribution (Groeneveld and Meeden, 1984), in this case, a more positive value indicates that the distribution of postures includes more data points with smaller angled postures. In other words, the participant spent less time leaning forward, and negative skewness values indicate more time spent leaning forward. Kurtosis evaluates the variability of the distribution (Groeneveld and Meeden, 1984), in this case, a higher kurtosis indicates that the participants' postures during each triage assessment are close together. A lower kurtosis indicates that the data points are more dispersed and individual differences are more apparent. Postural load measures the percentage of time participants spent with the flexion of their trunks at an angle of more than 45 degrees from vertical. The longer a participant spent with severe trunk flexion, the higher the physical workload (Paul et al., 1999). The posture variance indicates the change in posture or leaning throughout the victim assessment, similar to Lasley's measure of posture sway (Lasley et al., 1991). The posture variance calculation required averaging the variance of each participant's postural positions during each triage assessment.

Peak acceleration was measured in order to gather information regarding the participants' movement during the evaluation. Peak acceleration refers to the peak, per 1.008 second epoch, for the acceleration

vector sum along the three directions of movement (x, y, and z).

Movement count was determined to be a relevant metric because participants typically crouched near the victim during assessments. The number of times that participants crouched down and stood up was calculated, resulting in the number of times a participant stood, crouched down, and the total number of movement incidences. The video for each participant was coded to obtain the counts. The videos were recorded via two overhead digital cameras with pan and zoom capabilities. Trained technicians manipulated the cameras from the control room, updating camera positions as participants moved between victims. Video was available for 286 (92.9%) of the 308 victim assessments. The missing videos were due to errors in updating camera positions as participants moved throughout the evaluation area.

The secondary recognition task questions were based on a list of five names participants were asked to memorize during the pre-trial briefing: Kathy Johnson, Mark Thompson, Bill Allen, Tammy Hudson and Matt Smith. The names represented a hypothetical team that the participant needed to meet with for debriefing. Participants were given one minute to memorize the list before viewing a briefing video, and another minute to study the list after the briefing video. Thirteen questions incorporating the names were posed throughout the trial. An example question is: "Megan Garner is now setting up the medical treatment site. Was she on the list of names you were given?"

The secondary task failure rate was computed by dividing the number of incorrect responses to secondary task questions by the total number of questions asked. Three of the questions were associated with delayed victims, six of the questions were associated with immediate victims, and two were associated with expectant victims. The remaining two questions were not associated with a victim (i.e., one was asked before triage began, the other was asked between triage rounds), and these two questions are considered non-triage questions.

The Guided evaluation contained four principal components of the primary task with a measurable task failure rate. The primary task components each had a distinct task element: attention, effort, assessment, and detail-tracking (see Chapter IV for more information on primary task failure rate).

The Guided evaluation's four principal primary task components included providing the victim's breathing rate (i.e., victim breathing rate report task), injury status (i.e., victim injury report task), pulse regularity (victim pulse regularity report task), and age (i.e., victim age report task). Each primary task component was repeated a number of times by the participant, dependent on the six victims' associated triage levels during the two rounds of triage.

The victim breathing rate report task was performed eight times. The victim injury report task was performed five times, the sixth victim was expectant and, thus, injuries were not relevant. The victim pulse regularity report task was performed seven times, and the victim age report task was performed six times.

Table III.3: Summary of ground truth for each victim assessment with associated victim breathing rate in breaths per minute, injury status, pulse regularity, and age range.

| Assessment | Victim | Round | Breathing Rate | Injury Status | Pulse Regularity | Age Range |
|---|---|---|---|---|---|---|
| **1** | **1** | **1** | - | None | - | 0-1 years |
| **2** | **2** | **1** | 42 | Broken Thigh | - | 30-40 years |
| **3** | **3** | **1** | - | - | - | 6-8 years |
| **4** | **4** | **1** | 20 | Bruised Shin | Regular | 20-25 years |
| **5** | **5** | **1** | 18 | None | Regular | 1-2 years |
| **6** | **6** | **1** | 28 | None | Regular | 30-40 years |
| **7** | **1** | **2** | - | - | - | - |
| **8** | **2** | **2** | 19 | - | Regular | - |
| **9** | **5** | **2** | 18 | - | Regular | - |
| **10** | **6** | **2** | 28 | - | No Pulse | - |
| **11** | **4** | **2** | 11 | - | Regular | - |

Table III.3 provides information regarding the victim, the triage round, and the associated ground truth of each for the primary task components and associated mannequin settings throughout the evaluation.

The victim breathing rate report primary task failure rate was measured by determining a failure when the participant reported a breathing rate value that was greater than or less than one beat per minute of the ground truth breathing rate of the specified victim. The failure rate was calculated by dividing the total number of misses by eight (i.e., the total number of breathing reports given). A subset of breathing reports were excluded due to mannequin setting errors (i.e., breathing rate set to the incorrect number by the technician); thus, the total number of reports for participants with excluded reports were reduced by the number of exclusions. Specifically, there were 112 victim breathing rate assessments recorded for each condition. Errors in mannequin setup prevented the correct assessment of breathing rate in six cases in the H-H condition and 19 cases in the H-R condition.

The victim injury report primary task failure rate was measured by determining whether each participant reported the correct injury status of each victim during the first round of triage assessments (i.e., five reports total; the expectant victim did not require an injury report). For example, if the victim was uninjured but the participant reported an injury, the report was considered incorrect. The failure rate was calculated by dividing the number of misses for a participant by five. No victim injury report task data were excluded due to experimenter errors.

The victim pulse regularity report primary task failure rate was measured by determining whether or not each participant reported the correct pulse regularity (i.e., regular pulse rate, irregular pulse rate, no pulse rate detectable) when assessing the victim's pulse. The failure rate was calculated by dividing the number of misses by the number of times the participant assessed a victim's pulse regularity. There were 98 victim pulse regularity assessments in each condition. Errors in mannequin setup (i.e., pulse rate turned off) prevented correct pulse regularity assessment for 14 H-H condition assessments and 11 H-R condition assessments. A

subset of 11 additional pulse regularity assessments in the H-H condition were not considered due to an error in video recording, as pulse regularity assessment failure rate was determined via video coding.

The victim age report primary task failure rate was measured by assessing whether or not participants reported the age of each victim within the victims' ground truth age range. If the age reported by participants did not fit within the mannequin's ground truth range, the report was considered a failure. The failure rate was calculated by dividing the number of misses by six (i.e., the number of times a participant reported a victim's age). No victim age report task data were excluded due to experimenter errors.

The overall primary task failure rate was calculated by computing a failure rate for each participant composed of a sum of all failures from each of the four task components (i.e., victim breathing rate report task, victim injury report task, victim pulse regularity report task, and victim age report task) divided by the number of total reports given by the individual participant. The total number of reports sometimes varied by participant due to the described mannequin errors.

The victim breathing rate report task represents an attention task, because it requires participants to maintain direct observation and awareness of the breathing rate during the timed task. The participant must accurately hold the breath count in his or her mind during the task in order to be correct. The victim injury report task is an effort-sustaining task, because the participant extend his or her own effort in order to notice an injury on the victims' bodies, or the lack thereof. The pulse regularity report task represents an assessment task, during which the victim pulse rate is assessed to be "regular," "irregular," or "not present." The victim age report task is a detail-tracking task that required participants to use details present in the environment (i.e., the victims' clothes and sizes) to estimate the victim's age.

The Guided evaluation tasked participants to recall as many names from the memorized a list of first responders provided for the secondary task. After finishing the eleven triage assessments, participants were asked to provide any of the first responder names from the original list. Experimenter error prevented all of the responses from being gathered. Only seven participants in H-R condition and ten in H-H condition reported the names recalled upon evaluation completion. Instances existed where participants reported partially correct names (e.g., correct last name with an incorrect first name), and these instances were scored a half credit (i.e., each first and last name were considered separately as a failure).

After completing the triage steps for a particular victim, the participants completed the in situ workload ratings by rating six workload channels on a scale from 1 (little to no demand) to 5 (extreme demand). The six workload channels were Cognitive, Auditory, Visual, Tactile, Motor and Speech. Each channel was defined during the first set of questions. The questions were adapted from the Multiple Resources Questionnaire (Boles et al., 2007) and the channels were chosen to facilitate comparison to IMPRINT Pro's seven workload channels. In order to prevent confusion, Imprint Pro's fine and gross motor channels were combined into a

Motor channel rating. When comparing the results to the predicted values, the fine and gross motor channels were added together. The provided workload responses were normalized to the corresponding IMPRINT Pro scale to facilitate comparison. The total workload for each victim assessment for the models was calculated using a time-weighted average of all workload values experienced while assessing the particular victim. These totals were compared directly to the re-scaled in-task subjective overall workload results. The motor and tactile workload channels are the focus for evaluating physical workload.

The post trial questionnaire included eight statements. The participants rated their level of agreement with the statements on a Likert scale from 1 to 9, where 1 meant completely disagree and 9 meant that the participant completely agreed. The statements consisted of the following:

1. My teammate gave me clear instructions.
2. I trusted my teammate.
3. I felt comfortable communicating with my teammate.
4. My teammate understood what I was trying to communicate.
5. I did a good job on the tasks I was assigned.
6. I often felt confused about my teammate's instructions.
7. I often felt confused as to the purpose of my actions.
8. I felt stressed during the scenario.

The NASA-TLX questionnaire was completed at the end of the entire evaluation (Hart and Staveland, 1988). The overall workload results will present the overall NASA-TLX response score. Specific physical workload comparisons are derived from the Physical workload component.

#### III.1.2.5 Procedure

After donning the Bioharness heart rate monitor, participants viewed a four-minute primer video intended to set a scene of a mass-casualty incident. The video was comprised of scenes from David Vogler's live footage from the September 11th attacks in New York City (Vogler, 2001).

Participants completed two triage rounds. During the first round, participants visited each mannequin, of which three were to be classified as immediate (Victims 1, 2, and 5), two delayed (Victims 4 and 6) and one expectant (Victim 3). During the second round, the participants only visited the victims classified as immediate or delayed during the first round. During the second round, three victims were classified as

immediate (Victims 4, 5, and 6), one delayed (Victim 2) and one expectant (Victim 1). The human and robotic partner led participants through the triage steps for the eleven victim assessments. Upon completing a triage assessment, the participant verbally assessed the subjective workload channels via in situ subjective workload ratings. After the scenario completion, the participant completed a post-trial questionnaire and the NASA-TLX workload questionnaires.

### III.1.3 Results

Overall and mental workload were analyzed in Chapter III.1.3.1, physical workload was analyzed in Chapter III.1.3.2 and reaction and response time were analyzed in Chapter III.1.3.3.

#### III.1.3.1 Overall and Mental Workload Results

**Physiological Results**

The low frequency heart rate variability was analyzed by victim, triage level and condition. The overall H-H mean was 13.01 (St. Dev. = 12.56) and the H-R mean was 12.70 (St. Dev. = 14.77) $ms^2$. A t-test found no significant difference across conditions for overall heart rate variability. Mean and standard deviation (St. Dev.) for heart rate variability, heart rate and respiration rate by triage level and condition are provide in Table III.4.

A two-way ANOVA assessed the main effects and interaction of both condition and triage level, with heart rate variability as the dependent variable and both condition and triage level as independent variables. Results showed a significant main effect of triage level on heart rate variability, with $F(302,2) = 13.25$, $p < 0.01$. There was no main effect of condition on heart rate variability or an interaction effect of triage level and condition. A Tukey HSD post-hoc test indicated that all three triage levels had significantly different heart rate variability. Delayed victims elicited higher heart rate variability than both Immediate ($p = 0.01$) and Expectant ($p < 0.01$) victims. Immediate victims had higher heart rate variability than Expectant victims ($p < 0.01$).

The heart rate descriptive statistics by condition and triage level are provided in Table III.4. A t-test found that the H-H condition had significantly higher heart rate, $t(306) = 3.59$, $p < 0.01$. A two-way ANOVA assessed the main effects and interaction of both condition and triage level, with heart rate as the dependent variable and both condition and triage level as independent variables. Results showed that H-H heart rate was significantly higher than that of the H-R condition, with $F(302,1) = 12.78$, $p < 0.01$. There was no main effect of triage level and no interaction effect of triage level and condition.

The respiration rate descriptive statistics, by condition and triage level are also provided in Table III.4. A t-test found that the H-H condition had a significantly higher mean respiration rate, $t(306) = 2.65$, $p = 0.01$. A

two-way ANOVA assessed the main effects and interaction of both condition and triage level, with respiration rate as the dependent variable and both condition and triage level as independent variables. Results showed that the H- H respiration rate was higher than H-R, $F(302, 1) = 6.93$, $p < 0.01$. No main effect of triage level on respiration rate or interaction effect of triage level and condition were found.

**Peak Acceleration**

The peak acceleration descriptive statistics, by condition and triage level are provided in Table III.4. A t-test showed that the H-R condition had a significantly higher peak acceleration than the H-H condition, $t(306) = 5.47$, $p < 0.01$. A two-way ANOVA assessed the main effects and interaction of both condition and triage level, with peak acceleration as the dependent variable and both condition and triage level as independent variables. There was no significant interaction effect of condition and triage level. There was a main effect of condition, showing that the H-R condition had a significantly higher peak acceleration than the H-H condition, $F(288, 1) = 28.69$, $p < 0.01$. There was also a main effect of triage level, with $F(288, 2) = 4.786$, $p < 0.01$. A Tukey HSD post-hoc test showed that peak acceleration for Expectant victims was higher than for Immediate victims, $p = 0.01$. These results may indicate that the participants in the H-R condition had quicker movement, but since this is peak acceleration and not average acceleration, that cannot be stated as definite. The individual channels of acceleration (e.g., X-axis peak acceleration) was examined individually, but did not provide useful data for this dissertation, as participants were not restricted to specific axes of movement during specific time periods; thus, comparisons on single axes of movement are not relevant.

**Secondary Task Questions Correctness**

The number of correct answers to secondary task questions was compared between conditions. Thirteen questions (Q.) were asked in total one during the introduction to the task (Q. 1), one during the triage of each the victim during Round 1 (Q. 2 7), one between the two rounds (Q. 8), and one during the triage of each victim during the second round (Q. 9-13). Overall, the mean number of correct responses was 12.71 (St. Dev. = 0.61) during the H-H condition and 12.43 (St. Dev. = 0.65) for the H- R condition. T-tests across conditions found no significant difference. Analysis was conducted based upon triage level without any significant results. The division of correct answers by condition is provided in Table III.5.

**Secondary Task Failure Rate**

The Guided evaluation secondary task failure rate responses were analyzed by condition and triage level. Non-parametric analysis was used because the results were not normally distributed. The mean secondary task failure rate for each participant in the H-H condition was 2.20% (St.Dev. = 4.53%) and was 4.40%

Table III.4: Descriptive statistics for physiological metrics in the Guided evaluation.

| Metric | Triage Level | Statistic | H-H | H-R |
|---|---|---|---|---|
| **Heart Rate Variability** | Delayed | Mean | 17.50 | 18.09 |
| | | St. Dev. | 15.38 | 18.43 |
| | Immediate | Mean | 12.57 | 12.71 |
| | | St. Dev. | 10.82 | 13.81 |
| | Expectant | Mean | 7.61 | 4.56 |
| | | St. Dev. | 10.58 | 4.44 |
| | Overall | Mean | 13.01 | 12.70 |
| | | St. Dev. | 12.56 | 14.77 |
| **Heart Rate (beats per minute)** | Delayed | Mean | 85.03 | 81.13 |
| | | St. Dev. | 12.21 | 9.40 |
| | Immediate | Mean | 85.20 | 80.42 |
| | | St. Dev. | 12.00 | 10.27 |
| | Expectant | Mean | 87.36 | 82.31 |
| | | St. Dev. | 12.65 | 11.45 |
| | Overall | Mean | 85.55 | 80.96 |
| | | St. Dev. | 12.13 | 10.22 |
| **Respiration Rate (breaths per minute)** | Delayed | Mean | 18.88 | 18.01 |
| | | St. Dev. | 3.69 | 2.64 |
| | Immediate | Mean | 19.13 | 17.71 |
| | | St. Dev. | 3.91 | 3.25 |
| | Expectant | Mean | 19.01 | 18.69 |
| | | St. Dev. | 4.50 | 3.25 |
| | Overall | Mean | 19.04 | 17.97 |
| | | St. Dev. | 3.94 | 3.10 |
| **Peak Acceleration (gravitational force)** | Delayed | Mean | 0.11 | 0.16 |
| | | St. Dev. | 0.03 | 0.04 |
| | Immediate | Mean | 0.10 | 0.13 |
| | | St. Dev. | 0.05 | 0.07 |
| | Expectant | Mean | 0.11 | 0.16 |
| | | St. Dev. | 0.04 | 0.09 |
| | Overall | Mean | 0.11 | 0.15 |
| | | St. Dev. | 0.04 | 0.06 |

Table III.5: Average number of correct secondary task question responses, by triage level in the Guided evaluation.

| Triage Level | Statistic | H-H | H-R |
|---|---|---|---|
| **Delayed** | Mean | 2.93 | 2.86 |
| Q. 5, 7, 10 | St. Dev. | 0.27 | 0.36 |
| **Immediate** | Mean | 2.93 | 2.86 |
| Q. 2, 3, 6, 11, 12, 13 | St. Dev. | 0.27 | 0.36 |
| **Expectant** | Mean | 1.93 | 1.93 |
| Q. 4, 9 | St. Dev. | 0.27 | 0.27 |
| **Overall** | Mean | 12.71 | 12.43 |
| | St. Dev. | 0.61 | 0.65 |

Table III.6: Guided evaluation secondary task failure rate by condition and triage level

| Triage Level | H-H | H-R |
|---|---|---|
| Delayed | 1/42 | 2/42 |
| Immediate | 2/84 | 3/84 |
| Expectant | 1/28 | 1/28 |
| Non-triage | 0/28 | 2/28 |
| Total | 4/182 | 8/182 |

Table III.7: Mean reported age by condition and victim number, in years.

| Victim Number | H-H | H-R | Ground Truth |
|---|---|---|---|
| 2 | 33.21 (3.72) | 33.64 (6.21) | Adult, age 30-40 |
| 3 | 6.32 (1.73) | 7.50 (1.87) | Child, age 6-8 |
| 4 | 23.89 (6.57) | 22.93 (4.45) | Adult, age 20-25 |
| 5 | 0.92 (0.53) | 1.32 (0.97) | Toddler, age 1 |
| 6 | 31.46 (4.64) | 33.29 (6.34) | Adult, age 30-40 |

(St.Dev. = 4.79%) for participants in the H-R condition. Table III.6 presents the breakdown of incorrect responses to secondary task questions out of the total number questions asked by triage level and condition. A Kruskal Wallis test indicated no significant difference between conditions or triage levels for the secondary task failure rate.

**Task Performance**

The Guided evaluation required participants to report each victim's age and the breathing rate during 9 of the 11 triage assessments. The consistency of these reports represents task performance. The reported victim ages were only analyzed between the H-H and H-R conditions for Victims 2 through 6. The mean reported victim ages by condition are presented in Table III.7. T-tests between conditions found no significant differences.

The reported victim breathing rates were also compared by victim using t-tests between conditions. The mean reported breathing rates by condition are presented in Table III.8. Four participant breathing rate reports were omitted from the H-R condition for Assessment 2 and six participant H-H condition breathing rate reports from Assessment 9 were omitted due to an improper setting of the mannequins' breathing rates. Victim Assessment 8 resulted in the only significant difference between conditions. Participants in the H-H condition reported the victim's breathing as significantly higher than the H-R participants, $t(25) = 2.50$, $p = 0.02$. Participants from both conditions reported breathing rates very close to the ground truth values.

**Primary Task Failure Rate**

The means for each primary task failure rate component for the Guided evaluation are provided in Figure III.9. The overall primary task failure rate mean is also provided. A series of Shapiro-Wilks tests indicated

Table III.8: Mean reported respiration rate by condition and assessment number in breaths per minute.

| Assessment Number (of 11) | H-H | H-R | Ground Truth |
|---|---|---|---|
| 2 | 38.93 (6.73) | 38.86 (0.47) | 38 |
| 4 | 20.79 (0.97) | 20.80 (1.03) | 20 |
| 5 | 21.08 (2.23) | 19.44 (0.73) | 18 |
| 6 | 27.00 (2.91) | 26.00 (1.29) | 28 |
| 8 | 20.57 (1.45) | 18.15 (3.29) | 19 |
| 9 | 19.63 (2.64) | 18.71 (2.55) | 18 |
| 10 | 27.57 (1.50) | 28.86 (9.69) | 28 |
| 11 | 14.38 (8.68) | 11.89 (1.69) | 11 |

that all of the collected data were not normally distributed; thus, non-parametric Kruskal-Wallis tests are used. The mean attention primary task failure rate for the H-H condition was 41.96% (St.Dev. = 16.91%) and was 47.81% (St.Dev. = 23.85%) in the H-R condition. A Kruskal Wallis test indicated no significant difference between conditions. The mean effort primary task failure rate in the H-H condition was 3.21% (St.Dev. = 7.93%) and was 17.86% (St.Dev. = 14.23%) in the H-R condition. A Kruskal Wallis test indicated that the H-R condition participants had a significantly higher injury report primary task failure rate $\chi^2(1) = 7.88$, $p = 0.005$. The mean assessment report primary task failure rate in the H-H condition was 16.74% (St.Dev. = 18.52%) and was 28.88% (St.Dev. = 17.75%) in the H-R condition. A Kruskal Wallis test indicated that the H-R condition participants had a nearly-significantly higher pulse regularity report primary task failure rate, $\chi^2(1) = 2.63$, $p = 0.105$. The detail-tracking primary task failure rate for the H-H condition was 26.43% (St.Dev. = 17.39%) and was 27.38% (St.Dev. = 24.89%) in the H-R condition. A Kruskal Wallis test indicated no significant difference between conditions. The mean overall primary task failure rate for the H-H condition participants was 25.50% (St.Dev. = 10.25%) and was 30.86% (St.Dev. = 10.40%). A Kruskal Wallis test indicated no significant difference between conditions.

**Memory Recall**

Guided evaluation memory recall responses were analyzed by condition using non-parametric analysis, due to the lack of non-normally distributed values. The mean memory recall percentage for participants in the H-H condition was 86.0% (St.Dev. = 20.7%) and was 92.9% (St.Dev. = 7.6%) for the H-R participants. A Kruskal Wallis test indicated no significant difference between conditions.

**In Situ Subjective Workload Ratings**

The individual channel subjective workload ratings gathered at the completion of each victim assessment were combined into a total workload value. The overall mean for H-H workload was 16.26 (St. Dev. = 6.31), while the H-R workload mean was 13.48 (St. Dev. = 4.80). A t-test indicated that the H-H condition rated

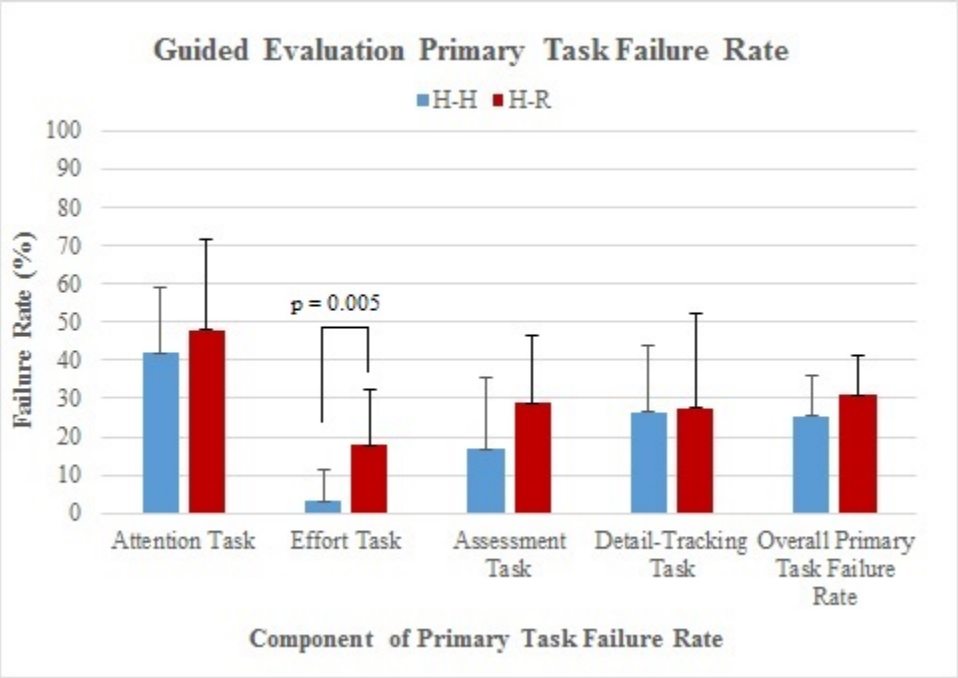



Figure III.9: Means for the four components of the Guided evaluation primary task failure rate with overall task failure rate, by condition. Error bars represent one standard deviation above the mean. Significant differences are represented with brackets and associated p-values.

workload higher, $t(306) = 4.35$, $p < 0.01$. The mean H-H condition workload for Delayed victims was 19.31 (St. Dev. = 5.94) and 15.02 (St. Dev. = 4.70) for the H-R condition. The mean for the H-H Immediate victims, was 16.06 (St. Dev. = 5.89) and for H-R was 13.48 (St. Dev. = 4.60). The Expectant victims H-H mean was 15.21 (St. Dev. = 5.97) and the H-R mean was 11.14 (St. Dev. = 4.74). Figure III.10 provides the total workload for each condition at each victim assessment point.

A two-way ANOVA assessed the main effects and interaction of both condition and triage level, with the total in-task subjective workload ratings as the dependent variable and both condition and triage level as independent variables. Results showed a significant main effect of triage level on workload ratings, with $F(302,2) = 10.29$, $p < 0.01$. There was a main effect of condition on the workload ratings with $F(302,1) = 29.86$, $p < 0.01$, showing that the H-H workload ratings were significantly higher than the H-R workload ratings. There was no interaction effect of triage level and condition. A Tukey HSD test showed that the significant difference between triage levels was due to the Delayed victims being rated significantly higher than both the Immediate ($p < 0.01$) and Expectant ($p < 0.01$) victims. There was no significant difference between the Expectant and Immediate workload ratings.

Figure III.10: Overall workload by victim and condition.

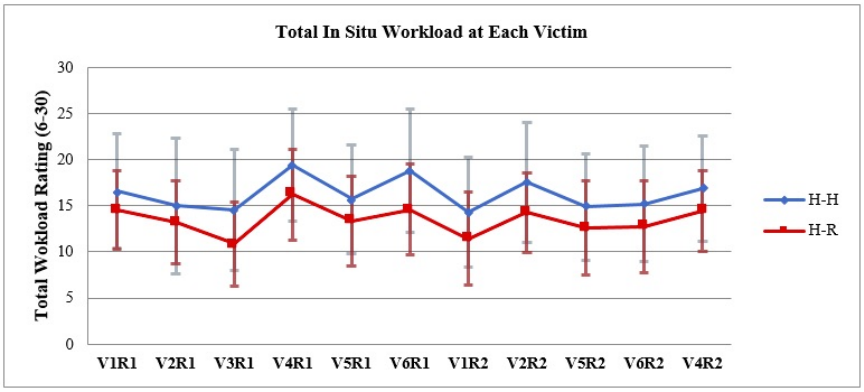


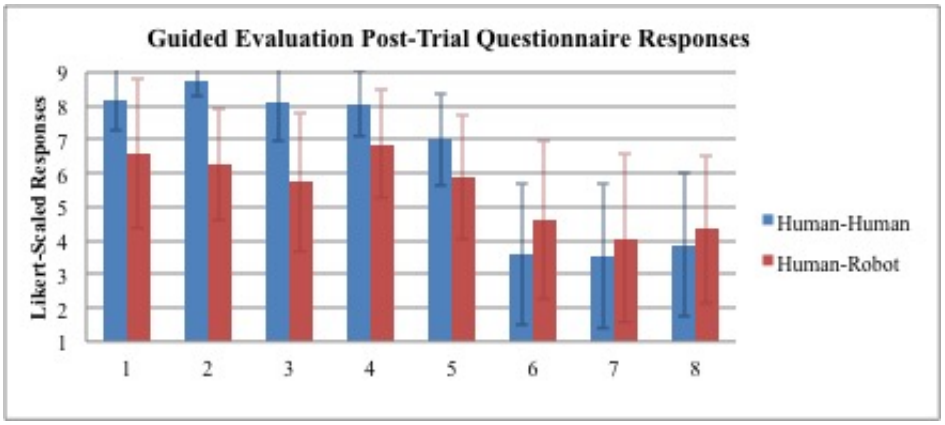

Table III.9: Descriptive statistics for Guided evaluation post-trial questionnaire data.

| Statement Number | Statement | H-H | | H-R | |
|---|---|---|---|---|---|
| | | Median | Range | Median | Range |
| 1 | My teammate gave me clear instructions. | 8.5 | 6-9 | 6.5 | 2-9 |
| 2 | I trusted my teammate. | 9 | 8-9 | 6.5 | 4-9 |
| 3 | I felt comfortable communicating with my teammate. | 9 | 6-9 | 6 | 1-9 |
| 4 | My teammate understood what I was trying to communicate. | 8 | 6-9 | 7 | 3-9 |
| 5 | I did a good job on the tasks I was assigned. | 7 | 5-9 | 6.5 | 2-8 |
| 6 | I often felt confused about my teammate's instructions. | 3 | 1-7 | 4.5 | 1-8 |
| 7 | I often felt confused as to the purpose of my actions. | 2.5 | 1-7 | 3 | 1-9 |
| 8 | I felt stressed during the scenario. | 4 | 1-8 | 4.5 | 1-9 |

**Post-Trial Questionnaire Responses**

The median responses from questions from the post-trial questionnaire (see Chapter III.1.2.4) are presented in Table III.9. A Kruskal-Wallis test indicated that H-R participants provided significantly lower agreement with the statements overall, $\chi^2(1) = 10.65, p = 0.001$. There was a main effect of question number ($\chi^2(7) = 83.65, p < 0.001$) and an interaction effect of condition and question number ($\chi^2(15) = 112.42), p < 0.001$). Post-hoc pairwise Wilcoxon tests were performed with Bonferonni corrections for family-wise error and indicated that Statement 2 was significantly higher in the H-H condition ($p = 0.01$). All mean results of the post-trial questionnaire are seen in Figure III.11.

Figure III.11: Mean Guided evaluation post-trial questionnaire responses. Post-trial questionnaire statements are provided in Table III.9.




**NASA-TLX Responses**

Each participant completed the NASA-TLX questionnaire. The mean overall weighted score for the H-H condition was 57.38 (St. Dev. = 14.00), while the mean for the H-R condition was 48.59 (St. Dev. = 11.98). A t-test found no significant difference between the overall scores. While this result is not significant, it indicates a trend that those in the H-H condition tended to rate their overall workload values slightly higher than the H-R condition participants.

**Correlations Analysis**

A partial Pearson's correlation was performed to analyze the correlation between heart rate variability, heart rate, respiration rate and the total in-task subjective workload rating from each victim while adjusting for the independent variables of victim being assessed and victim triage level. Across both conditions, in-task subjective workload ratings were significantly negatively correlated to respiration rate, $r(290) = -0.15$, $p = 0.01$ and had a significant positive correlation to heart rate, $r(290) = 0.16$, $p < 0.01$. The correlation between heart rate variability and subjective workload ratings was nearly significant with $r(290) = 0.10$, $p = 0.10$. The literature (Reiser and Schlenk, 2009; Vicente et al., 1987; Aasman et al., 1987; Roscoe, 1992) implies that these three physiological measures may be able to represent workload. Since the physiological metrics were correlated to the in-task subjective workload ratings, the trends shown by these three physiological measures can be considered when assessing the difference in workload between conditions. The literature also reports a positive correlation between both heart rate variability and heart rate and workload, and a negative correlation between respiration rate and workload.

Figure III.12: Overall workload: H-H model and H-H condition.

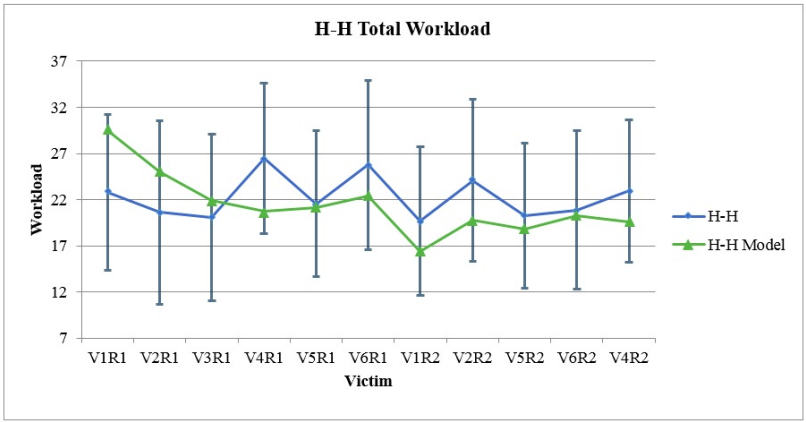
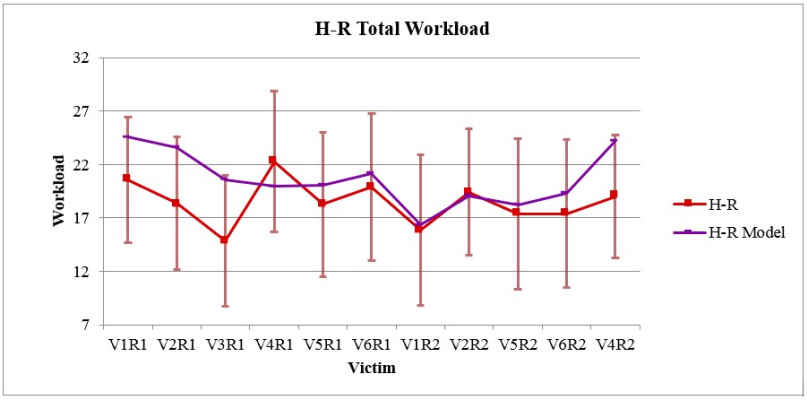



Figure III.13: Overall workload: H-R model and H-R condition.




**Model Analysis**

The participants' in-task subjective workload ratings and the average workload for each victim predicted by the IMPRINT Pro model were compared. Model total workload was calculated by adding together each workload channel. The empirical total workload was calculated by rescaling each of the channel results to the IMPRINT Pro's workload channel scales and then totaling the results at each time point. Figure III.12 compares the H-H condition results compared to the H-H model workload results. Figure III.13 provides the results for the H-R condition. As can be seen in the figures, the majority of the empirical workload data points by victim and round were higher than the predicted model values. The calculated difference between the modeled workload values and the in-task subjective workload values demonstrate how effective the models were at predicting human behavior. The average difference between the H-H subjective values and the model results at each time point was 3.23 (St. Dev. = 2.03). The mean delta between the H-R condition and the model was 2.64 (St. Dev. = 2.01). These data imply that the H-R model was slightly closer to empirical results than the H-H model.

Table III.10: The mean subtask time by condition and triage level. Table I and following tables provide the mean with standard deviation in parenthesis, unless otherwise noted.

| Triage Level | Subtask Time | |
|---|---|---|
| | H-H | H-R |
| Delayed | 190.92 (31.69) | 202.07 (24.02) |
| Immediate | 129.61 (54.75) | 150.69 (63.15) |
| Expectant | 114.92 (67.25) | 119.49 (85.99) |
| Overall | 136.42 (60.92) | 151.53 (70.73) |

#### III.1.3.2 Physical Workload Results

The presentation of the results begins with the subtask time results, which are presented to ground the physical workload discussion. These results are followed by the objective and subjective results, respectively. Finally, comparisons of these results to the computational models are provided. Cohen's d statistics are presented to represent the strength of the claim that significant results are not the product of simply chance. Larger effect sizes help to emphasize the strength of the significant results.

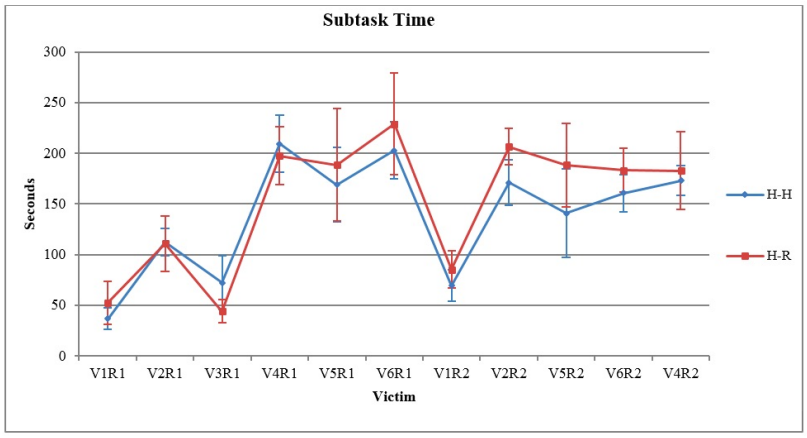
**Subtask Time**

A two-way ANOVA assessed the effects of independent variables triage level and condition on the dependent variable, subtask time. The H-R condition took significantly longer when triaging each victim, $F(288,1) = 4.51, p = 0.04$. A Cohen's d statistic indicated that the effect size of this difference was 0.2, or a small effect size (Cohen, 1988). This small effect size implies the difference in subtask time is significant between conditions, but the small effect size implies that the distributions of the subtask times do have overlap. Additionally, a main effect of triage level existed, $F(288,2) = 27.83, p < 0.01$. A Tukey HSD test determined that delayed victims (V4R1, V6R1, V2R2) required significantly longer subtask times than both immediate (V1R1, V2R1, V5R1, V4R1, V5R1, V6R1; $p < 0.01$) and expectant (V3R1, V1R2; $p < 0.01$) victims. The immediate victims also took significantly longer than expectant victims, $p = 0.02$. Table III.10 provides the descriptive statistics by condition and triage level for subtask time. Figure III.14 depicts the average subtask time by condition and it is apparent that both conditions followed a similar trend. The H-R participants had a longer subtask time for eight of the eleven victims.

**Other Objective Measures**

The descriptive statistics for vector magnitude, postural load, and posture variance are presented by triage level and condition in Table III.11. Vector magnitude is presented as raw values $x10^3$ for ease of reading. Postural load is presented as a percentage of time, while variance in posture is presented as units of degrees squared. Total movement count is presented in total number of times the participant stood up and crouched

Figure III.14: Mean subtask time for each victim assessment by condition.




down.

The H-R condition participants had a higher mean vector magnitude for all victim assessments. The mean vector magnitude for the H-H condition was 9.23 (St. Dev. = 7.10) $VMUx10^{-3}$ and was 15.67 (St. Dev. = 10.68) $VMUx10^{-3}$ for H-R condition. A Pearson's product-moment correlation of the subtask time and vector magnitude resulted in significant correlations between each condition's vector magnitude and the time spent triaging each victim (H-H: $r(139) = 0.29, t = 3.55, p < 0.01$ and H-R: $r(151) = 0.22, t = 2.79, p < 0.01$). These results show that the longer subtask time, the larger the resultant vector magnitude and the more participants moved. Figure III.15 depicts the vector magnitude for each victim assessment by each condition, and while the correlation is not extremely large, the trends are similar to those for subtask time; the values are lower for the first three victims and the first victim in the second triage round with higher values for victims 4, 5 and 6 in the first triage round.

Participants in the H-R condition had a larger postural load for the last seven victim assessments. The mean postural load for H-H condition participants was 16.89 (St. Dev. = 25.82), while the mean postural load for H-R condition participants was 21.00 (St. Dev. = 24.13). Figure III.16 presents the mean postural load for each condition by victim assessment and triage round. A Pearson's product-moment correlation evaluated the relationship between subtask time and postural load. There was no significant correlation for the H-H condition; however, there was a significant correlation in the H-R condition, $r(151) = 0.29, t = 3.78, p < 0.01$. Figure III.16 shows that trends in the two measures are comparable, starting with the fourth victim in the first round. The analysis shows that longer triage times in the H-R condition resulted in larger postural load.

The mean variance in posture for H-H condition participants was 396.74 (St. Dev. = 413.68) and was 792.29 (St. Dev. = 611.66) for H-R condition participants. Even though the H-R condition means are higher, they follow a similar trend as the H-H condition, as shown in Figure III.17. A Pearson's product-moment

correlation evaluated the relationship between subtask time and variance in posture. A significant correlation existed for both conditions; the H-H condition, $r(139) = 0.23, t = 2.83, p = 0.01$ and H-R condition, $r(151) = 0.53, t = 7.71, p < 0.01$. This result indicates that a longer triage assessment time resulted in higher variance in posture, independent of condition.

The overall mean total movement count during each victim assessment for the H-H condition was 1.75 (St. Dev. = 1.40) moves, while the overall mean for the H-R condition was 2.22 (St. Dev. = 1.98) moves. Figure III.18 presents the mean total movement count at each victim assessment for the conditions. A Pearson's product-moment correlation analyzed the relationship between subtask time and total movement count during each assessment. The H-H condition total movement count and time spent triaging each victim were significantly correlated, $r(135) = 0.48, t = 6.33, p < 0.01$, as were the results for the H-R condition, $r(147) = 0.32, t = 4.15, p < 0.01$. These results indicate that participants tended to have a higher total movement count for victim assessments that required longer subtask time.

A two-way MANOVA examined the effect of independent variables triage level and condition on total movement count and the three physiological dependent variables: vector magnitude, postural load, and variance in posture. The test revealed a significant main effect of condition, Wilks' $\lambda = 0.84$, $F(4, 282) = 13.17, p < 0.01$. An analysis of the univariate main effects found the H-R condition total movement count ($F(1, 280) = 5.76, p = 0.02$), vector magnitude ($F(1, 280) = 35.91, p < 0.01$) and variance in posture ($F(1, 280) = 40.48, p < 0.01$) were significantly higher than the H-H condition. Cohen's d statistics indicated that there was a large effect size of the difference between conditions variance in posture (d = 0.8), a medium effect size of vector magnitude (d = 0.7), and a small effect size of postural load (d = 0.2) and movement count (d = 0.3). The large effect size for variance in posture implies that 47.4% of the H-H and H-R condition distributions do not overlap. The medium effect size in vector magnitude has 43.0% non-overlap, while the small effect sizes show 14.7% and 21.3% non-overlap of distributions between conditions, respectively.

A significant main effect of triage level for both conditions on the four measures of physical workload was also present, Wilks' $\lambda = 0.84, F(4, 282) = 12.89, p < 0.01$. The significant effect was present for both conditions for total movement count ($F(2, 280) = 18.89, p < 0.01$), vector magnitude ($F(2, 280) = 3.31, p = 0.04$) and variance in posture, ($F(2, 280) = 8.12, p < 0.01$). A Tukey HSD post-hoc test was performed and determined that the vector magnitude elicited by delayed victim assessments was significantly higher than immediate victim assessments (p = 0.03) with a small effect size (d = 0.3) showing 21.3% non-overlap of distributions. In addition, the elicited variance in posture was significantly higher for delayed victims than both immediate (p = 0.01) victims, with a small effect size (d = 0.4), and expectant ($p < 0.01$) victims, with a medium effect size (d = 0.7). The Tukey HSD post-hoc test for total movement counts showed that all three triage levels were significantly different from one another. Delayed victims elicited a significantly higher

total movement count than both immediate victims ($p < 0.01$), with a medium effect size (d = 0.6), and expectant victims ($p < 0.01$), with a large effect size (d = 1.1). Immediate victims had a significantly higher total movement count than expectant victims (p = 0.03), showing a medium effect size (d = 0.5).

The MANOVA results demonstrate that the physiological measures and total movement counts were significantly affected by condition, indicating that physical workload was higher in the H-R condition. The physiological measures and movement counts were also affected by triage level, which demonstrates an experimental manipulation of overall workload and, in turn, physical workload.

Posture skewness for the H-H condition was 0.32, while overall skewness for the H-R condition was -0.01. These values imply that the H-H condition was skewed to the left and the distribution was filled with mostly smaller posture numbers. The close to zero H-R result indicates that the posture values were more evenly distributed. Skewness was calculated for each participant using the mean raw postures from each victim assessment in order to evaluate the difference in skewness. Mean skewness for the H-H condition was -0.28 (St. Dev. = 0.65) and mean skewness for the H-R condition was -0.23 (St. Dev. = 0.42). A t-test found no significant difference. These results indicate that the conditions are not significantly different in relation to mean posture skewness.

The overall posture kurtosis value for the H-H condition was 0.14, while the overall kurtosis value for the H-R condition was -0.71. These values imply that the H-R condition had a wider distribution of participant raw posture values. A new kurtosis value was calculated for each participant using the mean raw postures from each victim assessment in order to evaluate the kurtosis differences by condition. The mean kurtosis value for the H-H condition was -0.61 (St. Dev. = 0.68) and the H-R condition mean kurtosis value was -1.12, (St. Dev. = 0.43). A t-test indicated that the H-R condition had a significantly lower kurtosis mean than the H-H condition, $t(26) = 2.37$, $p = 0.03$, which implies that the H-R condition had a significantly more dispersed distribution of mean raw posture data for each participant. The Cohen's d statistic indicated a large effect size of condition on posture kurtosis (d = 0.9), implying that 51.6% of the distribution of kurtosis values between conditions did not overlap. The overall kurtosis value and the significantly lower H-R mean kurtosis value indicates that the distribution of postures was wider for the H-R condition.

**Subjective Measures**

The median subjective workload ratings for the motor and tactile channels are presented in Table III.12 by condition and victim assessed. The Likert-scaled subjective workload rating data is not normally distributed, therefore nonparametric analysis is used.

The median motor subjective workload ratings were the same or higher for every victim assessment in the H-H condition except for the first victim in the first round (V1R1 in Table III.12). The median motor subjec-

Figure III.15: Mean vector magnitude for each victim assessment by condition.

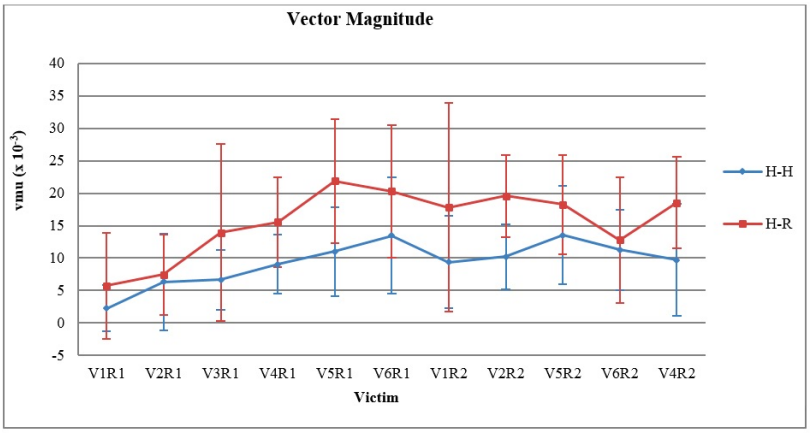



Figure III.16: Mean postural load for each victim assessment by condition.

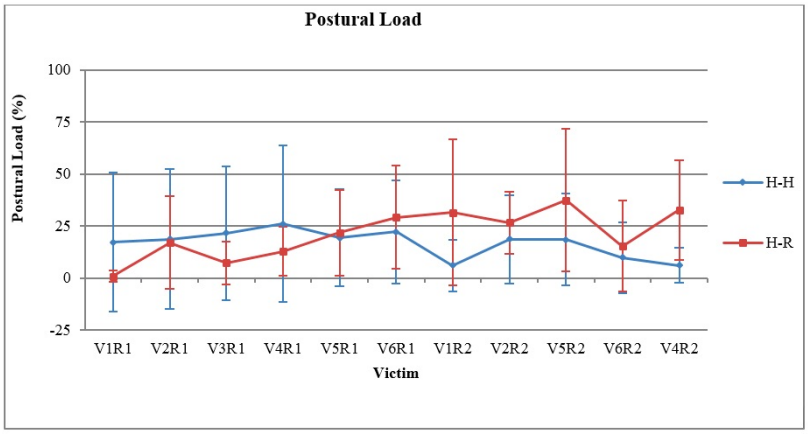



Figure III.17: Mean variance in posture for each victim assessment by condition.

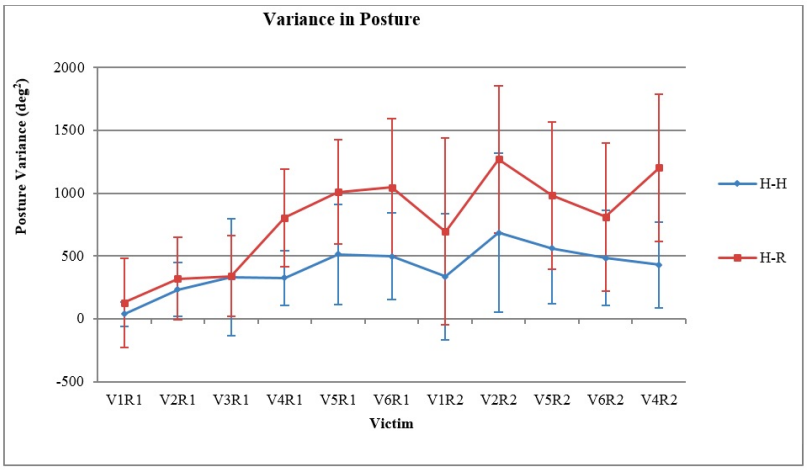


Figure III.18: Mean total movement count for each victim assessment by condition.

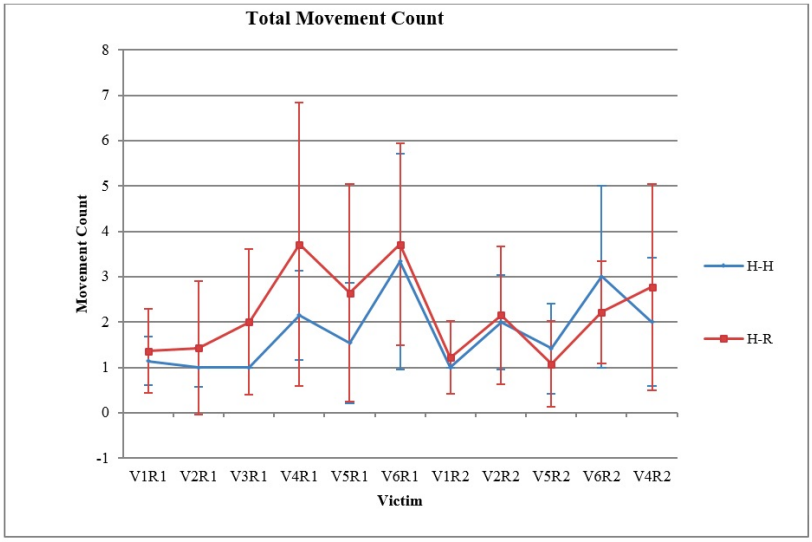



Table III.11: Descriptive statistics for vector magnitude, postural load, posture variance and total movement count by condition and triage level.

| | | Triage Level | | | Overall |
|---|---|---|---|---|---|
| | | **Delayed** | **Immediate** | **Expectant** | |
| **Vector Magnitude** | H-H | 10.85 (6.52) | 8.84 (7.64) | 8.02 (6.03) | 9.23 (7.10) |
| | H-R | 18.46 (8.13) | 14.17 (9.97) | 15.87 (14.77) | 15.67 (10.68) |
| **Postural Load** | H-H | 22.46 (28.35) | 15.18 (24.65) | 13.81 (25.11) | 16.89 (25.82) |
| | H-R | 22.83 (19.25) | 20.59 (25.07) | 19.49 (28.23) | 21.00 (24.13) |
| **Variance in Posture** | H-H | 498.19 (444.30) | 368.41 (370.01) | 333.03 (474.42) | 396.74 (413.68) |
| | H-R | 1035.88 (534.49) | 744.80 (614.26) | 518.09 (591.07) | 792.29 (611.66) |
| **Total Movement Count** | H-H | 2.49 (1.66) | 1.65 (1.34) | 1.00 (0.00) | 1.75 (1.40) |
| | H-R | 3.22 (2.45) | 2.01 (1.71) | 1.32 (1.25) | 2.22 (1.98) |

Table III.12: The median subjective workload ratings for the motor and tactile channels are presented in Table III.12 by condition and victim assessed. The Likert-scaled subjective workload rating data is not normally distributed, therefore nonparametric analysis is used.

| Victim Assessment | Condition | Workload Channel | |
|---|---|---|---|
| | | Motor | Tactile |
| V1R1 | H-H | 1.5 | 3 |
| | H-R | 2.5 | 2 |
| V2R1 | H-H | 2 | 2 |
| | H-R | 2 | 2 |
| V3R1 | H-H | 3 | 3 |
| | H-R | 2 | 2 |
| V4R1 | H-H | 3 | 4 |
| | H-R | 3 | 3 |
| V5R1 | H-H | 2.5 | 2.5 |
| | H-R | 2 | 2.5 |
| V6R1 | H-H | 3.5 | 4 |
| | H-R | 2.5 | 2.5 |
| V1R2 | H-H | 2 | 3 |
| | H-R | 1 | 1.5 |
| V2R2 | H-H | 3 | 3 |
| | H-R | 2.5 | 3 |
| V5R2 | H-H | 2 | 3 |
| | H-R | 2 | 2 |
| V6R2 | H-H | 3 | 3 |
| | H-R | 2 | 2.5 |
| V4R2 | H-H | 3 | 3 |
| | H-R | 2 | 2 |

tive workload rating for the H-H condition was 2 and was also 2 in the H-R condition. The sum of the motor subjective workload ratings in the H-H condition was 414 and was 342 for the H-R condition. A Pearson's product-moment correlation found a significant correlation between H-H subjective motor workload ratings and the subtask time, $r(139) = 0.32$, $t = 3.91$, $p < 0.01$. The analysis also found a significant correlation for the H-R results, $r(151) = 0.21$, $t = 2.63$, $p = 0.01$. The longer subtask time resulted in higher subjective motor workload ratings for both conditions.

The median H-H condition tactile subjective workload rating for each victim assessment was the same or higher than the H-R condition ratings (Table III.12). The median rating was 3 for the H-H condition and 2 for the H-R condition. The sum of all tactile subjective workload ratings in the H-H condition was 462 and was 372 for the H-R condition. A Pearson's product-moment correlation analyzed the relationship between the subtask time and tactile workload ratings, and the H-H condition was not significant. The correlation for the H-R condition was significant, $r(151) = 0.35$, $t = 4.54$, $p < 0.01$, thus indicating that the H-R condition tactile ratings increased with longer subtask times.

The main effects of independent variables condition and triage level were tested on motor and tactile

Table III.13: Median motor and tactile workload channel ratings by condition and triage level.

| | **Motor Workload** | | **Tactile Workload** | |
|---|---|---|---|---|
| **Triage Level** | H-H | H-R | H-H | H-R |
| Delayed | 3 | 3 | 4 | 3 |
| Immediate | 2 | 2 | 3 | 2 |
| Expectant | 3 | 1.5 | 3 | 2 |
| Overall | 2 | 2 | 3 | 2 |

subjective workload ratings. Kruskal-Wallis tests indicated that H-H condition workload was higher than in the H-R condition for both motor ($\chi^2(1) = 10.01$, $p = 0.002$) and tactile ($\chi^2(1) = 13.92$, $p < 0.001$) subjective workload ratings. There were also medium Cohen's d effect sizes for both motor (d = 0.4) and tactile (d = 0.5) subjective workload ratings. The motor subjective workload ratings had 27.4% of the distributions of the two conditions not overlapping and the tactile ratings had 33.0% of non-overlap. Table IV provides the median values for both motor and tactile subjective workload ratings by triage level and condition.

A significant main effect of triage level across both conditions on the subjective measures of physical workload was also present in both the motor ($\chi^2(2) = 19.70$, $p < 0.001$) and tactile ($\chi^2(2) = 11.02$, $p = 0.004$) workload channels. A series of Mann Whitney U post-hoc tests reported that delayed victim assessments were significantly higher than the immediate victim assessments for both the motor ($p < 0.001$) and tactile ($p = 0.01$) channel ratings. There were small effect sizes for both motor (d = 0.1) and tactile (d = 0.2) ratings. In addition, the workload ratings were significantly higher for delayed victims than expectant victims for both the motor ($p < 0.001$) and tactile ($p = 0.01$) channel ratings, with small effect sizes for both the motor (d = 0.4) and tactile (d = 0.2) ratings.

The NASA-TLX physical demand responses resulted in a mean weighted demand in H-H condition participants of 4.63, (St. Dev. = 6.37) and was 2.64 (St. Dev. = 2.51) for the H-R condition. There was no significant effect of condition on the ratings.

**Comparison with Modeled Physical Workload**

The modeled motor workload predictions were compared with the evaluation motor workload ratings. The subjective motor workload rating scale did not match the modeled scale, thus the evaluation ratings were converted to match the modeled scale. Fig. 9 provides both the H-H modeled motor workload predictions and the rescaled H-H participant motor ratings, which very closely match one another. Nine of the eleven model data points are within one standard deviation of the mean evaluation rating. The mean delta between the H-H empirical data and the model was -0.89 (St. Dev. = 1.31). The H-R model data is provided in Fig. 10 along with the rescaled participant ratings. The H-R results follow a similar trend as found for the

Figure III.19: H-H modeled and mean empirical motor workload for each victim assessment by condition.

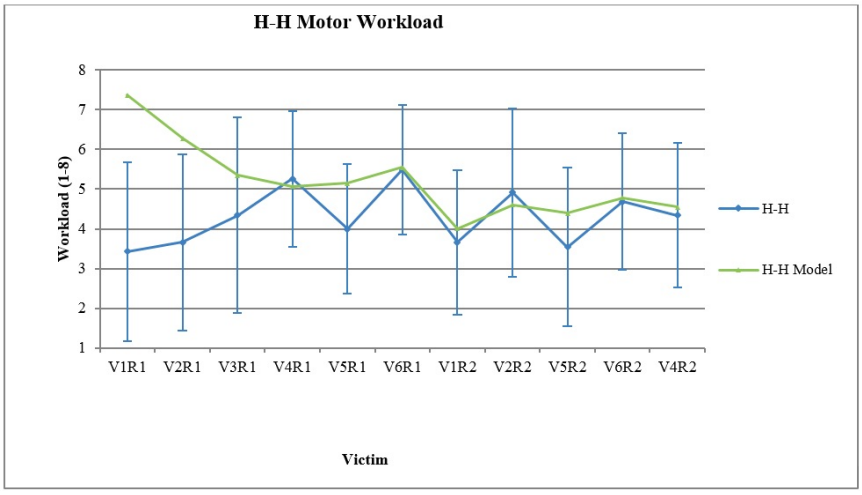
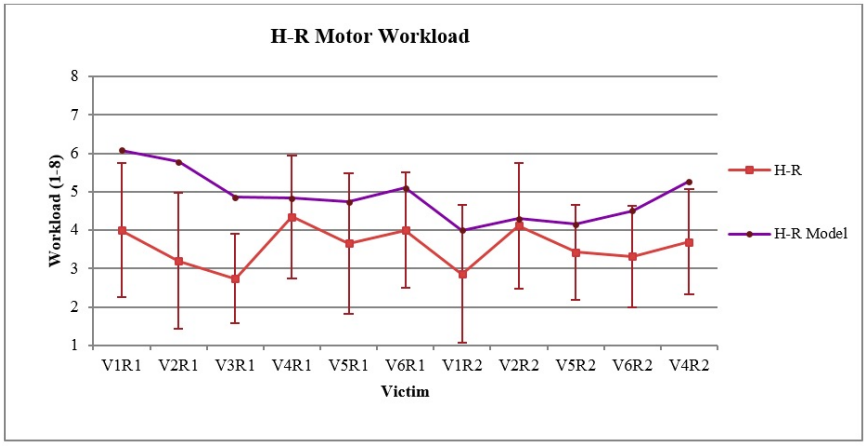



Figure III.20: H-R modeled and mean empirical motor workload for each victim assessment by condition.




H-H condition. Seven of the eleven model data points are within one standard deviation of the empirical ratings. The mean delta between the model and the participant ratings was -1.30 (St. Dev. = 0.73). T-tests comparing the deltas between each model and the corresponding participant ratings were not significantly different; therefore, the two models predicted motor workload with the same ability.

Tactile workload ratings from the IMPRINT Pro models incorporated monitoring and making distinctions using the sense of touch. The mean tactile modeled workload for the H-H condition was 1.42 (St. Dev. = 0.41), while the H-R condition was 1.29 (St. Dev. = 0.33). A t-test of the overall results indicated that this difference was not statistically significant. Even though there was no significant difference between the two conditions, the overall data trend indicates that the H-R condition has lower motor workload than the H-H condition. This difference corresponds with the trend observed in the participant-rated subjective tactile workload ratings.

Figure III.21: H-H modeled and mean empirical tactile workload for each victim assessment by condition.

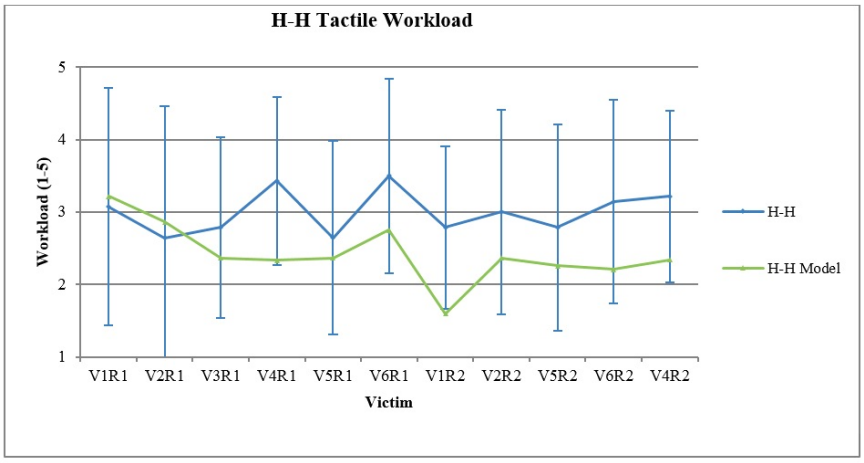
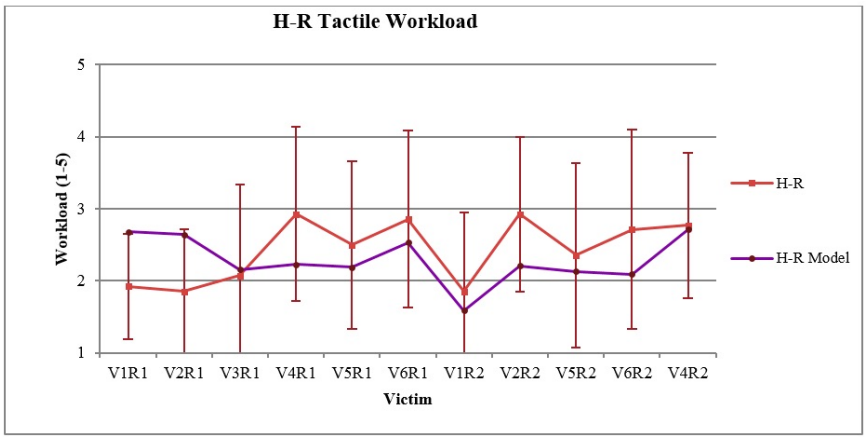



Figure III.22: H-R modeled and mean empirical tactile workload for each victim assessment by condition.




The modeled tactile workload predictions were compared with the evaluation tactile workload ratings. Figure III.21 provides the modeled H-H tactile workload predictions and the associated participant ratings. Both the evaluation results and the model values were on the same scale, thus no re-scaling was necessary. All the modeled tactile workload data points are within one standard deviation of the participant ratings. The mean delta between the H-H evaluation results and the model was 0.58 (St. Dev. = 0.47). The H-R model data and the participant ratings are plotted in Figure III.22. Ten of the eleven model data points are within one standard deviation of the evaluation ratings. The mean delta between the model and participant ratings was 0.15 (St. Dev. = 0.52). A t-test indicated that the H-R condition did a nearly significantly better job of predicting tactile workload, $t(20) = 2.05$, $p = 0.05$. Both models were able to predict actual tactile workload ratings well.

Table III.14: The primary and secondary reaction and response time medians (IQR in parentheses) in seconds.

| Metric | | Collaborative Evaluation | | Guided Evaluation |
|---|---|---|---|---|
| | | H-H | H-R | H-R |
| Primary Task Reaction Time | $< 30s$ | 3 (2-5) | 4 (3-9) | 1 (1-2) |
| | All Data | 3 (2-5) | 6 (3-15) | |
| Secondary Task Reaction Time | | 1 (0-1) | 1 (0-2) | 1 (1-2) |
| Primary Task Response Time | | 1 (1-12) | 2 (1-4) | 3 (1-7) |
| Secondary Task Response Time | | 1 (1-3) | 3 (1-9) | 2 (1-6.75) |

#### III.1.3.3 Reaction Time and Response Time Results

The H-H condition results were not analyzed due to a lack of audible sound recording for the human responder partner. The results are provided for comparison to the Collaborative evaluation H-R condition results in Chapter III.3.

The primary task reaction time median was 1 with an Inter-Quartile Range (IQR) = 1-2s (see Table III.14). The median secondary task reaction time was 1 (IQR = 1-2)s.

The median primary task response time was 3 (IQR = 1-7)s, while the secondary task response time median was 2 (IQR = 1-6.75)s. The descriptive statistics are provided in Table III.14.

Each of the analyzed metrics involved verbal tasks with explicit primary and secondary task prompts. The primary and secondary task reaction times were one second each, while the response times were longer, encompassing decision-making time.

### III.1.4 Discussion

$H_G^1$ focused on determining that a difference exists in overall workload between H-H and H-R teams. This hypothesis was partly supported based upon the evaluation results and comparison to the IMPRINT Pro model of human-robot team workload values. The H-R condition resulted in a slightly lower workload level, apparent in the model predictions and the subjective and physiological measures of workload from the evaluation.

Two theories were investigated as to why the H-R condition resulted in lower workload. First, the embodiment of the robot may directly result in lowering the human's workload during the H-R condition. The human partner was not present with the participant, while the robot partner was in the room; this experimental confound may have affected results. The second theory to explain results is that the robot's slower movement speed from victim to victim and slower interaction with the participant during the triage tasks may result in a lower workload for the H- R condition participants. These hypotheses were explored in the Collaborative evaluation (see Chapter III.2) that required the participant to complete joint, peer-based decision making and task activities while collocated with either a human or a robotic partner.

$H_G^2$ focused on determining the applicability of human performance modeling to human-robot peer-based teaming. The results from this analysis supported the hypothesis; the IMPRINT Pro models predicted the trend in workload very similarly to the actual in-task subjective workload rating results for both conditions. Overall, the models provided a valuable tool for prediction of workload. The data imply that the H-R condition participants experienced a lower level of workload than participants in the H-H condition, and the models more accurately predicted the H-R subjective workload. Current workload HPMFs may be applicable to human-robot peer based teams with the understanding that H-R teams may experience slightly less workload than H-H teams; however, the results may be affected when considering more complex tasks, increased dynamism in the relationship between teammates, or time-restricting tasks.

Modeled physical workload was also investigated as part of $H_G^2$, in the physical workload results. These model predictions did not correspond with the analysis of the objective results, but the model results matched the subjective measures. The correspondence between modeled results and subjective ratings is expected because the motor and tactile workload channels represent the mental assessment of physical demand and not the body's experience. These results indicate that the model may not provide accurate physical workload predictions that correspond with objective results, but do accurately represent the disparity between experienced physical work and its mental assessment within human-robot teams.

A specific investigation into physical workload was prompted by an observation that participants appeared to move more in the H-R condition, which appeared to contradict the physical workload subjective results. $H_G^3$ stated that physical workload will be higher in the H-R condition. This hypothesis was supported by the objective results; the objectively measured H-R condition physical workload was significantly higher than that for the H-H condition.

Despite the strong support for $H_G^3$ with objective data, the analysis of the subjective physical workload data did not lend its support. The H-H condition participants rated physical workload higher than the H-R condition participants. The modeled motor and tactile workload values also predicted slightly higher physical workload values for the H-H condition. One cause for the perceived workload difference may be due to the physical presence of the robot. The H-R condition participants were instructed to walk behind the robot when the robot moved. The participants may have anticipated the robot's next move. Additionally, the participants may have felt a need to step back in order to give the robot the opportunity to investigate, whereas the participants in the H-H condition did not have a partner physically present. Often human-robot evaluations focus primarily on subjective workload metrics. Based on the presented results, it appears that humans misjudge their perceived physical workload efforts; thus, subjective ratings alone may be an unreliable measure of physical workload and validation with objective measures may be necessary. Again, the medium effect size for both motor and tactile subjective workload ratings indicates that the difference in partner condition

affected the outcome in workload ratings, rather than chance. Additionally, the lower subjective ratings of physical workload in the H-R condition match the trend seen in overall workload ratings. Even though experienced physical movement and workload were higher in the H-R condition, the lower subjective ratings may imply that participants considered mental factors, such as time pressure, when rating physical workload.

$H_G^4$ stated that reaction time and response time are different measurements in mobile, non-laboratory environments with human-robot peer-based teams. This hypothesis was supported by the Guided evaluation data. The complex nature of the environment and tasks makes it a difficult challenge to develop methods of measuring reaction and response time, but the presented metrics offer ways of capturing verbal reaction and response times. A limitation of the metrics presented was their measurement via video coding; thus, it is an aim in future work to move away from video coding measurements for reaction and response time measurements for this domain. Another limitation of the presented evaluation was the inability to compare reaction and response time results between H-H and H-R teams, but the data presented were useful for comparison analysis with the Collaborative evaluation results.

Overall, the Guided evaluation formed initial insight into workload measurement techniques and differences between H-H and H-R teams. Potential experimental confounds (e.g., collocation of partners) were discovered and remedied for later evaluations.

### III.2 The Collaborative Evaluation

The Collaborative scenario involved the reconnaissance of an academic building following a bomb threat. The primary task was an investigation task. The participants searched six investigation areas for out-of-place items, determined if the items were suspicious or non-suspicious, and took photographs of a subset of items. The participants were paired with a responder partner (i.e., human or robot) and each partner had distinct tasks. Participants were responsible for searching bulletin boards and trash cans, while the human or robot partner was tasked with checking fire extinguishers and air quality. The participant and the responder also collaborated to make joint decisions about the environment and suspicious items. Twenty-six items were investigated and were either placed by experimenters or were inherent to the environment.

Generally, upon receiving a credible bomb threat, an area is evacuated and carefully, systematically, and thoroughly searched for suspicious items and bomb materials. During preliminary reconnaissance, teams of two search for unusual objects, while describing and taking note (e.g., pictures) of strange sounds, smells, and anything out of place (Humphrey and Adams, 2009; Mahoney, 1994). The teams check for items in containers (e.g., trash cans), behind and underneath items in the environment (e.g., furniture), and in the building structure (e.g., ceiling). If a suspicious object is found, it is not disturbed and information regarding its whereabouts and characteristics are reported immediately, including to incident command, who are located

at a safe distance from the incident area.

The hypotheses of the Collaborative evaluation included:

- $H_C^1$ = H-R workload is lower than H-H workload.
- $H_C^2$ = Human performance models will represent workload in human-robot peer-based teams.
- $H_C^3$ = Task performance will not suffer due to a participant having a robot teammate.
- $H_C^4$ = Physical workload will be higher in human-robot teams.
- $H_C^5$ = There is a difference in reaction time and response time between H-H and H-R teams.

### III.2.1 Modeling Description

The dynamic nature of the Collaborative scenario requires modeling for uncertainty. The Collaborative scenario models represent the scenario subtasks and uncertainty related to exactly which tasks are performed during each model simulation run. Adding alternate scenario paths created a range of mental workload values for each investigation area assessment. Each path had an associated probability of being followed, based on the estimated likelihood of an individual performing the action. The untrained human can spot the bag first and report it to the human or robot first responder partner (i.e., 50 percent), but there was also an equal chance (50 percent) that the untrained human is looking in the nearby trashcans or investigating another item at the time that the first responder partner reached the area near the bag. These probabilities created different outcomes based on running the model with differing random-value seeds. The results of the IMPRINT Pro models report predicted mental workload values for each workload channel at each time step and a predicted scenario completion time.

Reaction time and response time were also modeled for the Collaborative evaluation using IMPRINT Pro. Simple reaction and response time models were created for the primary and secondary tasks using IMPRINT Pro (Allender et al., 1995) for both conditions. The models represented the subtasks to complete the investigation and incorporated a level of uncertainty related to the exact task performance during each model run. For example, when answering a question, sometimes the simulated human responded yes and completed a follow-up action, while in other runs the response was no. The H-H and H-R team models differed only in the timing of some tasks. The robot spoke 1.3 times slower and traveled the same distance 1.5 times slower than the human partner.

Reaction time, which modeled the same aspects as the evaluation was modeled using IMPRINT Pro's micromodels of behavior (United States Army Research Laboratory, 2009). IMPRINT Pro provides values for simple reaction times including binary responses, physical matching, name matching or class matching.

The micromodel does not represent the time required for the visual system to recognize items in a setting like the modeled scenario, but does incorporate the reaction to an item and the decision to respond. Since these values do not incorporate all aspects of the reaction time, namely the recognition by the visual system, the modeled reaction combined micromodel time values and the time of each primary task reaction time was the same. The IMPRINT Pro's reaction time result represents the sum of all the individual micromodel reaction times.

#### III.2.1.1 Modeling Results

The model results were based on ten simulation trials for each model (H-H and H-R), where a different random number seed provided variability. The same ten random number seeds were used across the two models for individual trials (e.g., number seed 1 was used for H-H model trial 1 and H-R model trial 1). The output from each trial was divided by investigation area and mean workload values for each channel were computed.

The Fine and Gross Motor workload channels were combined for analysis. Gross Motor demands vary based on what is experienced for heavy physical exertion or movement of the entire body. The tasks involved in the Collaborative scenario that involve Gross Motor demand (e.g., walking or crouching) were rated a 1.0 on the 6.0 scale. The values above 1.0 on the Gross Motor scale are not relevant for this specific scenario and were not considered. Participants in the Collaborative evaluation rated the experienced demands in a combined Motor scale for simplicity, because Gross Motor scaled demand was limited in the task. The model predictions for the Fine and Gross Motor values were summed and subsequently considered to be on a scale from one to eight. Thus, these two sets of values were combined prior to calculating the total modeled mental workload.

Each mental workload channel value was normalized to a value between one and five prior to calculating the total modeled mental workload. After normalization, the total modeled mental workload was determined by summing the mental workload channels at each time point and calculating a time-weighted mean for each investigation area (the same calculation that was used for the Guided evaluation). The minimum possible modeled mental workload value was six and the maximum value was 30.

The mean modeled total mental workload across investigation areas for the H-H condition was 13.18 (St. Dev. = 0.58) and 12.96 (St. Dev. = 0.47) in the H-R condition. A Kruskal-Wallis test indicated that the H-H model's total mental workload was significantly higher than the H-R model's total mental workload, $\chi^2(1) = 9.02$, $p < 0.01$.

The mean total modeled mental workload values across both models by investigation area index are presented in Table III.15. A Kruskal-Wallis test indicated a significant main effect of investigation index for

Table III.15: Mean modeled mental workload by investigation area index and condition. Standard deviations are listed in parentheses.

| Investigation Index | Across Both Models | H-H Model | H-R Model |
|---|---|---|---|
| Low | 12.78 (0.71) | 12.86 (0.80) | 12.70 (0.61) |
| Medium | 13.24 (0.33) | 13.41 (0.31) | 13.08 (0.27) |
| High | 13.19 (0.36) | 13.28 (0.35) | 13.10 (0.35) |

Table III.16: Investigation area size ($m^2$).

| Investigation Area | Size |
|---|---|
| Area 1 | 52.43 |
| Area 2 | 53.28 |
| Area 3 | 84.74 |
| Area 4 | 57.32 |
| Area 5 | 43.73 |
| Area 6 | 124.04 |

total modeled mental workload, $\chi^2(2) = 7.02$, p = 0.03. Mann-Whitney U tests, with a Bonferroni adjusted $\alpha$, revealed no significant comparisons across the investigation indices.

The mean total modeled mental workload by condition and investigation index are also presented in Table III.15. A Kruskal-Wallis test indicated a significant interaction effect of condition and investigation index on total modeled mental workload, $\chi^2(5) = 17.00$, $p < 0.01$. Mann-Whitney U tests, with a Bonferroni adjusted $\alpha$, revealed no significant results within conditions or between conditions for the same investigation index.

A task-density ratio was calculated using model data based on the number of items in an area, the size of the area, and the time to search the area, as shown in Equation III.1:

$$TaskDensity = NumberofItems * SizeofArea/AreaTime \tag{III.1}$$

Area Time refers to the amount of time the model predicted that participants required to investigate each area; the Number of Items per area is provided in Table III.23; and the Area Size (meters squared) is presented in Table III.16. Task density is a calculation of the number of tasks completed in a given amount of time (Weinger et al., 1994). The investigation tasks in this scenario include investigating a given area and finding a number of items.

The mean overall task density was 0.41 (St. Dev. = 0.16) for the H-H condition, while the H-R condition was 0.34 (St. Dev. = 0.13). A Kruskal-Wallis test indicated that the predicted task density was significantly higher in the H-H condition, $\chi^2(1) = 13.59$, $p < 0.01$, and as a result, mental workload was predicted to be lower in the H-R condition.

The mean task density for low investigation index areas was 0.45 (St. Dev. = 0.13), 0.38 (St. Dev. =

0.13) for medium investigation index areas, and 0.41 (St. Dev. = 0.20) for high investigation index areas in the H-H model. The mean task density was 0.37 (St. Dev. = 0.09) for low investigation index, 0.30 (St. Dev. = 0.09) for medium investigation index areas, and 0.35 (St. Dev. = 0.17) for high investigation index areas in the H-R model. A Kruskal-Wallis test indicated no significant main effect of investigation index. These results indicate that the investigation areas did not have significantly differing levels of task density.

Workload (not to be confused with mental workload) is the amount of work completed in the amount of time available to complete it (Wickens et al., 2003). Based on this definition, an equation can be formed to estimate workload and solve for the work completed, using the model predictions of workload values and area completion times. The amount of work completed in each investigation area was computed based on the work completed and the time to complete the work (Equation III.2):

$$WorkCompleted = MentalWorklod * AreaTime \tag{III.2}$$

The mental workload values are known based on the predictions from the model. The time to complete the work (AreaTime) is also known based on model predictions. Overall, the mean work completed in the H-H Model was 2535.06 (860.88) workload seconds and 2933.15 (893.40) workload seconds for the H-R Model. Computing the work completed permits assessment of what was completed by the human by factoring out time. A workload mathematical unit can be work per time unit. Equation III.2 factors out timing by multiplying workload by area time and results in only the work completed, as shown in Equation III.3:

$$Work = Work/Time * Time \tag{III.3}$$

Equation III.3 is a simple estimation, but offers insight into whether or not the time the teams' required to complete the task significantly affects the teams' performance. A significantly lower work-completed value for a team reveals that the longer time taken was not spent productively.

Overall, predicted mental workload was significantly higher in the H-H Model than in the H-R Model. While there were significant effects of investigation index and an interaction effect between condition and investigation index, individual comparisons found no significant results.

The modeled motor workload combined the fine and gross motor channels from the IMPRINT Pro model using the same method as in the Guided evaluation (see Chapter III.1). The mean motor workload rating, from the six investigation areas, in the H-H model was 2.48 (St. Dev. = 0.14) and was 2.40 (St. Dev. = 0.12) in the H-R model. A t-test indicated that the motor workload from the two conditions were not significantly different. Despite a lack of significance, the model's higher values in the H-H condition echo the results shown in the subjective motor workload ratings.

The reaction time and response time models were also developed simultaneously with the evaluation. The models incorporated the primary and secondary task reaction times, and a general question response time model for comparison with the primary and secondary task response times. The primary task reaction time was predicted to be 1.41s, based on a combination of eye fixation time (0.30s), eye movement time (0.10s), head movement time (0.20s), decision time (0.07s), search time in the environment (0.60s), and the time to determine if an item was out-of-place (0.24s). The time components were not affected by the human's teammate and were based only on the human participant and the environment, thus the primary task reaction time was modeled in the same way in both conditions.

The H-H model secondary task reaction time included the time to comprehend a question after it was spoken (0.72 s) and utter a one-word response (0.34s), for a total time of 1.06s. The H-R model required an additional 0.21s for the modeled human to process the robot's speech; thus the H-R model predicted secondary task reaction time to be 1.27s. The secondary task response time was modeled to account for matching the given name to the memorized list (0.45s). The H-H model secondary task response time was 1.51s and the H-R time was 1.72s. Reaction time and response times predicted by the model are compared to the Collaborative evaluation results in Chapter III.2.3.3.

#### III.2.1.2 Modeling Discussion

The Collaborative scenario models provided predictions of mental workload changes for both conditions during the actual user evaluations. These models account for uncertainty by incorporating probability-based alternative model paths, and random number seeds provided variance between trials. The same ten random number seeds were used across the two models for individual runs. Any differences between the two models are attributed to the modeled slower robot speech and movement tasks.

The model does not predict a significant discrepancy between the H-H and H-R teams for work completed. If an H-R team were to take a longer time to complete the task (as predicted) and has lower workload values (as predicted), but has a significantly lower work-completed value (not as predicted), then this result implies that the H-R team will not be as successful at performing the task as the H-H team.

The model development occurred simultaneously with the evaluation, thus it was not known that mental workload was not manipulated by investigation index. This confound created an inconsistent experimental manipulation of mental workload; however, the evaluation results are analyzed by investigation index in order to determine which mental workload measures may reflect the intended goal of the investigation index measure. The Collaborative evaluation workload model results are compared with Collaborative evaluation results in Chapter III.2.3.1.

Table III.17: Descriptive statistics (median and inter-quartile range) for the H-H and H-R models and evaluation results.

| Area | H-H Model | H-H Condition | H-R Model | H-R Condition |
|---|---|---|---|---|
| 1 | 169.20 (164.98-172.11) | 204.5 (172.75-245.25) | 196.34 (179.79-199.27) | 284.0 (268.25-345.25) |
| 2 | 105.80 (103.88-27.23) | 179.0 (149.00-239.75) | 148.47 (137.19-160.62) | 254.0 (228.75-282.00) |
| 3 | 157.51 (152.47-161.89) | 300.0 (267.50-360.50) | 185.52 (185.52-192.87) | 457.0 (398.00-500.25) |
| 4 | 309.48 (305.17-313.09) | 424.0 (352.25-479.00) | 332.50 (325.85-368.26) | 483.0 (432.00-530.50) |
| 5 | 183.11 (177.24-186.22) | 256.5 (218.50-282.00) | 219.41 (214.98-223.77) | 279.5 (254.50-318.00) |
| 6 | 211.84 (205.14-225.89) | 422.0 (376.25-463.00) | 251.35 (248.43-256.37) | 463.0 (420.00-543.00) |
| Overall | 175.92 (154.92-210.97) | 291.0 (218.75-389.75) | 211.73 (185.52-251.35) | 368.0 (274.00-464.00) |

#### III.2.1.3 Secondary Modeling Analysis

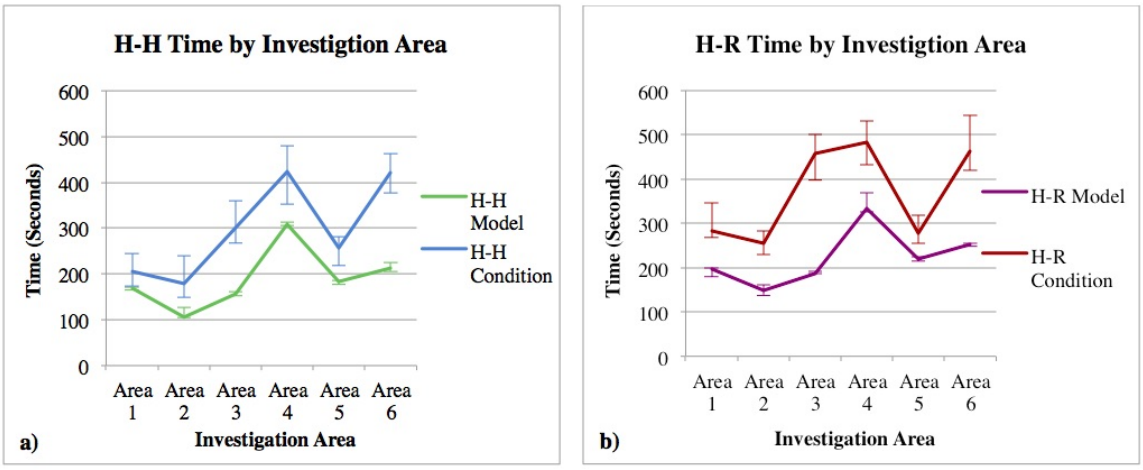
Prior to conducting further investigations following the Collaborative evaluation, it is necessary to investigate why the Collaborative evaluation H-R models were not representative of the actual evaluation workload results (see Chapter III.2.3.1). It is crucial to also analyze whether timing was accurate in the Collaborative model. Task timing is a crucial factor in determining workload levels, and inaccurate timing models may prevent the use of any model of workload for H-H or H-R teams.

The time taken to complete the search of each investigation area, or the subtask time, was examined for the H-H condition, the H-H model, the H-R condition, and the H-R model. The median subtask times are presented in Table III.17, with the associated inter-quartile ranges (IQRs) for the H-H model predictions, the H-H condition evaluation results, the H-R Model predictions, and the H-R condition evaluation results. Figure III.23 provides the median and IQRs for the model and evaluation results. A Mann-Whitney U test assessed whether modeled and evaluation result time values were from the same distributions. The H-H model and the H-H evaluation results were significantly different, $U = 1241, p < 0.001$. The H-R model and the H-R evaluation results were also significantly different, $U = 5694, p < 0.001$. The H-H and H-R models significantly underestimated the evaluation results in both conditions.

These results indicate that neither model accurately predicted the timing of the area investigations in the Collaborative evaluation. Figure III.23 demonstrates that the H-H model and the H-H condition results follow a very similar trend. The H-R model and the H-R results are similar, except for the large deltas between the data points for Areas 3 and 6.

The model's timing was subsequently adjusted in order to address the large discrepancy. Following the adjustment of H-H model's task time adjustment, the H-R model was adjusted in the same way and analyzed. The extra time left represents robot communication cost. Finally, workload was reanalyzed using the model's new confidence intervals to assess whether the increased accuracy in timing also improved representation of workload in either condition.

Figure III.23: Median time spent per investigation area for model predictions and evaluation condition results in (a) the H-H condition and (b) the H-R condition. Error bars represent the IQR for each median data point (25 percentile below and 75 percentile above).




**Updating H-H Model Timing**

Timing was adjusted in the Collaborative H-H model. The first step required examining the video coding data from the evaluation and comparing the time taken to assess a subset of representative items to the time modeled for the investigation. The modeled time was within a mean of 4.90 s (St. Dev. = 3.23). The time differences are explainable by examining the time when participants needed to describe an item, answer an open-ended question, make a decision, or have time to walk around on his or her own. The participants took longer to do these types of tasks than modeled, as extra time for a lack of efficiency was not included. The timing for each task after the adjustment included time for decision making of 3 or 6 s, depending on the task requirements.

The second adjustment for timing was to add the option investigating extra items not set out by experimenters during the evaluation. H-H participants reported a total of 41 extra items. The likelihood of a participant in the H-H participant reporting an extra item in any of the investigation areas was 37.96% and this likelihood dictated how often the newly added extra item functions were executed in the model.

Finally, search time was grossly underestimated in the initial models. Participants took longer to search the areas (e.g., examining all bulletin boards in the hallway, looking under tables) and find the evaluation items. Participants were, on average, very thorough in the search. The extra search time added to the models is presented in Table III.18. Areas 3 and 6 required the longest additional time. These areas also offered participants the most time to freely roam the environment, adding time to the investigation.

The resulting updated H-H model resulted in times that were very close to the mean models. All modeled times were within the 95% confidence interval of the evaluation results (see Table III.19). Figure III.24

Table III.18: The search time added to the Collaborative H-H model during the timing adjustment analysis.

| Area | Search Time |
|---|---|
| **Area 1** | 10 s |
| **Area 2** | 60 s |
| **Area 3** | 110 s |
| **Area 4** | 45 s |
| **Area 5** | 20 s |
| **Area 6** | 110 s |

Table III.19: Mean H-H Condition subtask time with confidence interval and mean subtask time value for the adjusted H-H model.

| | Area 1 | Area 2 | Area 3 | Area 4 | Area 5 | Area 6 |
|---|---|---|---|---|---|---|
| **H-H Condition Mean** | 206.28 | 206.11 | 315.22 | 428.61 | 254.44 | 420.00 |
| **H-H Condition St. Dev.** | 45.87 | 74.05 | 61.63 | 77.65 | 47.79 | 84.52 |
| **H-H Condition 95% C.I. Low** | 185.09 | 171.90 | 286.75 | 392.74 | 232.36 | 380.95 |
| **H-H Condition 95% C.I. High** | 227.47 | 240.32 | 343.69 | 464.48 | 276.52 | 459.05 |
| **H-H Model Mean** | 202.02 | 204.33 | 307.65 | 409.91 | 250.32 | 404.84 |
| **H-H Model St. Dev.** | 24.39 | 32.86 | 18.31 | 22.85 | 23.41 | 27.02 |
| **Within C.I.?** | Yes | Yes | Yes | Yes | Yes | Yes |

presents the adjusted timing model along with the initial model's timing data in comparison to the evaluation data. All three sets of data show similar trends, but the initial model severely underestimated subtask time. The new model used the evaluation data in order to create a realistic model of what the H-H participants experienced during the evaluation.

**Assessment of H-R Model after Timing Adjustment**

The updated H-H model was used to examine whether the H-R condition participants truly took longer because of the speed of the robot or the cost and impact of interacting with the robot. If the H-H model, adjusted for the robot's speech and movement speed, significantly underestimated the H-R condition subtask time, then a known, measurable, cost of robot communication exists. If the new H-R model accurately represented the H-R subtask time, the longer time spent in the H-R condition was due to the robot's slower speech and movement speed and the model does not need further time adjustment.

The H-R condition subtask time, the initial modeled subtask time, and the adjusted subtask time model are presented in Figure III.25. The descriptive statistics for the H-R condition and the adjusted model are available in Table III.20. Areas 1, 2, and 3 were not modeled within the 95% confidence interval of evaluation results. Area 2 is close to the low cutoff value, with approximately an eight second difference, but Areas 1 and 3 are over one minute away from being within the interval. Areas 4, 5, and 6 were modeled very closely to the mean and were within the confidence interval.

Figure III.24: Mean subtask time for the H-H Condition with the error bars representing the 95% confidence interval. Initial model shown as well as model after timing adjustments.

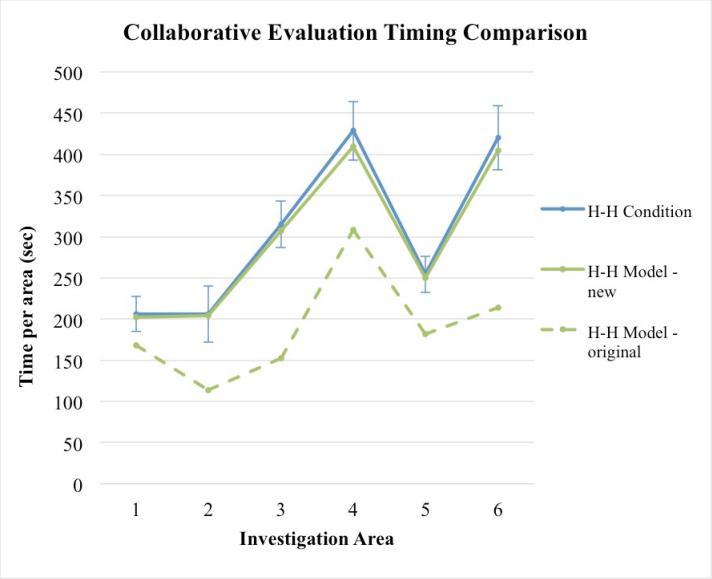



Figure III.25: Mean subtask time for the H-R Condition with the error bars representing the 95% confidence interval. Initial model shown as well as model after timing adjustments.

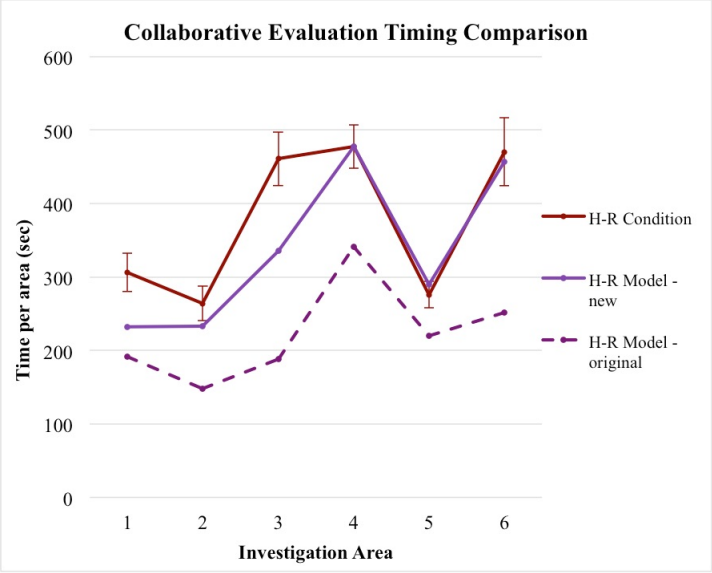


Table III.20: Mean H-R Condition subtask time with confidence interval and mean subtask time value for the adjusted H-R model.

| | **Area 1** | **Area 2** | **Area 3** | **Area 4** | **Area 5** | **Area 6** |
|---|---|---|---|---|---|---|
| **H-R Condition Mean** | 306.06 | 263.89 | 460.50 | 477.06 | 275.78 | 470.24 |
| **H-R Condition St. Dev.** | 56.09 | 50.83 | 78.48 | 63.74 | 38.45 | 99.91 |
| **H-R Condition 95% C.I. Low** | 280.15 | 240.41 | 424.24 | 447.61 | 258.02 | 424.09 |
| **H-R Condition 95% C.I. High** | 331.97 | 287.37 | 496.76 | 506.51 | 293.54 | 516.39 |
| **H-R Model Mean** | 232.08 | 232.82 | 335.21 | 477.67 | 289.51 | 470.24 |
| **H-R Model St. Dev.** | 29.11 | 31.83 | 23.75 | 29.96 | 23.44 | 38.30 |
| **Within C.I.?** | No | No | No | Yes | Yes | Yes |

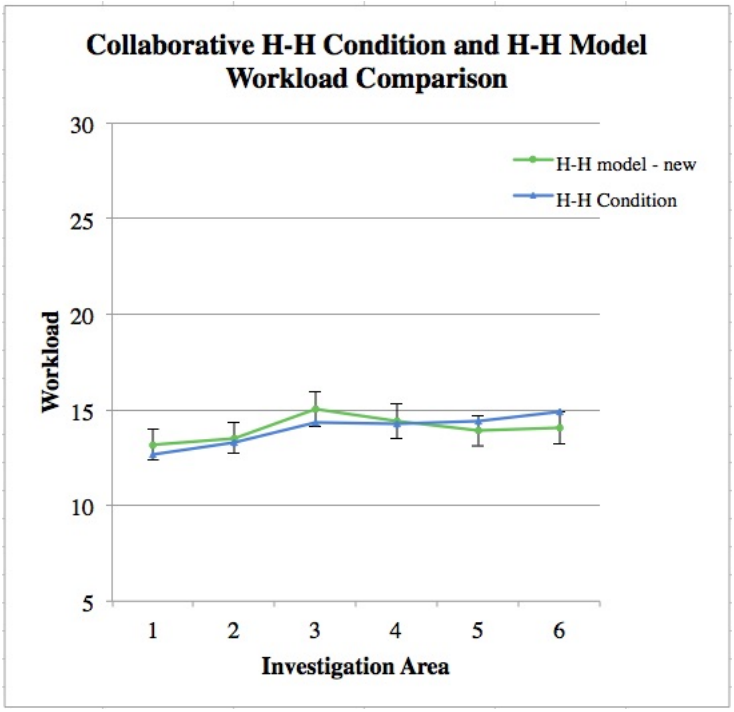
These results indicate that two areas in particular, Areas 1 and 3, were not represented well in the H-R model. Area 1 is the first point of interaction for the participant and robot, and there may be a communication cost when the participant first adjusts to the robot's speaking style (e.g., 1.3 times slower than human speech, occasionally ill-timed interruptions, momentary delays in some responses). The majority of Area 3's investigation took place inside a laboratory room that had audible fans, further inhibiting the communication between the participant and robot. The robot compensated for the fan noise by raising the volume of its speakers within this room, but a communication cost was incurred with many requests for repeated instructions by the participant, and time taken to move closely to the robot during its utterances.

The unmodeled communication costs were calculated as a percentage of the modeled subtask time that was missing from the model (i.e., delta between modeled mean and evaluation mean). The robot communication cost from Area 1 was 31.88% (73.98s), and was 37.38% for Area 3 (125.29s). Area 2 also was not within the 95% confidence interval of the evaluation results and had a 13.35% communication cost (31.07s). The mean of Areas 1, 2, and 3 was 27.53% (76.78s). The communication cost of working with the robot was also calculated over the entire evaluation; even though the latter three investigation areas were modeled accurately, building future models may factor in the general communication cost found overall, 15.09%.

Reasons other than communication costs were considered for the inaccurate models in Areas 1 and 3, but they were not supported. The alternate reasons include the more difficult physical actions taken by the robot (e.g., turning corners, driving through doorways). The movements were not time sinks, as Areas 5 and 6 include similar corner turns as Area 1 and Area 6 includes a doorway entry that is similar to the one found in Area 3. These actions were accounted for by modeling robot partner movement speed as 1.5 times slower than human partner movement speed.

Modeling human-robot peer-based teams generally may include a 15.09% communication cost, but when examined further, higher costs are due to adjusting to robot communication (e.g., meeting the robot), and adjusting to environmental conditions (e.g., noise). If the participant is familiar with the robot and there are

Figure III.26: Mean workload for the H-H Model with the error bars representing the 95% confidence interval with mean H-H Condition in situ workload ratings.




low noise levels, the model may need no adjustment for communication costs, as seen in the Collaborative evaluation model for Areas 4, 5, and 6.

**Updated Workload Analysis**

The final step in analyzing the timing of the Collaborative evaluation models is assessing the accuracy of the adjusted workload representation. Following the improvement of the task timing, the workload values changed. The adjusted H-H modeled workload values and the H-H condition in situ workload ratings from the Collaborative evaluation are presented in Figure III.26. The descriptive statistics for the H-H model, its 95% confidence interval, and the H-H condition are provided in Table III.21. All H-H condition means are within the 95% confidence interval of the H-H model.

The adjusted H-R workload values and H-R condition ratings are in Figure III.27. The descriptive statistics for the H-R model, its 95% confidence interval, and the H-R condition are provided in Table III.22. The H-R model followed the workload trend seen in the H-R condition results; however, the model overestimated the H-R condition results. None of the H-R condition data points were within the 95% confidence interval of the model results.

One theory for the H-R model's overestimation of the H-R condition's subjective workload ratings is

Table III.21: Descriptive statistics for the H-H condition and model, and the 95% confidence interval for the H-H model, after timing adjustments.

| | Area 1 | Area 2 | Area 3 | Area 4 | Area 5 | Area 6 |
|---|---|---|---|---|---|---|
| **H-H Model Mean** | 13.18 | 13.53 | 15.03 | 14.40 | 13.91 | 14.06 |
| **H-H Model St. Dev.** | 1.28 | 1.31 | 1.46 | 1.47 | 1.27 | 1.39 |
| **H-H Model 95% C.I. Low** | 12.39 | 12.72 | 14.13 | 13.49 | 13.12 | 13.2 |
| **H-H Model 95% C.I. High** | 13.97 | 14.34 | 15.93 | 15.31 | 14.70 | 14.92 |
| **H-H Condition Mean** | 12.67 | 13.33 | 14.33 | 14.28 | 14.44 | 14.89 |
| **H-H Condition St. Dev.** | 4.64 | 5.02 | 5.60 | 4.89 | 5.50 | 5.25 |
| **Within C.I.?** | Yes | Yes | Yes | Yes | Yes | Yes |

Figure III.27: Mean workload for the H-R Model with the error bars representing the 95% confidence interval with mean H-R Condition in situ workload ratings.

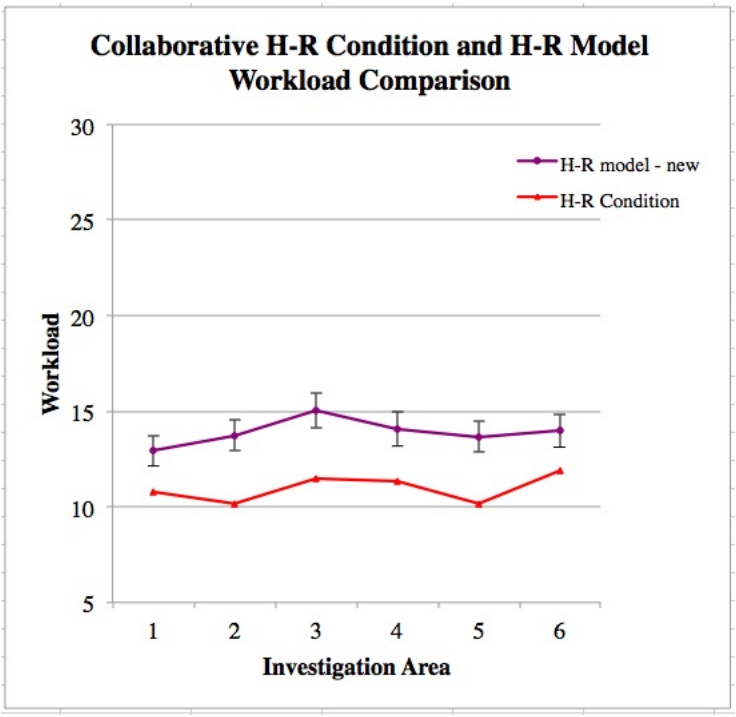


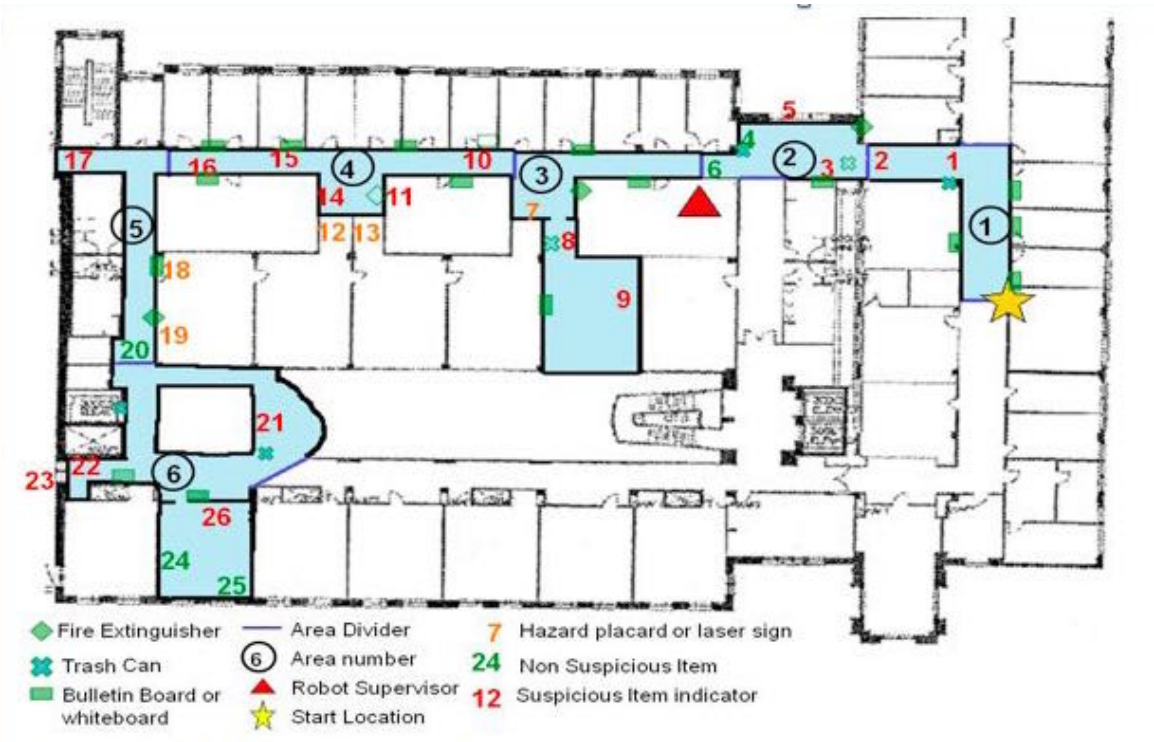
Table III.22: Descriptive statistics for the H-R condition and model, and the 95% confidence interval for the H-R model, after timing adjustments.

| | **Area 1** | **Area 2** | **Area 3** | **Area 4** | **Area 5** | **Area 6** |
|---|---|---|---|---|---|---|
| **H-R Model Mean** | 12.93 | 13.75 | 15.02 | 14.09 | 13.67 | 13.98 |
| **H-R Model St. Dev.** | 1.23 | 1.36 | 1.51 | 1.37 | 1.35 | 1.40 |
| **H-R Model 95% C.I. Low** | 12.17 | 12.91 | 14.08 | 13.24 | 12.83 | 13.11 |
| **H-R Model 95% C.I. High** | 13.69 | 14.59 | 15.96 | 14.94 | 14.51 | 14.85 |
| **H-R Model Mean after search adjustment** | 12.44 | 11.42 | 11.37 | 12.95 | 12.81 | 12.21 |
| **H-R Model St. Dev. after search adjustment** | 1.21 | 1.35 | 1.50 | 1.34 | 1.33 | 1.39 |
| **H-R Model 95% C.I. Low after search adjustment** | 11.69 | 10.58 | 11.66 | 12.12 | 11.99 | 11.35 |
| **H-R Model 95% C.I. High after search adjustment** | 13.19 | 12.26 | 13.52 | 13.78 | 13.63 | 13.07 |
| **H-R Condition Mean** | 10.78 | 10.17 | 11.50 | 11.33 | 10.17 | 11.89 |
| **H-R Condition St. Dev.** | 3.95 | 3.92 | 4.30 | 3.71 | 3.17 | 4.25 |
| **Within C.I. before searching adjustment?** | No | No | No | No | No | No |
| **Within C.I. after searching adjustment?** | No | No | No | No | No | Yes |

the slower movement speed of the team. As the team moves more slowly through the environment, a more thorough search takes place; thus, the extra searching times added to the H-H model, with the associated workload levels for searching, are unnecessary for the H-R condition. The time was taken for moving around the environment, but the participant may not have been actively searching during that time, and experienced lower workload during searching. Additional analysis was performed to remove unnecessary searching time in the H-R model and the results are provided in Table III.22. Only the Area 6 condition results were within the model's 95% confidence interval.

Overall, the H-H model predicted workload levels very similarly to the participants' rating of workload in the H-H condition. The H-R model was created in two ways. First, by simply adjusting the timing of movement and speech tasks, which significantly overestimated workload ratings. Second, the added search time was lowered; however, this second workload model resulted in only one investigation area's evaluation condition's results accurate representation by the model. H-R participants rated workload, on average, 10.20% lower (St. Dev. = 6.63%) for each investigation area than was predicted by the H-R model. This information was used when creating the Time-Structured H-R models, and can be used in other future human-robot peer-based team models, to account for the participants' lower rating of perceived workload.

Figure III.28: The map of the scenario environment with each investigation area shaded and labeled. All item locations are shown.




### III.2.2 Experimental Method

#### III.2.2.1 Environment

The scenario environment was a single floor of a Vanderbilt University academic building. The hallway and two laboratories were divided into six investigation areas (1-6), as shown in Figure III.28. Each area was assigned an investigation index. The investigation index is a point value assigned based on the investigation area contents and required tasks. One point was assigned for a non-suspicious item that participants were trained to investigate (e.g., an empty trash can). Two points were assigned for an item that required discussion by the team members (e.g., a message on a whiteboard). Three points were assigned for items found by physically moving objects around (e.g., removing a recycling bin lid). An additional point was added for each necessary joint decision. Each of the two low investigation index areas (Areas 1 and 3) had scores of 10, the medium investigation index area (Areas 2 and 5) scores were 15, and the high investigation index areas (Areas 4 and 6) each scored 20 points. The teams, independent of condition, followed the same path through the hallway and laboratories, while traversing each investigation area in numerical order. The investigation began at the location labeled by the star in Figure III.28.

Nineteen suspicious and non-suspicious items were placed in the environment by the experimenters; see

Table III.23: The items listed by investigation area with a description and indication as to whether or not the item was suspicious.

| Investigation Area | Item Number | Item Description | Suspicious? |
|---|---|---|---|
| Area 1 | 1 | Map in recycling bin | X |
| | 2 | backpack on bench | X |
| Area 2 | 3 | Soda bottle with suspicious material in recycling bin | X |
| | 4 | Math equations written on white board | |
| | 5 | Box of bomb-making supplies on windowsill | X |
| | 6 | Box of textbooks on floor | |
| Area 3 | 7 | Hazard placard and laser sign over closed lab door | |
| | 8 | Bomb-making instructions in a trash can in the large lab | X |
| | 9 | Bomb-making materials on counter in the large lab | X |
| Area 4 | 10 | Printout discussing C4 on bulletin board | X |
| | 11 | Fire extinguisher sitting outside of its case | |
| | 12 | Hazard placard next to closed lab door | |
| | 13 | Additional hazard placard next to another lab door | |
| | 14 | Bomb underneath eyewash station | X |
| | 15 | Wires hanging from the ceiling | X |
| | 16 | Map on bulletin board | X |
| Area 5 | 17 | Note in emergency door | X |
| | 18 | Laser sign over closed lab door | |
| | 19 | Laser sign over another closed lab door with a keypad lock | |
| | 20 | Lunchbox under water fountain | |
| Area 6 | 21 | Box of bomb-making supplies under table in hall | X |
| | 22 | Message written on whiteboard: “rendezvous code green” | X |
| | 23 | Note on windowsill | X |
| | 24 | Computer cables in the small lab | |
| | 25 | Cleaning supplies in the small lab | |
| | 26 | Unknown machine or experimental equipment in the small lab | X |

Table III.23 for a complete list and the identification of suspicious items. Additionally, seven items in the environment were incorporated, including: hazard placards, laser warning signs, a fire extinguisher, and a piece of lab equipment. Items normally present, but not included in the scenario, are not listed in Table III.23, for example, bulletin boards and white boards, recycling bins and trashcans, and fire extinguishers. Figure III.28 labels each item in the table with the corresponding item number while also delineating locations of benign objects.

#### III.2.2.2 Apparatus

All H-H condition participants completed the evaluation prior to the H-R condition participants. During the H-H condition, an evaluator played the role of the first responder and human teammate. A script dictated verbal interactions between the human participant and human experimenter. The same female experimenter was partnered with all participants.

The H-R condition paired the participant with a semi-autonomous Pioneer 3-DX robot equipped with a laser range finder for navigating through the environment autonomously on a pre-planned path. The robot was supervised by the remote experimenter from the room labeled with the triangle in Figure III.28. Radio frequency-identification (RFID) tags in the environment allowed the robot to identify specific objects and trigger appropriate robot behaviors, such as sensing an RFID tag marking a suspicious box on a windowsill that resulted in the robot announcing the identification of the box. The robot's speech was scripted using the same script as the H-H condition. The experimenter controlled the script, and the experimenter was allowed to repeat statements and insert customized utterances, if needed. The robot spoke using a digital female voice with an American accent, which was chosen to be similar to the human experimenter's voice in the H-H condition.

The Collaborative evaluation used a mixed design with the participants as a random element. The experimental condition, H-H or H-R, differed in a between subjects comparison. The within-subjects element was the series of investigation areas.

#### III.2.2.3 Participants

The evaluation included 36 participants. The 19 male and 17 female participants, eighteen in each condition, ranged in age between 18 and 56 years old. The H-H condition mean age was 27.4 (standard deviation (St. Dev.) = 8.4), while the mean H-R participant age was 24.1 (St. Dev. = 4.6). Participants rated their experience with search and rescue response and robots as little or no experience with no significant difference between conditions, based on Likert scales from 1 (little or no experience) to 5 (very experienced).

#### III.2.2.4 Metrics

The objective metrics included: physiological data from a portable BioHarness ECG monitor (Biopac Systems, 2013) (heart rate variability, breathing data, R-R data, heart rate, respiration rate, skin temperature, posture, vector magnitude data and acceleration data), subtask time, correctness of responses to the secondary task questions, secondary task failure rate, primary task failure rate, working memory task, accuracy of search task, primary task reaction time, secondary task reaction time, primary task response time, secondary task response time, and pedometer data. Subjective metrics included in situ workload ratings collected after triaging each victim, post-experimental questionnaire responses, and post-experimental NASA-TLX (Hart and Staveland, 1988) responses. The physiological responses presented in this chapter include heart rate variability, heart rate, respiration rate, posture, vector magnitude, and pedometer data. Pedometer values were recorded using a Garmin footpod and watch.

The reaction and response time metrics were collected using the same definitions as were defined in Chapter III.1.2.4. Primary task reaction time recorded the time required to react to an out-of-place item in the participant's field of vision, as recorded by the participant's point-of-view camera. The primary task stimulus onset time occurred when an out-of-place item entered the recorded field-of-view. The participant's reaction time corresponded with reacting with a verbal utterance, identifying the item with the laser pointer, touching the item, fixating the camera view directly on the item, or beginning to take a photograph. Secondary task reaction time was measured by recording the time when participants first reacted to the On List secondary task question. The stimulus onset time occurred when the responder partner completed asking the question, while the reaction was the participant's first verbal response (e.g., "uh...").

Primary task response time relied on directed prompts related to the primary task (e.g., deciding an item's suspiciousness, prompt to take a photograph). The stimulus onset time was the time when the prompt was completed and the response occurred at the moment a participant provided an appropriate response. Stall words (e.g., "uh") were not considered an appropriate response. Secondary task response time was measured by recording when participants provided an appropriate answer to the On List secondary task questions. Due to the potential impact of priming, the Danger Level question response times are not included in the determination of secondary task response time. Two video coders determined primary and secondary task reaction times and secondary task response times. The inter-coder reliability had a Cohen's kappa value of 0.81. A second set of video coders (also completed video coding of Guided evaluation reaction and response time data) recorded the primary task response time data with an inter-coder reliability score with a Cohen's kappa value of 0.90.

The subtask time represented the time participants spent completing the search of each of the six inves-

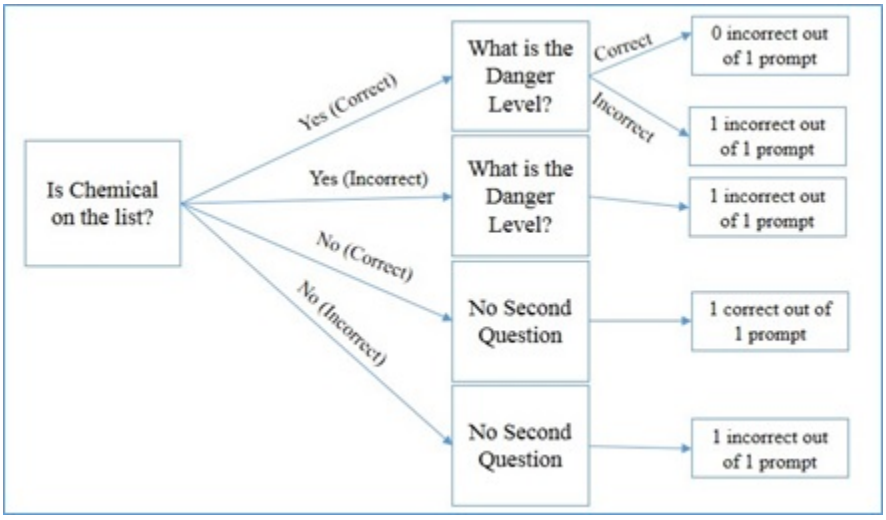



Figure III.29: A flow chart of the procedure for determining secondary task failure rate in the Collaborative evaluation

tigation areas. Timing began when participants arrived in the area, following the completion of the previous area's in situ workload ratings, and completed when the responder partner (human or robot) began requesting the in situ workload ratings for that investigation area.

The secondary recognition task questions involved memorizing a list of six chemical names, each with an associated numerical danger level. Twelve secondary question prompts were asked during the evaluation, with two per each investigation area. Each secondary question prompt consisted of an On List question, asked first, asking if a specific chemical name was on the memorized list. If the participant replied affirmatively, a second Danger Level question was asked, regardless of the correctness of the On List question response, regarding the chemicals corresponding danger level.

The Collaborative evaluation's secondary task failure rate considered responses to the two questions, On List and Danger Level. Figure III.29 presents the method of determining correct and incorrect responses to the On List and Danger Level questions, which establishes the secondary task failure rate values. The On List questions were asked 209 times in the H-H condition and 210 times in the H-R condition. This discrepancy is due to the experimenter neglecting to ask a question during the evaluation. Danger Level questions were asked 84 times in the H-H condition and 92 times in the H-R condition. Situations where the Danger Level question was asked, but the chemical was not on the list (i.e., the second path in Figure 1 showing an incorrect Yes answer to the On List question with a follow-up Danger Level question) were not considered.

There were four components of primary task performance in the Collaborative evaluation: participant items report task, team items first report task, suspicious item report task, and photographs taken task. Twenty-six items were placed in the six investigation areas. Table III.24 presents all of the items from

the investigation with the corresponding categorization into participant items, team items, whether or not the item is suspicious, and whether or not it is to be photographed. The participant items are the six items that can only be found by the participant, based on the difference in responsibilities between the participant and the responder partner. The remaining twenty items are considered team items. The participants' partner gave the participants opportunities to discover the team items first (i.e., providing time for the participant to find the item when the item was in the participant's vicinity) before the responder pointed out the item. After either the participant or the human/robot partner found an item, the item was categorized as suspicious or non-suspicious. If the item was deemed as suspicious, the participant took a photograph. Additionally, photographs were taken of hazard placards and laser signs, though signs did not fall into the suspicious category.

The participant items report task determines whether the participant found the items that he or she was specifically responsible for during the search. Participant items are those items that were only investigated when found by participants. Primary task failure rate for the participant item report task was calculated by considering a failure when a participant did not find a participant item. The primary task failure rate was calculated by dividing the number of misses by the total of participant items investigated (i.e., six). Five of the participant items were suspicious; one item was non-suspicious.

The team items first report task determined whether the participant found one of the items in the search, that either teammate was able to find, before his or her responder partner found it. The participant was always given opportunities to discover the team items before the responder partner (i.e., human or robot). The primary task failure rate for finding team items first report task was calculated by counting a failure each time that a participant did not find the team items first (i.e., before the responder partner). There were twenty team items; ten team items were suspicious and ten team items were not suspicious.

The suspicious item report task determined whether the participant provided the correct categorization of items as suspicious or not suspicious. Twenty-one of the items found by the participants and responder partners were categorized as either suspicious or not suspicious. The five remaining items were hazard placards and laser signs, which were not categorized. Primary task failure rate for the suspicious item report was calculated by recording a miss when a participant incorrectly categorized an investigated item as either suspicious, when it was truly not suspicious, or vice versa.

The photographs taken task determined whether the participant took the photographs that he or she was responsible for and requested to take by his or her partner. The participant was asked to photograph all of the items deemed suspicious and each of the hazard placards and laser signs. Primary task failure rate for photographs taken was calculated by counting a miss each time that the participant did not photograph an item on the list. The number of misses was divided by twenty (i.e., the number of required photographs) in order to calculate the failure rate. If the participant did not investigate the item with the responder partner

in their specific investigation, the item was not required for that participant (i.e., if the participant did not find Item 1, they are penalized with a failure for not having a photograph of Item 1 and the total number of photographs required for that participant is reduced from 20 to 19).

Overall primary task failure rate was calculated for each participant by summing each of their total task failures from each of the four primary task components. This total number of failures was divided by the total number of that specific participant's total number of required tasks.

The participant items report task is a detail-tracking task. The participant was searching on his or her own to look for an out of place item, given the details present in the environment. The team items first report task is an attention task. The participant is working with the responder partner to find team items, but a higher level of attention in monitoring the surroundings will manifest in a lower failure rate for the participant when finding the team items first. The suspicious item report task is an assessment task that classifies each investigated item as either "suspicious" or "not suspicious." The photographs taken task is an effort task, indicating whether or not participants were correctly following through with instructions to photograph items when asked. The general primary task component types are the same as in the Guided evaluation and the results will be presented using the generic categories for easier comparison between evaluations.

Immediately following the Collaborative evaluation, participants were asked to recall all suspicious items that they investigated and write them down (see Table III.24 for a list of the items and which were considered suspicious). Participants responded with varying levels of detail, which necessitated a method for scoring participant responses along four dimensions: the item's primary attribute, the item's primary location, suspicious components, and global location. Each response from each participant was compared to ground truth information for the recalled suspicious item along these four dimensions and a score between 0 to 1 was given (i.e., maximum score of four for each item recalled).

The primary attribute is the suspicious item's most important and obvious descriptor and/or container; for example, Item 2's primary attribute is that it is a backpack. The primary location is a short description or note to identify the item in the relative location in the hallway or room. For example, Item 2's primary location is on a bench. Each item also had a number of suspicious components represented by the suspicious elements of the item necessary to fully describe it. Item 2's suspicious components include four items: wires, tape, batteries, and C4 putty. Reporting one of the four items earned a participant a score of 1/4. The last dimension is global location that locates the item in the perspective of the entire search area. Item 2's global location includes descriptions of being near the first corner, on the right side of the second hallway, near the lobby area, near the map in the recycling bin, etc. Offering any global location secured the participant a score of one for this dimension. For example, if a participant's response was "backpack with wires and tape in second hall" he or she scored 1 for primary attribute (i.e., backpack), 0 for primary location (i.e., the bench it

Table III.24: The items with categorization by participant item, team item, suspiciousness, and need for being photographed: the four components of primary task failure rate in the Collaborative evaluation.

| Item No. | Item Description | Participant Items | Team Items | Suspicious? | Photograph? |
|---|---|---|---|---|---|
| 1 | Map in trashcan | X | | X | X |
| 2 | Backpack on bench | | X | X | X |
| 3 | Soda bottle with suspicious material in recycling bin | X | | X | X |
| 4 | Math equations written on whiteboard | | X | | |
| 5 | Box of bomb-making supplies on windowsill | | X | X | X |
| 6 | Box of textbooks on floor | | X | | |
| 7 | Hazard placard and laser sign over closed lab door | | X | NA | X |
| 8 | Bomb-making instructions in trashcan in large lab | X | | X | X |
| 9 | Bomb-making materials on counter in large lab | | X | X | X |
| 10 | Printout discussing C4 on bulletin board | X | | X | X |
| 11 | Fire extinguisher out of its case | | X | | |
| 12 | Hazard placard next to closed lab door | | X | NA | X |
| 13 | Hazard placard next to another lab door | | X | NA | X |
| 14 | Bomb underneath eyewash station | | X | X | X |
| 15 | Wires hanging from ceiling | | X | X | X |
| 16 | Map on bulletin board | X | | X | X |
| 17 | Note in emergency door | | X | X | X |
| 18 | Laser sign over closed lab door | | X | NA | X |
| 19 | Laser sign over another closed lab door with a keypad lock | | X | NA | X |
| 20 | Lunchbox under water fountain | | X | | |
| 21 | Box of bomb-making supplies under table in hall | | | X | X |
| 22 | "Rendezvous code green" written on whiteboard | | | X | X |
| 23 | Note on windowsill | | | X | X |
| 24 | Computer cables in the small lab | | X | | |
| 25 | Cleaning supplies in the small lab | X | | | |
| 26 | Unknown experimental equipment in the small lab | | X | X | X |

sits on was not mentioned), 2/4 for suspicious components (i.e., putty and batteries were not mentioned), and 1 for global location (i.e., second hallway). The total score for this item was 2.5 out of a possible 4.

Participants in the Collaborative evaluation sometimes responded to the memory recall task with generic terms (i.e., descriptors that possibly indicate multiple items investigated by the participant). Scores were given for each item that was reflected by the statement, and the result was divided by the total number of items reflected by the statement. For example, if a participant wrote down simply, "tape," this statement reflects a suspicious component of Items 2, 5, and 21. The participant received 0 credit for primary attribute, primary location, and global location for all three items. The participant received a score of $[(1/4)+(1/6)+(1/6)]/3$ for suspicious components; the participant reported one of four suspicious components in Item 2, one of six in Item 5, and one of six in Item 21. The score was divided by three because "tape" was related to three items. The generic term "bomb-making materials" was given the automatic score of 0.5 for suspicious components.

Collaborative evaluation memory recall percentage was calculated by computing a ratio of the recall score divided by the maximum possible recall score for each individual participant. The participant's recall score was computed by summing the earned points along each of the four dimensions for each item and dividing by the participant's maximum possible score. The maximum possible recall score was calculated for each participant by totaling the number of suspicious items found during the evaluation. Suspicious items consisted of items found by the participant and his or her human or robot partner on the list in Table III.24 that were considered suspicious, and any extra suspicious items reported by the participant during the evaluation. The maximum possible score was calculated by taking the number of suspicious items found and multiplying it by four, representing the four scored dimensions of working memory recall.

Memory recall scores for the Collaborative evaluation participants were adjusted for "incorrect" responses (i.e., recalling non-suspicious items). Participants were not penalized for recalling hazard and laser signs, as these may have been ambiguous during the investigation due to the time spent describing them and noting their details for incident command. If a non-suspicious item was recalled by a participant, it was given a score of negative four (i.e., a negative one for each of the four components of recall).

The in situ workload questions were completed (see Chapter III.1.2.4) following the search of each investigation area. Definitions for each workload channel were given during training.

The post trial questionnaire included the same eight statements as the Guided evaluation (see Chapter III.1.2.4), with the addition of ten additional statements. Responses were rated on a Likert scale from 1 (totally disagree) to 5 (totally agree). The additional statements consisted of the following:

9. My teammate helped me identify possible alternatives.

10. I understood the problem.

11. I understood my teammate's concerns (or suggestions).
12. I made contribution to the decision-making process.
13. I felt I was in charge of the team.
14. My teammate led the team in the investigation.
15. I felt I had greater responsibility in the team than my teammate.
16. My teammate asked for help from me when it was needed.
17. I understand my teammate's strengths and abilities.
18. My teammate was collaborative.

The NASA-TLX questionnaire was completed at the end of the entire evaluation; overall and physical workload were analyzed.

#### III.2.2.5 Procedure

Following a demographic survey, participants were informed of the anonymous bomb threat and that the task was to search for out of place items and report anything suspicious to their partner. The participants donned a BioHarness monitor and viewed a three-minute training video explaining how typical searches are executed (e.g., trashcan lids are lifted), what types of items were to be deemed suspicious, and indicating that photographs were to be taken after assessing a suspicious item or hazard sign. Following the video, in situ subjective workload questions were explained to the participant (see Chapter IV for more details). Definitions were provided and the participants were provided an opportunity to ask any questions. The participants were given one minute to memorize a list of chemicals for the secondary task. Upon briefing completion, participants donned the remaining equipment, including a neon reflective vest to indicate that the participant was a part of the investigation, a Garmin footpod pedometer and watch, a Shure microphone headset, and a Looxcie head-mounted video camera (attached to the microphone headset). The participants were provided with a point-and-shoot digital camera and a laser pointer. The microphone headset was used to record the participants' speech. Participants were instructed to use the laser pointer to indicate what they were investigating (to be recorded by the head-mounted camera). The point-and-shoot camera captured images of potentially suspicious items and hazard placards.

Participants were responsible for checking bulletin boards and trashcans, while the partner (either human or robot) checked fire extinguishers and monitored air quality. Periodically, the responder partner informed the participant of air sample readings. The participants were informed of their specific duties and the duties of

their partners during an introductory speech from their partners. It is common to collect air sample readings while working in the field, and our experimental scenario incorporated the measurement of methane that was found to increase near a research laboratory as was located in investigation area 4 (See Figure III.28).

The responder halted the immediate investigation area search when the area border was reached, as shown in Figure III.28. The responder first verified with the participant whether or not any additional items required investigation and then proceeded to ask the participant to provide the in situ workload ratings. Two secondary task questions were posed during each of the six areas for a total of 12. After all six investigation areas were completed, participants responded to a post-trial questionnaire and the NASA-TLX mental workload questionnaire.

### III.2.3 Results

Overall and mental workload were analyzed in Chapter III.2.3.1, physical workload was analyzed in Chapter III.2.3.2 and reaction and response time were analyzed in Chapter III.2.3.3.

#### III.2.3.1 Overall and Mental Workload

Evaluation results were not normally distributed; thus, nonparametric analysis techniques were employed.

**Physiological Measures**

Higher levels of low-frequency heart rate variability have been shown to correspond with lower mental workload (Aasman et al., 1987; Castor, 2003; Roscoe, 1992). The mean low-frequency heart rate variability for the H-H condition was 402431 (St. Dev. = 1081004) $ms^2$ and 249469 (St. Dev. = 1025228) $ms^2$ for the H-R condition. There were no significant effects of condition or investigation index. This result does not correspond with the model's prediction of lower mental workload and expected significantly higher heart rate variability for the H-R condition.

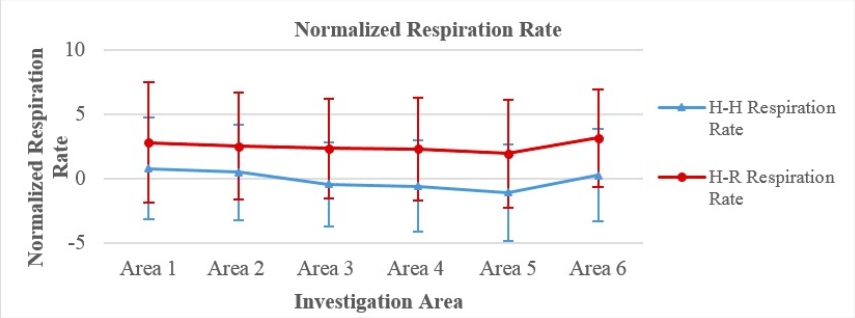
The mean normalized heart rate for the H-H condition was -3.15 (St. Dev. = 26.95) beats per minute and 3.83 (St. Dev. = 10.69) beats per minute for the H-R condition. H-H participants tended to have negative normalized heart rate values, while the H-R participants tended to have positive values. There was no main effect of condition or investigation index on normalized heart rate.

The mean normalized respiration rate for the H-H condition participants was -0.07 (St. Dev. = 3.62) breaths per minute and 2.54 (St. Dev. = 4.03) breaths per minute for the H-R condition. Figure III.30 provides the mean normalized respiration rate experienced by condition and investigation area. Respiration rate tends to decrease when experiencing high levels of mental workload (Roscoe, 1992); thus, the results indicate that the H-R condition resulted in lower mental workload levels than the H-H condition. A Kruskal-Wallis test

Table III.25: Mean normalized respiration rate by condition and investigation index.

| Investigation Index | Across Both Conditions | H-H | H-R |
|---|---|---|---|
| Low | 1.40 (4.08) | 0.20 (3.62) | 2.60 (4.18) |
| Medium | 0.99 (4.13) | -0.27 (3.77) | 2.25 (4.14) |
| High | 1.31 (3.96) | -0.14 (3.56) | 2.76 (3.86) |

Figure III.30: Mean normalized respiration rate by condition and investigation area.




indicated that the mean normalized respiration rate was significantly higher during the H-R condition, $\chi^2(1)$ = 21.45, p $<$ 0.01. The mean normalized respiration rates by condition and investigation index are presented in Table III.25. A Kruskal-Wallis test found no significant main effect of investigation index. A third Kruskal-Wallis test identified a significant interaction effect of condition and investigation index, $\chi^2(5)$ = 22.13, p $<$ 0.01. Mann-Whitney U tests with a Bonferroni adjustment found that H-H participants had a significantly lower normalized respiration rate for the high investigation index area than did the H-R participants, U = 380, Z = -3.02, p $<$ 0.01. The model results predicted lower mental workload in the H-R condition, which is supported by these results. Additionally, the results align with the claim that negative correlation happens between mental workload and respiration rate (Keller et al., 2001; Roscoe, 1992).

### Distance Traveled

A pedometer was used to measure distance traveled. The pedometer attached to the laces of the shoe; however, participants did not always wear athletic shoes while completing the evaluation, as the system is intended for wear. The footpod was attached using small zip ties when laces were not available on the shoe. The type of shoe that the participants wore was not recorded, so the impact on results is not known. Additionally, the Garmin pedometer system is generally meant for use during high levels of physical activity, and not for the short bursts of low level activity (e.g., walking 10 feet, standing still 2 minutes, walking 50 feet). Usable pedometer data was collected from eleven H-H participants and nine H-R participants. The remaining sixteen participants' walking data was not successfully recorded by the Garmin pedometer.

Mean distance traveled during each investigation area by the H-H condition participants was 14.76 (St.

Table III.26: Total number of correct (C) and incorrect (I) responses to On List questions, by investigation index and condition.

| Investigation Index | Both C | Both I | H-H C | H-H I | H-R C | H-R I |
|---|---|---|---|---|---|---|
| **Low** | 119 | 21 | 55 | 15 | 64 | 6 |
| **Medium** | 111 | 28 | 61 | 8 | 50 | 20 |
| **High** | 98 | 42 | 56 | 14 | 42 | 28 |

Dev. = 13.05) meters (m) and was 18.94 (St. Dev. = 19.20) m in the H-R condition. There was a significant interaction effect between investigation area and condition on distance traveled, $\chi^2(8) = 34.27, p < 0.001$, but none of the results for the post-hoc pairwise Wilcoxon comparisons were significant between conditions.

Total distance traveled by participant was also analyzed. The mean total distance traveled in the H-H condition was 77.85 (St. Dev. = 39.36) m and was 111.25 (St. Dev. = 43.50) m in the H-R condition. There was no significant difference, but the results support the trend that H-R participants may move more.

**Secondary Task Questions**

A total of 172 correct and 37 incorrect responses to the secondary task questions were provided during the H-H condition. The H-R condition participants provided 156 correct and 54 incorrect responses. The number of responses differs across conditions, because six questions were not asked during the H-H condition and four questions were not asked in the H-R condition. A Pearson's Chi-squared test with Yates' continuity correction found no significant main effect of condition for the On List metric.

The number of correct and incorrect On List responses by condition and investigation index are provided in Table III.26. The On List question responses, independent of condition, were analyzed by investigation index. A Chi-squared test indicated a significant main effect of investigation index on the number of correct responses, $\chi^2(2) = 9.57, p < 0.01$. Mann-Whitney comparisons were performed with a Bonferroni adjustment for family-wise error. The number of correct responses for the low investigation index was significantly higher than that for the high investigation index, $U = 11573, Z = -2.930, p < 0.01$. No other comparisons were significant.

A Pearson's Chi-squared test with Yates' continuity correction found a significant interaction effect of condition and investigation index for the correct On List responses, $\chi^2(11) = 152.28, p < 0.01$. The Mann-Whitney pairwise comparisons with Bonferroni family-wise adjustments found that the H-R condition low investigation index responses were correct significantly more often than those for the high investigation index areas in the H-R condition, $U = 2730, Z = -3.94, p < 0.01$. No other comparisons were significant.

The On List Correct results demonstrate that mental workload reserves were not significantly different

Table III.27: Total number of correct (C) and incorrect (I) responses to Danger Level questions, by investigation index and condition.

| Investigation Index | Both C | Both I | H-H C | H-H I | H-R C | H-R I |
|---|---|---|---|---|---|---|
| **Low** | 66 | 26 | 31 | 11 | 35 | 15 |
| **Medium** | 43 | 18 | 20 | 9 | 23 | 9 |
| **High** | 19 | 13 | 11 | 2 | 8 | 2 |

between conditions. Additionally, the number of correct responses was significantly lower in high investigation areas than in the low investigation index areas. While this result offers evidence that mental workload reserve levels were manipulated by investigation index, the data does not fully support this claim. The result matches the model result, which demonstrated no significant effect of investigation index. The H-R condition participants were dramatically affected by the changes in investigation index, while the H-H condition participants' correct responses were not significantly impacted by the investigation index.

The secondary task questions incorporate a second question requiring the participants to provide the danger level associated with a chemical when the participants indicated that a chemical was on the provided list (whether the response was correct or not). The number of correct and incorrect responses to the Danger Level questions, by condition and investigation index are provided in Table III.27. 101 Danger Level questions were asked during the H-H condition and 132 were asked in the H-R condition. This sum includes Danger Level questions asked in response to an incorrect response to the On List Correct question. Danger Level results were only analyzed by condition and investigation index for the Danger Level questions asked with a possible correct answer (i.e., the chemical was on the list and had an associated danger level). There were 84 total Danger Level questions in the H-H condition and 92 Danger Level questions in the H-R condition with a possibility for correct responses. The H-H condition resulted in 62 correct and 22 incorrect responses, while the H-R condition participants had 66 correct and 26 incorrect responses. The effect of the Pearson's Chi-squared test with Yates' continuity correction found no significant main effect of condition or investigation index on the Danger Level Correct metric.

Overall, there was no significant main effect of condition on the number of correct responses to the Danger Level questions. Both H-H and H-R condition participants answered the secondary task questions correctly with similar frequency, showing that the two conditions resulted in similar levels of spare mental capacity.

#### Secondary Task Failure Rate

The Collaborative evaluation secondary task failure rate responses were analyzed by condition and investigation index. Non-parametric analysis was used because the results were not normally distributed. The mean for

Table III.28: Collaborative evaluation secondary task failure rate by condition and investigation index

| Investigation Index | H-H | H-R |
|---|---|---|
| Low | 26/72 | 22/72 |
| Medium | 20/72 | 31/72 |
| High | 14/72 | 29/72 |
| Total | 60/216 | 82/216 |

the H-H condition was 28.24% (St.Dev. = 14. 74%). The mean for the H-R condition was 37.04% (St.Dev. = 20.45%). A Kruskal Wallis test showed no significant difference between conditions. The secondary task failure rate was also analyzed by investigation index. Table III.28 provides the number of incorrect responses over the total number of prompts by investigation index and condition. There was no significant difference in secondary task failure rate between investigation indexes.

**Task Performance**

Task performance was based on the number of items assessed by participants, including both the items in Table III.23 and any extra items. The mean number of items from Table III.23 assessed by the H-H condition participants was 24.11 (St. Dev. = 0.81), while the H-R condition participants assessed an average of 24.33 items (St. Dev. = 0.82). A Kruskal-Wallis test found no significant effect of condition for the number items found.

Some participants identified additional items that they suspected may be related to the bomb threat. The H-H participants found a mean of 2.28 extra items (St. Dev. = 1.73), while the H-R participants found a mean of 1.44 extra items (St. Dev. = 1.21). A Kruskal-Wallis test found no significant effect of condition. This result shows that participants did not respond significantly differently by condition to extra items in the environment and neither the human nor robot partner encouraged participants to find extra items at a significantly higher rate. Overall, there was no significant difference in task performance by condition.

**Primary Task Failure Rate**

The means for each of the four primary task failure rate components for the Collaborative evaluation are provided in Figure III.31. The overall primary task failure rate mean is also provided. A series of Shapiro-Wilk tests indicated that the collected data was not normally distributed; thus, nonparametric Kruskal-Wallis tests are used. The attention primary task failure rate in the H-H condition was 65.83% (St.Dev. = 11.58%) and was 61.86% (St.Dev. = 14.60%) in the H-R condition. A Kruskal-Wallis test indicated no significant difference between the two conditions. The effort primary task failure rate for the H-H condition participants was 8.03% (St.Dev. = 8.11%) and was 17.82% (St.Dev. = 12.78%) for the H-R condition participants.

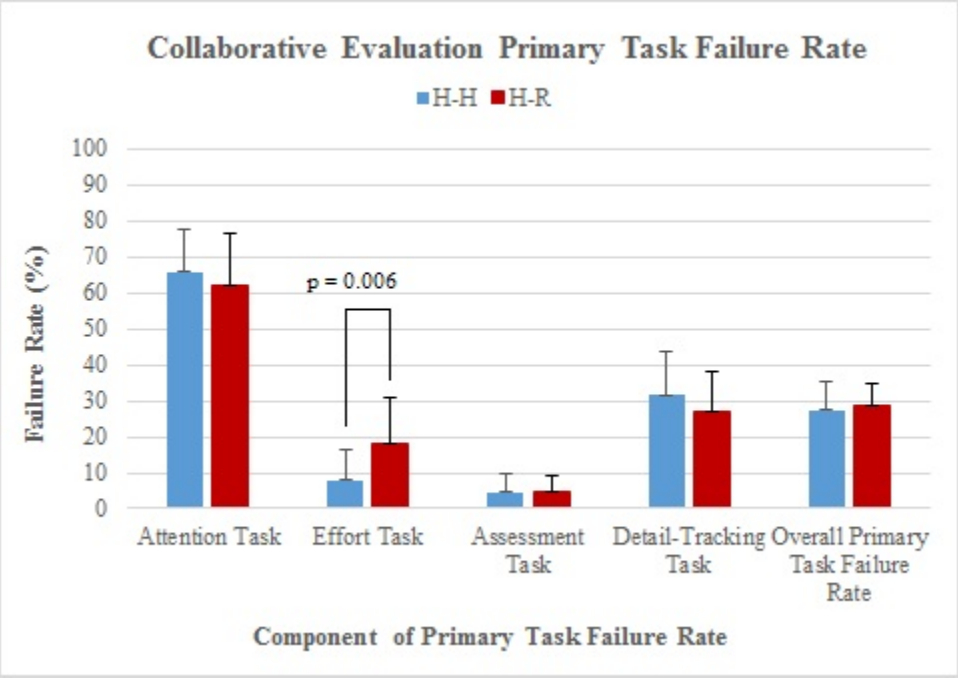


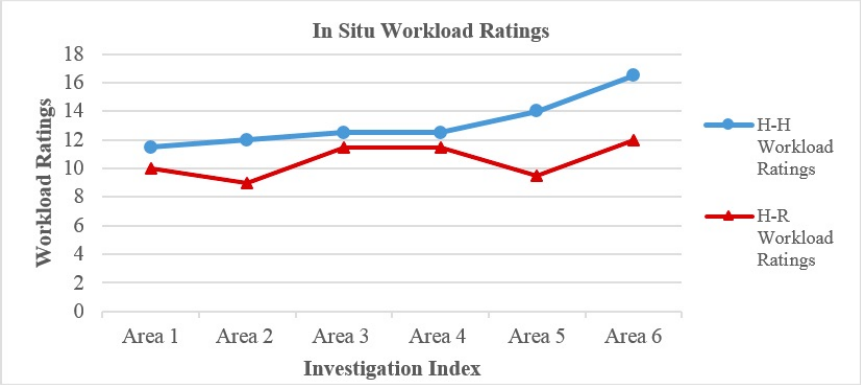

Figure III.31: Means for the four components of the Collaborative evaluation primary task failure rate with overall task failure rate, by condition. Error bars represent one standard deviation above the mean. Significant differences are represented with brackets and associated p-values.

A Kruskal-Wallis test indicated that the failure rate was significantly higher in the H-R condition, $\chi^2(1) = 7.536$, p = 0.006. Participants in the H-R condition missed taking the assigned photographs more frequently. The assessment report primary task failure rate for the H-H condition participants was 4.54% (St.Dev. = 4.97%) and was 4.97% (St.Dev. = 4.49%) for H-R condition participants. A Kruskal-Wallis test indicated no significant difference between the two conditions. The mean detail-tracking primary task failure rate in the H-H condition was 31.48% (St.Dev. = 12.28%) and was 26.85% (St.Dev. = 11.30%) in the H-R condition. A Kruskal-Wallis test indicated no significant difference between the two conditions. The mean overall primary task failure rate for the H-H condition participants was 27.34% (St.Dev. = 4.16%). The mean overall primary task failure rate for the H-R condition participants was 28.47% (St.Dev. = 6.42%). A Kruskal-Wallis test indicated no significant difference between conditions.

**Memory Recall**

Collaborative evaluation working memory recall responses were analyzed by condition using non-parametric analysis as well. The median number of suspicious items found (i.e., the denominator of the memory recall ratio) in both the H-H and H-R conditions was 16. The minimum number of items found in both the H-H and H-R condition was 13 and the maximum was 21 items found in the H-H condition and 19 items in the H-R

Figure III.32: The median in situ workload ratings by condition and investigation area.




condition. A Kruskal Wallis test indicated no significant difference between the two conditions in the number of suspicious items found during the investigation. This information is provided in order to demonstrate that the participants investigated a similar number of suspicious items in both conditions. Additionally, five participants recalled one item each that was non-suspicious; three participants recalled a non-suspicious item in the H-H condition and two participants recalled a non-suspicious item in the H-R condition. The H-H condition mean memory recall percentage was 40.68% (St.Dev. = 15.41%) and the H-R condition memory recall was 49.25% (St.Dev. = 12.84%). A Kruskal Wallis test indicated that the H-R condition memory recall was nearly significantly higher than the H-H condition recall, $\chi^2(1) = 3.25$, $p = 0.071$.

**In Situ Subjective Workload Ratings**

The median total mental workload rated by the H-H participants was 13, while the H-R participants' median was 11. The median, minimum, and maximum in situ mental workload ratings for each investigation index in both conditions are provided in Table III.29. A Kruskal-Wallis test showed that the H-H participant ratings were significantly higher than the H-R condition ratings, $\chi^2(1) = 18.64$, $p < 0.01$. A second Kruskal-Wallis test indicated no significant effect of investigation index on total in situ subjective workload ratings. Figure III.32 provides the median in situ mental workload rating values for each investigation area.

A Kruskal-Wallis test indicated a significant interaction effect of investigation index and condition, $\chi^2(5) = 22.22$, $p < 0.01$. Mann-Whitney U tests with Bonferroni adjustments determined that none of the results were significantly different between investigation indices for either condition. Generally speaking, the H-H condition participants rated their mental workload higher than the H-R participants.

Table III.29: In situ subjective mental workload ratings by condition and investigation index. Med = median, Min = minimum, and Max = maximum.

| Investigation Index | Across Both Conditions | | | H-H | | | H-R | | |
|---|---|---|---|---|---|---|---|---|---|
| | Med. | Min. | Max. | Med. | Min. | Max. | Med. | Min. | Max. |
| Low | 11.5 | 6.0 | 29.0 | 12.0 | 7.0 | 29.0 | 11.0 | 6.0 | 23.0 |
| Medium | 11.5 | 6.0 | 29.0 | 12.0 | 7.0 | 29.0 | 11.0 | 6.0 | 23.0 |
| High | 11.5 | 6.0 | 29.0 | 12.0 | 7.0 | 29.0 | 11.0 | 6.0 | 23.0 |

**Post-Trial Questionnaire**

Analysis of the post-trial questionnaire for the Collaborative evaluation was performed using non-parametric analysis because of the Likert scale ratings provided. Median agreement responses for each of the eighteen questions are provided in Table III.30. A Kruskal-Wallis test found a significant interaction effect of the question number and condition $\chi^2(35) = 342.41$, $p < 0.001$; however, the follow-up pairwise Wilcoxon tests with Bonferonni corrections for significance indicated no significant differences between conditions for any individual questions.

**NASA-TLX Responses**

The mean total NASA-TLX score for the H-H condition was 41.32 (St. Dev. = 17.80) and 30.89 (St. Dev. = 15.12) for the H-R condition. A two-sided t-test found no significant difference ($p = 0.08$), but the H-H responses, across the NASA-TLX components, tended to be higher.

**Correlations Analysis**

Pearson's product-moment correlations were performed to correlate physiological data with the in situ mental workload ratings. Heart rate variability and normalized heart rate were not significantly affected by either condition or investigation index; thus, they were not evaluated. Normalized respiration rate was significantly higher in the H-R condition ($p < 0.01$), notably in areas with high investigation indices. There was a significant negative correlation between normalized respiration rate and the in situ mental workload ratings, $r(214) = -0.17$, $p = 0.01$. This result indicates that when in situ mental workload ratings were high and normalized respiration rate was low; thus, normalized respiration rate can be used to analyze mental workload levels.

Subtask time was also correlated to in situ mental workload ratings to determine whether higher mental workload ratings in the H-H condition were due to the shorter time frame in which H-H condition participants completed tasks. There was no significant correlation.

Table III.30: Descriptive statistics for Collaborative evaluation post-trial questionnaire data.

| Statement Number | Statement | H-H | | H-R | |
|---|---|---|---|---|---|
| | | Median | Range | Median | Range |
| 1 | My teammate gave me clear instructions. | 5 | 4-5 | 5 | 3-5 |
| 2 | I trusted my teammate. | 5 | 3-5 | 5 | 2-5 |
| 3 | I felt comfortable communicating with my teammate. | 5 | 4-5 | 4 | 2-5 |
| 4 | My teammate understood what I was trying to communicate. | 5 | 4-5 | 4 | 3-5 |
| 5 | I did a good job on the tasks I was assigned. | 4 | 2-5 | 4 | 3-5 |
| 6 | I often felt confused about my teammate's instructions. | 1 | 1-3 | 2 | 1-5 |
| 7 | I often felt confused as to the purpose of my actions. | 2 | 1-4 | 1.5 | 1-4 |
| 8 | I felt stressed during the scenario. | 2 | 1-4 | 2 | 1-4 |
| 9 | My teammate helped me identify possible alternatives. | 4 | 3-5 | 4 | 3-5 |
| 10 | I understood the problem. | 5 | 3-5 | 5 | 3-5 |
| 11 | I understood my teammate's concerns (or suggestions). | 4 | 2-5 | 4 | 3-5 |
| 12 | I made contribution to the decision-making process. | 5 | 3-5 | 4 | 2-5 |
| 13 | I felt I was in charge of the team. | 3 | 2-3 | 2 | 1-5 |
| 14 | My teammate led the team in the investigation. | 4 | 2-5 | 4 | 2-5 |
| 15 | I felt I had greater responsibility in the team than my teammate. | 3 | 2-4 | 2.5 | 1-4 |
| 16 | My teammate asked for help from me when it was needed. | 4 | 3-5 | 4 | 4-5 |
| 17 | I understand my teammate's strengths and abilities. | 4 | 2-5 | 4 | 3-5 |
| 18 | My teammate was collaborative. | 4 | 3-5 | 4 | 2-5 |

**Comparison of Modeled Mental Workload and In Situ Subjective Workload Ratings**

The total mental workload is calculated based upon the same total mental workload scale for both the model results and the in situ mental workload ratings. Figure III.33 compares the in situ mental workload ratings with the modeled mental workload predictions. The H-H condition's mean total subjective mental workload rating was 13.99 (St. Dev. = 5.24) and the H-R condition's result was 10.97 (St. Dev. = 3.98). The mean total mental workload from the H-H Model was 13.18 (St. Dev. = 0.58), while the H-R model resulted in a mean of 12.96 (St. Dev. = 0.47). The mean total modeled mental workload was 13.07 (St. Dev. = 0.53) across both models, while the mean total subjective mental workload rating was 12.48 (St. Dev. = 4.88) across both conditions. A Kruskal-Wallis test indicated that the modeled mental workload across both conditions was significantly higher than the subjectively rated mental workload across both conditions, $\chi^2(1) = 15.08$, $p < 0.01$.

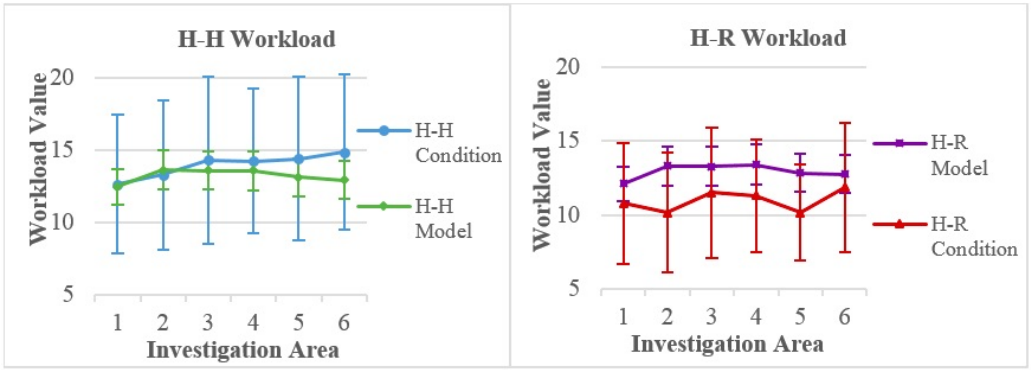
Mental workload was compared across the model results and the subjective ratings for the four data sets: the H-H condition in situ ratings, the H-R condition in situ ratings, the H-H model and the H-R model. A Kruskal-Wallis test indicated a significant main effect of data set, $\chi^2(3) = 38.89$, $p < 0.01$. A pairwise Wilcoxon rank sum test with Holms p-value correction indicated that the H-H condition subjective ratings were significantly higher than the H-R condition subjective ratings ($p < 0.01$). As well, the H-H model mental workload was significantly higher than the H-R model mental workload ($p < 0.01$). The H-H condition subjective ratings and the H-H model mental workload were not significantly different. Finally, the H-R condition subjective ratings were significantly lower than the H-R model mental workload ($p < 0.01$). The H-R model overestimated participants' mental workload.

Overall, these results indicate that the H-H modeled mental workload was in-line with the H-H in situ mental workload ratings. Thus, further analysis is required to determine how closely the two data sets match. Ninety-five percent confidence intervals were computed for each of the model data points. If the H-H condition evaluation results fall into these windows, then the model can be considered a good fit (Roberts and Pashler, 2000). Table III.31 presents the results for both conditions. Four of the six the H-H investigation area mental workload predictions were good fits, whereas only one value was a good fit in the H-R condition, Area 6.

**Calculation of Task Density**

The overall task-density ratio was calculated for each condition. The H-H condition had a mean overall task density of 0.26 (St. Dev. = 0.09), while the H-R task density was 0.20 (St. Dev. = 0.07). A Kruskal-Wallis test found that the H-H condition ratio was significantly higher, $\chi^2(1) = 27.62$, $p < 0.01$. This result indicates that the longer task times in the H-R condition lower the task density and result in lower mental workload

Figure III.33: (a) H-H model predictions and H-H condition in situ subjective workload ratings; (b) H-R model predictions and H-R condition in situ subjective workload ratings.

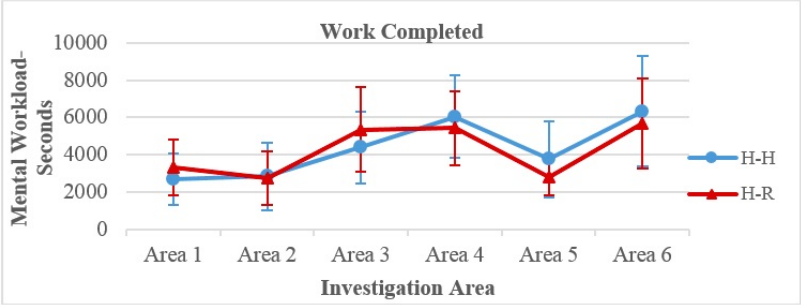



a. b.

Table III.31: Confidence interval (C.I.) results of the model and evaluation results.

| | Area 1 | | Area 2 | | Area 3 | | Area 4 | | Area 5 | | Area 6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | H-H | H-R | H-H | H-R | H-H | H-R | H-H | H-R | H-H | H-R | H-H | H-R |
| Model Mean | 12.6 | 12.1 | 14.1 | 13.3 | 13.8 | 13.3 | 13.7 | 13.4 | 13.1 | 12.8 | 13.1 | 12.8 |
| Model St. Dev. | 1.24 | 1.20 | 1.34 | 1.31 | 1.34 | 1.32 | 1.34 | 1.34 | 1.31 | 1.28 | 1.29 | 1.26 |
| C.I. Low Cutoff | 11.7 | 11.3 | 13.2 | 12.4 | 12.8 | 12.4 | 12.8 | 12.5 | 12.2 | 11.9 | 12.2 | 11.9 |
| C.I. High Cutoff | 13.5 | 12.9 | 15.1 | 14.3 | 14.7 | 14.2 | 14.7 | 14.4 | 14.1 | 13.8 | 14.0 | 13.7 |
| Evaluation Mean | 12.7 | 10.8 | 13.3 | 10.2 | 14.3 | 11.5 | 14.3 | 11.3 | 14.4 | 10.2 | 14.9 | 11.9 |
| Within C.I.? | Yes | No | Yes | No | Yes | No | Yes | No | No | No | No | Yes |
| Within St. Dev. of Model? | Yes | No | Yes | No | Yes | No | Yes | No | Yes | No | No | Yes |

Figure III.34: Amount of work completed by investigation area and condition.




levels. This evaluation result also echoes the model's result of lower task density for the H-R condition.

Task-density ratios were calculated by investigation area. The mean task density in low investigation areas was 0.24 (St. Dev. = 0.07), 0.23 (St. Dev. = 0.08) in medium investigation areas, and 0.23 (St. Dev. = 0.10) in high investigation areas. A Kruskal-Wallis test found no significant difference in the task-density proportion by investigation index; therefore, there was no difference in mental workload by investigation index. The model results also showed no significant main effect by investigation index.

**Work Completed**

The amount of work completed reflects the degree to which participants experienced similar levels of work based on the proportion of mental workload and subtask time. The amount of work completed by H-H participants was 4344.23 (St. Dev. = 2512.33) mental workload seconds and 4226.55 (St. Dev. = 2205.72) mental workload seconds for H-R participants. There was no significant difference between these two values. The amount of work completed by investigation area and condition is presented in Figure III.34. The values between the conditions are very similar for the individual investigation areas; thus, both conditions resulted in similar levels of work.

### III.2.3.2 Physical Workload

As with the Guided evaluation results presentation (see Chapter III.1.3.2), analysis of the time spent in each Investigation Area is presented first in order to ground the following physical workload discussion, which includes objective and subjective results. The timing and physiological measurements were not normally distributed; therefore, nonparametric analysis is used.

**Subtask Time**

The mean subtask time for the H-H condition was 305.11 (St. Dev. = 114.21) sec, while the average time in the H-R Condition was 386.65 (St. Dev. = 169.97) sec. A Kruskal-Wallis test indicated that H-R Condition

Figure III.35: Mean subtask time spent in each investigation area by condition.

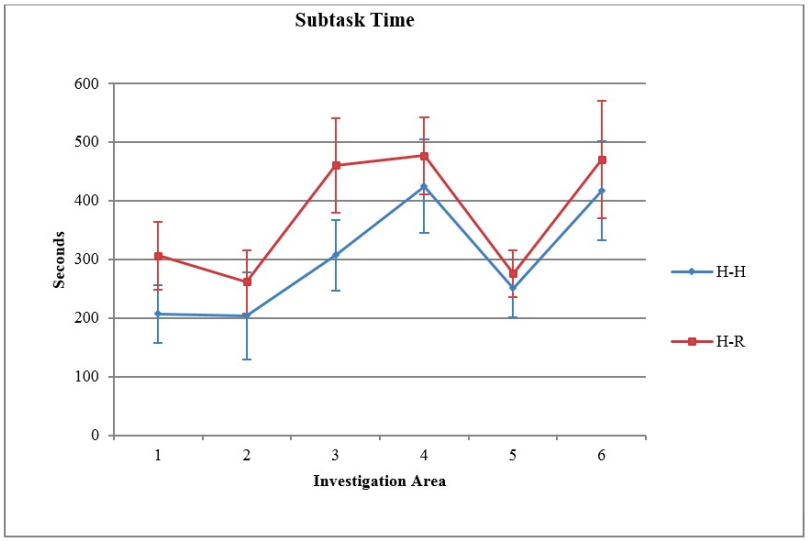



Table III.32: Subtask time by condition and investigation index.

| Investigation Index | Subtask Time | |
|---|---|---|
| | H-H | H-R |
| Low | 260.75 (78.02) | 383.28 (104.50) |
| Medium | 230.28 (67.79) | 269.83 (46.10) |
| High | 424.31 (82.43) | 506.83 (215.33) |
| Overall | 305.11 (114.21) | 386.65 (169.97) |

participants took significantly longer to investigate areas, $\chi^2(1) = 18.08$, $p < 0.001$. A Cohen's d statistic indicated a medium effect size (d = 0.6) of condition on subtask time, indicating that 38.2% of the two conditions' distributions did not overlap. The mean subtask time for each investigation area is presented in Figure III.35.

A Kruskal-Wallis test revealed a significant main effect of investigation index on subtask time, $\chi^2(2) = 102.75$, $p < 0.001$. Descriptive statistics for subtask time by investigation index and condition are presented in Table III.32. A series of Mann Whitney U tests showed that all three mean subtask times were significantly different from one another. The low investigation index mean time were significantly higher than the medium investigation indexes, U = 3604.5, Z = 4.05, $p < 0.001$. Mean subtask times were significantly higher for high investigation indexes than the medium investigation index areas, U = 164.5, Z = -9.70, $p < 0.001$ and the low investigation index area times, U = 939, Z = -6.07, $p < 0.001$.

Overall, subtask times were longer in the H-R condition. The main effect of Investigation Index on subtask time reflected that High Investigation Indexed areas took longest, as there was the largest amount of items present.

Table III.33: Descriptive statistics for vector magnitude and variance in posture by condition and investigation index.

| | | Investigation Index | | | Overall |
|---|---|---|---|---|---|
| | | Low | Medium | High | |
| **Vector Magnitude** | H-H | 68.6 (29.8) | 73.1 (31.9) | 75.7 (23.2) | 72.5 (28.4) |
| | H-R | 96.0 (27.2) | 90.0 (27.6) | 83.0 (27.9) | 89.7 (27.8) |
| **Variance in Posture** | H-H | 237.49 (286.12) | 300.30 (340.23) | 245.62 (228.43) | 261.13 (287.23) |
| | H-R | 477.49 (463.54) | 339.65 (414.47) | 279.35 (202.16) | 365.50 (383.12) |

**Other Objective Measures**

The descriptive statistics for vector magnitude and variance in posture are presented in Table III.33. As in Chapter III.1.3.2 vector magnitude is presented as $VMUx10^3$ for ease of reading, and variance in posture is presented in degrees squared.

The mean vector magnitude was 72.5 (St. Dev. = 28.4) $VMUx10^{-3}$ for participants in the H-H Condition, while participants in the H-R Condition had a mean vector magnitude of 89.7 (St. Dev. = 27.8) $VMUx10^{-3}$. The mean vector magnitude during each investigation area is presented in Figure III.36. The figure shows that the H-R data is higher than the H-H data, but it does not directly correspond to the trend seen in subtask time (Figure III.35). A Pearson's product-moment correlation assessed the relationship between subtask time and vector magnitude, but the correlation was not significant in either condition. This result shows that the increases in mean vector magnitude are not due to longer investigation times.

The mean variance in posture was 261.13 (St. Dev. = 287.23) degrees squared for participants in the H-H condition and was 365.50 (St. Dev. = 383.12) degrees squared for H-R condition participants. Figure III.37 provides the mean variance in posture during each investigation area. A Pearson's product-moment correlation between variance in posture and subtask time was not significant for either condition. This result shows that the increases in mean posture variance are not due to longer investigation times.

The main effects of independent variables condition and investigation index were tested on the physiological measures of vector magnitude and variance in posture. The presented objective measures do not meet the assumptions of analysis of variance. One-way Kruskal Wallis tests indicated that both vector magnitude ($\chi^2(1) = 20.85$, $p < 0.001$) and posture variance ($\chi^2(1) = 5.685$, $p = 0.017$) were significantly higher for participants in the H-R condition. Cohen's d statistics showed a medium effect size of vector magnitude (d = 0.6) with 38.2% non-overlap in the distributions and a small effect size for variance in posture (d = 0.3), with 21.3% non-overlap. Two additional Kruskal Wallis tests indicated that there were no significant main effects of investigation index on either vector magnitude or variance in posture.

The posture skewness of the distribution of H-H condition participants' posture values was 1.03 and was 0.63 for H-R condition participants. The positive skewness values in both conditions indicate that the distri-

Figure III.36: Mean vector magnitude during each investigation area by condition.

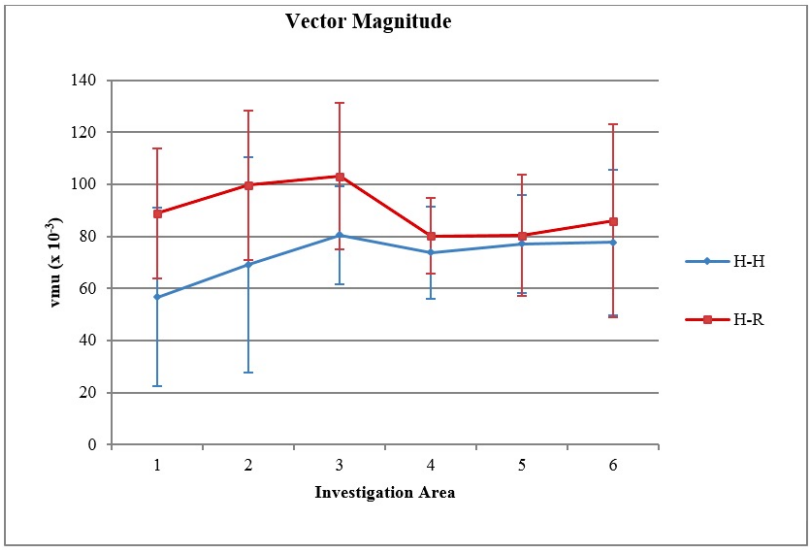
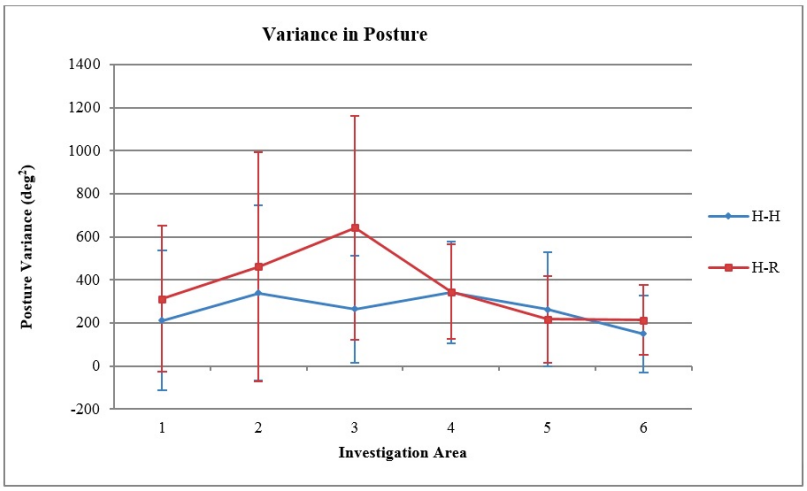



Figure III.37: Mean variance in posture during each investigation area by condition.



bution of posture values is skewed by having a higher concentration of smaller posture values. A skewness value closer to zero indicates a trend towards equal distribution of posture values. The larger skewness for H-H condition participants suggests that H-H condition participants had fewer instances of leaning far forward. Skewness was also calculated for each participant over their mean posture values in each Investigation Area. Mean skewness by participant in the H-H condition was 0.66 (St. Dev. = 0.81) and was 0.68 (St. Dev. = 0.90) in the H-R condition. None of these skewness values were significantly different. The overall skewness values imply that H-H condition participants leaned forward less often, but mean skewness values by participant were not significantly different between conditions.

The posture kurtosis of the raw posture values was 3.93 for the H-H condition and 0.06 for the H-R condition. The lower value in the H-R condition indicates that H-R condition participants had a wider distribution of raw posture values. The mean kurtosis of raw posture values from each Investigation Area was calculated for each participant. The mean kurtosis value from the H-H condition participants was 0.09 (St. Dev. = 2.26) and was 0.59 (St. Dev. = 2.15) for H-R condition participants. These kurtosis values were not significantly different. The smaller mean value in the H-H condition implies that participants tended to have a wider distribution of postures and H-R condition participants each had a peaked posture distribution about a distinct mean, but as the overall kurtosis was smaller than the H-H condition kurtosis value, this implies that H-R condition participants had more individual differences.

**Subjective Measures**

The median subjective workload ratings for the motor and tactile channels are presented in Table III.34 by condition and investigation index. As in the Guided evaluation, the Likert-scaled subjective workload rating data is not normally distributed, therefore nonparametric analysis is used.

The median H-H condition motor subjective in situ workload ratings are the same as or higher than the H-R condition ratings for each of the six investigation areas. The median motor in situ subjective workload rating for the H-H condition was 2 and was 1 for the H-R condition. The sum of the motor subjective workload ratings in the H-H condition was 223 and was 153 for the H-R condition. The correlation between area investigation time and motor workload ratings was tested via Pearson's product-moment test. There was no significant result, thus motor workload ratings did not change significantly with changes in investigation time.

The tactile subjective in situ workload ratings followed the same trend as motor workload ratings: the median H-H condition in situ ratings are the same as or higher than the H-R condition ratings in each investigation area. The median tactile in situ subjective workload rating was 1 for each condition. The sum of the tactile subjective workload ratings in the H-H condition was 201 and was 147 for the H-R condition. There

Table III.34: Median subjective workload ratings for motor and tactile workload by investigation index and condition.

| Investigation Area | Condition | Workload Channel | |
|---|---|---|---|
| | | Motor | Tactile |
| Area 1 | H-H | 1 | 1 |
| | H-R | 1 | 1 |
| Area 2 | H-H | 2 | 2 |
| | H-R | 1 | 1 |
| Area 3 | H-H | 2 | 2 |
| | H-R | 1 | 1 |
| Area 4 | H-H | 2 | 1.5 |
| | H-R | 1.5 | 1 |
| Area 5 | H-H | 2 | 2 |
| | H-R | 1 | 1 |
| Area 6 | H-H | 2 | 1.5 |
| | H-R | 1 | 1 |

was no significant correlation between the tactile workload ratings and area investigation time, as evidenced by a Pearson's product-moment correlation test. Tactile workload ratings were not significantly dependent on the length of time it took to investigate each area.

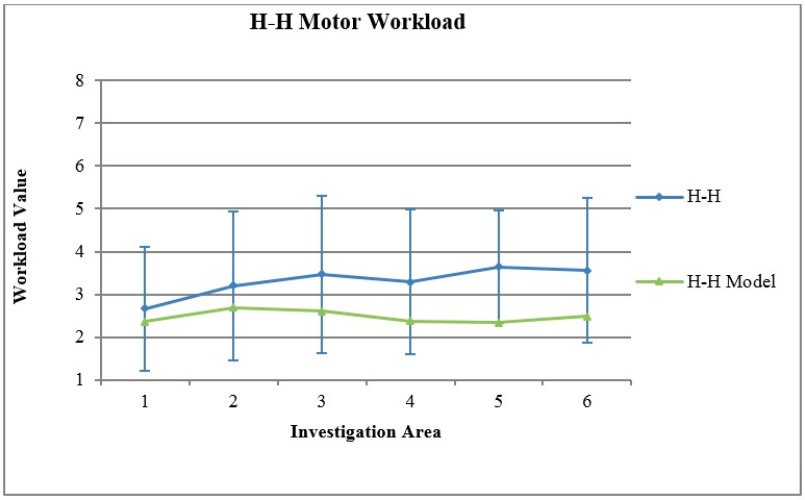
The main effects of condition and investigation index were tested for the motor and tactile subjective workload ratings. The presented subjective measures do not meet the assumptions of analysis of variance. A Kruskal Wallis tests indicated that H-H condition motor workload ratings ($\chi^2(1) = 26.41$, $p < 0.001$) and tactile workload ratings ($\chi^2(1) = 9.92$, $p = 0.002$) were significantly higher than H-R condition ratings. There was a large Cohen's d statistic effect size for the motor ratings (d = 0.8) with 47.4% non-overlap between the distributions and a medium effect size for the tactile subjective workload ratings (d = 0.6) with 38.2% of non-overlap. Kruskal-Wallis tests indicated that there was no significant effect of investigation index on either motor or tactile workload ratings. This result indicates that the investigation index measure may not be a successful manipulation of workload.

The NASA-TLX physical demand responses resulted in a mean weighted demand for H-H condition participants of 1.97 (St. Dev. = 2.55) and was 0.70 (St. Dev. = 0.94) for the H-R condition participants. A t-test indicated no significant difference between the two conditions.

**Comparison with Physical Workload Model Results**

The modeled workload was compared to the rescaled participant motor and tactile ratings, presented in Figure III.38. Five of the six model data points are within one standard deviation of the H-H participant ratings. The H-H model follows a similar trend as the subjective ratings. The mean delta between the H-H model and the H-H ratings is 0.83 (St. Dev. = 0.37). The H-R modeled motor workload values are plotted with the

Figure III.38: H-H modeled and mean participant-rated motor workload in each investigation area, by condition.




associated participant responses in Figure III.39. The values are very close and the mean delta is -0.13 (St. Dev. = 0.31). As with the H-H model, five of the six H-R model data points are within one standard deviation of the H-R participant ratings. A t-test indicated that the H-R model produced significantly smaller deltas, $t(10) = 4.87$, $p < 0.001$. This significant difference implies that the H-R model provided a better prediction of participant motor workload ratings than the H-H model.

Modeled tactile workload is on the same scale as the subjective tactile ratings, therefore a direct comparison is possible without rescaling. The mean modeled tactile workload predicted by the H-H model was 1.20 (St. Dev. = 0.10) and was 1.17 (St. Dev. = 0.07) for the H-R model. A t-test indicated no significant difference between the two models.

The modeled H-H tactile workload predictions for each investigation area compared to the corresponding subjective workload values in Figure III.40. All six of the model data points are within one standard deviation of the participant subjective tactile workload ratings, demonstrating the accuracy of the model's predictions. The mean delta between the model and corresponding participant ratings was 0.66 (St. Dev. = 0.26). The H-R model data is presented alongside the participant ratings in Figure III.41. The model's six data points are each within one standard deviation of the participants' ratings as well. The mean delta between the model's predictions and participant ratings is 0.19 (St. Dev. = 0.11). The predictions of the H-R model were shown to have a significantly smaller delta from the participant ratings, $t(10) = 4.08$, $p = 0.002$. These results indicate that the H-R model provided a more accurate set of predictions than the H-H model.

Figure III.39: H-R modeled and mean participant-rated motor workload in each investigation area, by condition.

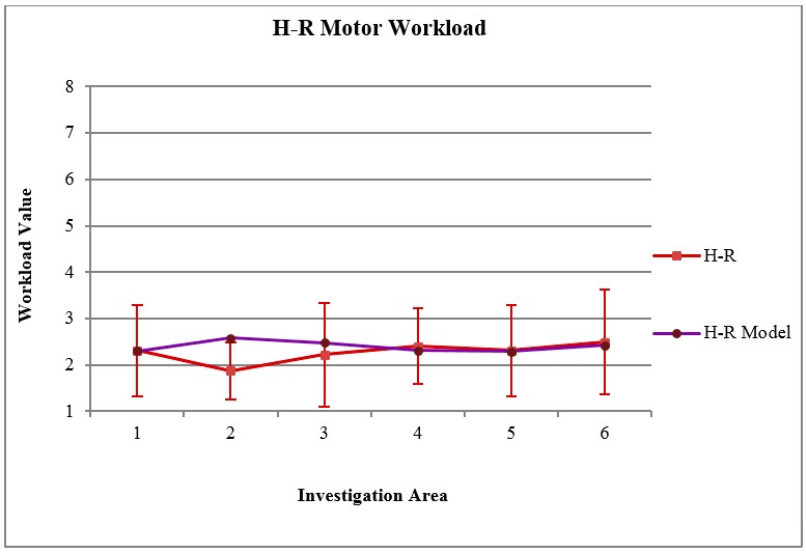



Figure III.40: H-H modeled and mean participant-rated tactile workload in each investigation area, by condition.

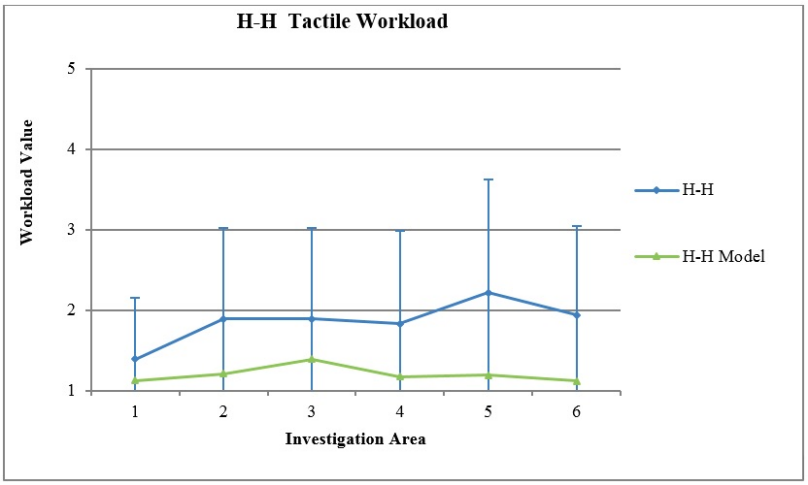
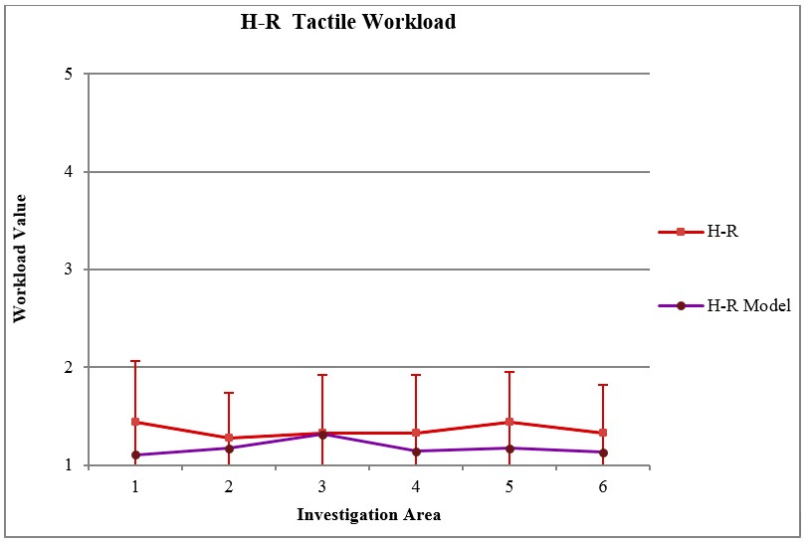


Figure III.41: H-R modeled and mean participant-rated tactile workload in each investigation area, by condition.




### III.2.3.3 Reaction Time and Response Time

The reaction and response time medians and inter-quartile ranges (IQR) are provided in Table III.14. A non-parametric analysis was conducted, as the data was not normally distributed, determined by Shapiro-Wilks tests.

The median H-H condition primary task reaction time was 3 (IQR = 2-5)s and was 6 (IQR = 3-15)s for the H-R condition (Table III.14). Reactions to items resident in the environment that were not part of the evaluation were excluded. A Kruskal-Wallis test found that the H-R participants took significantly longer to react to the primary task, $\chi^2(1) = 50.18$, $p < 0.001$. The H-R participants' primary task reaction time was more than 30s in 38 instances, but there were only five cases for H-H participants. Considering only reaction times less than 30s, the H-H participants had a median primary task reaction time of 3 (IQR = 2-5)s and required significantly less time (Kruskal-Wallis test, $\chi^2(1) = 28.17$, $p < 0.001$) than the H-R participants, whose median was 4 (IQR = 3-9)s. The H-R participants took significantly longer to react to the primary task with the entire data set and the set excluding times longer than 30s; the extremely long primary task reaction times did not change the outcome. The H-H condition median secondary task reaction time was 1 (IQR = 0-1)s, and the H-R condition required 1 (IQR = 0-2)s. A Kruskal-Wallis test found that the H-R participants reacted significantly slower to secondary task questions, $\chi^2(1) = 11.05$, $p < 0.001$.

The H-H condition median primary task response time was 1 (IQR = 1-12)s and was 2 (IQR = 1-4)s for the H-R condition (Table 1). The H-R response times were determined by a Kruskal Wallis test to be significantly longer, $\chi^2(1) = 35.49$, $p < 0.001$ to select an appropriate response. The median secondary task

response times for the H-H condition was 1 (IQR = 1-3)s, while the H-R condition mean was 3 (IQR = 1-9)s. The H-R participants required significantly more time to respond, $\chi^2(1) = 50.82$, $p < 0.001$.

The evaluation results were compared with model results by using a 95% confidence interval equivalence test with a threshold of a 1s difference. The model predicted that primary task reaction time was 1.41s for both conditions and the confidence interval ranged from 1.14 to 1.68s. The evaluation results were not equivalent to the model for either condition. The comparison for the primary task reaction time results $< 30s$ were not equivalent to the model either. The predicted secondary task reaction time was 1.06s for the H-H condition and 1.27s for the H-R condition. The model predictions for both conditions were very close to the evaluation results. The confidence interval for the H-H condition secondary task reaction time model was 0.84 to 1.28s. The H-H evaluation confidence interval was 0.42 to 1.56s. The H-H model and the evaluation results were demonstrated to be statistically similar, within one second. The H-R model confidence interval was 1.02 to 1.52s and the evaluation confidence interval was 0.11 to 3.35s. The H-R model and evaluation results were not equivalent within the 1s threshold, but were statistically similar, within a 2s threshold.

The H-H modeled response time was 1.51s and was 1.72s for the H-R model, which was compared to primary and secondary task response times. The confidence interval for the H-H model was 1.22 to 1.80s and the confidence interval for the H-R model was 1.39 to 2.05s. There were no results for either evaluation condition for primary or secondary task response time that were equivalent to the model within 1s. H-H condition secondary task response time was statistically similar, within a threshold of 2s.

Primary task reaction time is the only metric that used a non-verbal, tacit stimulus onset. A large number of H-R participants took $> 30s$ to react to the primary task; thus, that analysis considered for times $< 30s$, as well. Both results indicated that H-R participants took a longer time to notice the out-of-place items. The secondary and primary task response times, and the secondary task response time indicated that the H-R condition participants took longer to respond. The stimulus onsets for these three metrics are explicit verbal prompts and the feedback times are measured using verbal responses. The model was able to reasonably predict secondary task reaction time for both conditions, and H-H condition secondary task response time. The primary task reaction time was greatly underestimated by the model.

### III.2.4 Discussion

$H_C^1$ stated that H-R teams have lower levels of workload than H-H teams. This hypothesis was supported. Objective and subjective results indicated that mental workload, in particular, was lower in the H-R condition. The nature of the relationship between the human and robot in the H-R condition seemingly lowered mental workload. One theory is that participants may have overestimated the robot's capabilities. Participants may have also felt less pressure, assuming that the robot was able to correct mistakes made by a participant.

The participants were told that the robot was completely autonomous, thus the lack of human involvement in judging their performance may have alleviated pressure and led to lower mental workload. This theory may seem to account for some of the drop in workload; however, the calculation of work completed from the subjective mental workload responses provided by participants did not reveal any significantly differing patterns between conditions; the participants seemed to accurately account for the timing of tasks when rating their mental workload.

The secondary task question results, however, did not demonstrate a significant difference in the number of correct answers in either condition. The H-R condition resulted in lower secondary task question performance as investigation index increased. H-R participants answered twice the number of On List questions incorrectly in high investigation index areas than H-H condition participants. This trend can result from a number of factors. Specifically, the H-R participants performed the evaluation tasks over a longer time period due to the robot capabilities. Task interruption can cause a multi-second lag time, called the interruption lag (Trafton et al., 2003). There is a high-cost disruptive effect on perceptual and memory processes when being interrupted during a complex task. Additionally, robots cannot always use the same social cues as humans. The Pioneer 3-DX robot is not considered to be a socially oriented robot with social cues, such as eye contact, to indicate when the robot may begin speaking (Breazeal, 2003). For these reasons, there was likely a higher "switching cost" involved when conversing with a robot. Thus, it may be more difficult to quickly switch between conversation topics to answer an unexpected secondary task question, especially when working in an investigation area with a higher investigation index.

$H_C^2$ stated that the models are a good predictor of mental workload. This hypothesis was not supported for the H-R condition. The mental workload values produced by the H-R model were significantly higher than the H-R condition results, even though the model accounted for the robot's slower speech and movement.

The results of the Collaborative evaluation demonstrate that collaboration between a human and a robot may take longer than the model predicted; extra time may be required for clarifying questions and allowing for awkward interruptions and pauses. Future work will investigate just how much extra time was required and how the extra time was distributed. The time analysis can be completed via video coding. The specific time spent may be robot-dependent, but in collaborative tasks with robots and inexperienced human collaborators, there may be common time sinks. The longer task time is an undesired product of working with a robot partner; however, robotic technology will improve continually, which will alleviate some of these problems (e.g., improved speech recognition, robot navigation speed, and obstacle avoidance). It is important to note that the human and robot teams saw no significant detriments to task performance nor significantly lower levels of work completed.

$H_C^3$ was supported; there was no significant difference in task performance levels between the H-H and

H-R teams. This result demonstrates that the lower mental workload experienced by the H-R condition participants did not reflect lower levels of task performance.

The analysis demonstrates that while the H-H and H-R teams were both able to complete the tasks, the H-R teams took longer. Some tasks may not suffer when a collaborative team takes extra time to complete the task. In the case of this task, the difference in evaluation time between conditions was approximately eight minutes, which can be inconsequential or can be the time in which a bomb explodes.

$H_C^4$ stated that physiologically measured physical workload will be higher in the H-R condition than in the H-H condition. The objective results supported this hypothesis. The subjective results were similar in trend to the Guided evaluation and demonstrate opposite results from the objective findings; the H-H condition participants rated their motor and tactile workload significantly higher than the H-R condition participants. Additionally, the models were shown to correspond to the workload levels rated by participants.

$H_C^5$ investigated by the Collaborative evaluation postulated that a difference in reaction time and response time is measurable between the H-H and H-R teams. This hypothesis was supported by demonstrating the use of non-laboratory measurement techniques and measuring H-R reaction and response time longer. Considering the use of the presented reaction and response time metrics in human-robot peer-based teams, secondary task reaction time was arguably the least ambiguous measurement technique, followed by secondary task response time. Due to the flexibility of the primary task in the Collaborative evaluation, measurement of the primary task response time was less consistent than with primary task reaction time. These four metrics are promising for future research and use in the non-laboratory environments.

Distance traveled was a metric used for comparing the movement of participants during the Collaborative evaluation. This metric was collected after analyzing peak acceleration in the Guided evaluation and not finding the granularity needed for correct results; however, the measurement technique used in the Collaborative evaluation was not successful. Almost half of the participants provided no useful pedometer data. The data provided had very large standard deviations, but the trend indicated that distance traveled may be higher in the H-R condition.

Additionally, a concern from the Guided evaluation regarded the lack of collocation of the human responder partner. The results of the Collaborative evaluation demonstrate that the addition of the collocation of the human partner did not affect the general trends in results predicted by the hypotheses. The objective results indicated that physical workload was higher in the H-R condition, while the subjective results indicated that perceived physical workload was higher in the H-H condition. This dissonance necessitates the use of physiological measures to assess physical workload, as participants' mental assessments of their experienced physical load were not aligned with their physiological changes. The concept of relying on objective measures in combination with, or instead of, subjective measures is not new, but it is important to demonstrate

this claim in multiple domains.

Finally, investigation index was designed to provide a manipulation of workload throughout the evaluation; however, results did not support the claim that workload was consistently manipulated by investigation index. This result may have been caused by tasks that were too easy, a lack of a time constraint on the task, and modeling that was not completed as a check of workload levels before running the evaluation. These potential experimental confounds were considered when planning the subsequent Time-Structured evaluation (see Chapter V).

## III.3 Comparison of Guided and Collaborative Evaluations

### III.3.1 Overall and Mental Workload Analysis

This chapter seeks to determine if the same workload trends occurred across the Guided and Collaborative interaction relationships. New analysis is provided for some metrics, while summaries of previously presented physiological measures, secondary task questions, NASA-TLX ratings, and correlations results (Harriott et al., 2011a,b, 2012a) are provided to support the comparison analysis. The hypothesis of this investigation is that the different relationship type between the participant and teammate (i.e., master-slave vs. collaborative) affected the results; differing trends will correspond with different team relationships.

#### Modeling

The Guided scenario model did not include uncertainty, thus the results are derived from a single simulation trial. The Guided scenario total modeled mental workload values were normalized to the 630 range as a new presentation of these results. The mean mental workload for the H-H model across the 11 triage assessments was 15.73 (St. Dev. = 2.36) and 15.17 (St. Dev. = 1.74) in the H-R model. A Kruskal-Wallis test resulted in no significant difference. The model predictions indicate that the Guided scenario H-H condition mental workload may be slightly higher than the H-R condition mental workload, but the lack of statistical significance prevents strong assertions.

#### Physiological Measures

The Guided evaluation analysis (Harriott et al., 2011a) determined that heart rate and respiration rate were significantly higher in the H-H condition, but were not significantly affected by triage level. A new analysis of low-frequency heart rate variability found a mean H-H condition low-frequency heart rate variability of 122681.1 $ms^2$ (St. Dev. = 424825.9), while the mean low-frequency heart rate variability for the H-R condition was 264364.5 (St. Dev. = 875409.6) $ms^2$. A Kruskal-Wallis test found no significant main effect of condition, while another Kruskal-Wallis test resulted in a main effect for triage level, $\chi^2(2) = 6.93$, p

= 0.03. Pairwise Wilcoxon tests with Bonferroni adjustments indicated low-frequency heart rate variability when triaging delayed victims was significantly higher than for expectant victims (p = 0.02).

Counter to the Guided evaluation results, the Collaborative evaluation resulted in no significant difference between conditions for heart rate. The Collaborative evaluation results also indicated that respiration rate was higher for the H-R condition. This result is the opposite of the Guided evaluation, which found significantly lower respiration rate for the H-R condition.

Neither scenario found a significant difference between conditions for low-frequency heart rate variability. The Guided evaluation heart rate variability results were affected by triage level, but this result was not duplicated in either the heart rate or respiration rate results. The lack of consistent effects of triage level shows that the mental workload manipulations by triage level were not measurable using all of the channels of physiological data. Overall, the physiological measures were inconsistent between the two evaluations and may have been influenced by differences in physical movements between the two evaluations. Physiological measures of mental workload can be subject to interference from physical activity (Kramer, 1990).

**Secondary Task Questions**

The secondary task questions for the Guided evaluation were based on a list of five names that participants were asked to memorize during the pre-trial briefing. Thirteen questions incorporating the names were posed throughout the trial. Previous analysis of response correctness indicated that there were no significant differences between the H-H and H-R conditions (Harriott et al., 2011a). The results from both evaluations demonstrated a lack of significant difference between secondary task question responses between the two conditions in either scenario.

**Task Performance**

The reported ages and breathing rates represent an aspect of task performance. The Collaboration evaluation's number of items found also represents task performance. The Guided evaluation performance can be determined by assessing the quality of the victim assessments, based on participant responses to the triage step responses. If the results are not significantly different, task performance cannot be considered different between conditions in either evaluation. There were no significant differences between the H-H and H-R conditions in either evaluation.

**In Situ Subjective Workload Ratings**

The Guided evaluation total in situ subjective workload ratings ranged from 6 to 30 and were gathered after each victim assessment in the same manner as in the Collaborative evaluation.

Table III.35: Median total subjective mental workload ratings by condition and triage level for the Guided evaluation.

| Investigation Index | Across Both Conditions | | | H-H | | | H-R | | |
|---|---|---|---|---|---|---|---|---|---|
| | Med. | Min. | Max. | Med. | Min. | Max. | Med. | Min. | Max. |
| Delayed | 15.0 | 6.0 | 30.0 | 15.0 | 6.0 | 30.0 | 14.5 | 8.0 | 26.0 |
| Immediate | 14.0 | 6.0 | 30.0 | 15.0 | 6.0 | 30.0 | 13.0 | 6.0 | 24.0 |
| Expectant | 13.0 | 6.0 | 30.0 | 16.0 | 8.0 | 30.0 | 10.0 | 6.0 | 22.0 |

A new analysis of the Guided evaluation in situ mental workload results was performed in order to determine the median total subjective mental workload values; the median, minimum, and maximum mental workload ratings by condition and triage level are provided in Table III.35. The median in situ mental workload for the H-H participants was 15 and 13 for the H-R participants. A Kruskal-Wallis test indicated that the H-H condition was significantly higher than the H-R condition, $\chi^2(1) = 13.07$, $p < 0.01$. A second Kruskal-Wallis test found no significant effect of triage level. A Kruskal-Wallis test indicated the presence of a significant interaction effect of condition and triage level, $\chi^2(5) = 23.98$, $p < 0.01$. A pairwise Wilcoxon post hoc test revealed that the H-H participants rated mental workload significantly higher than the H-R participants ($p = 0.003$ after Holm's correction).

The Collaborative evaluation results also showed lower levels of mental workload in the H-R condition but were not significantly different across conditions. However, the total subjective workload ratings from both evaluations indicate the same trend: The H-R condition participants reported lower reported subjective mental workload values.

**NASA-TLX Responses**

No significant difference between conditions for the NASA-TLX results was found for either evaluation, but both sets of results demonstrated lower total perceived mental workload for the H-R condition (Harriott et al., 2011a). Both evaluations showed that the individual mental workload component mean scores were higher in the H-H condition than the H-R condition, except for in the Frustration component.

**Comparing Modeled Workload and In Situ Subjective Workload Ratings**

The Guided evaluation mental workload analysis found that the H-R model was a closer fit to H-R condition in situ workload ratings than to the H-H condition ratings and the H-H model (Harriott et al., 2011b). The mean of the Guided evaluation modeled workload (H-H and H-R models) was 15.45 (St. Dev. = 2.04) and was 14.87 (St. Dev. = 5.77) across both evaluation conditions (H-H and H-R conditions) for the in situ workload ratings. A Kruskal-Wallis test by dataset (i.e., H-H model, H-R model, H-H condition, and H-R condition) indicated that there was a significant effect of dataset on mental workload values for the

Guided scenario, $\chi^2(3) = 15.15, p < 0.01$. Post hoc pairwise Wilcoxon tests indicated that the only significant difference was that the H-H condition participants rated mental workload significantly higher than the H-R condition participants ($p < 0.01$). There were no significant differences between the H-H model and the H-H condition, the H-R model and the H-R condition, or the H-H model and the H-R model. Both Guided scenario models appeared to be good predictors of the evaluation mental workload results. This result differs from the Collaborative scenario results, since the Collaborative scenario models predicted significantly higher mental workload than was found for the H-R condition evaluation results.

**Correlations Analysis**

The Guided evaluation analysis yielded a significant positive correlation between heart rate and total in situ mental workload ratings, r(290) = 0.16, p $<$ 0.01, and a significant negative correlation between respiration rate and total in situ mental workload ratings, r(290) = -0.15, p $<$ 0.01 (Harriott et al., 2011a). A significant negative correlation between normalized respiration rate and total in situ mental workload ratings was found for both evaluations.

An analysis using a Pearson's product-moment correlation found a significant correlation between time taken to assess each victim and the corresponding in situ mental workload ratings, r(293) = 0.16, p $<$ 0.01. This result indicates that the lower in situ mental workload ratings in the H-R condition may be due to the fact that the triage assessments took a longer time. This result is the opposite of the same correlation for the Collaborative evaluation, which showed that longer investigation time was not significantly correlated to higher mental workload. The Collaborative evaluation indicated that the lower H-R mental workload was not due to more time spent performing investigation tasks.

**Work Completed**

The amount of work completed during each victim assessment was computed as defined in Chapter III.2. The H-H participants had a mean work completion value of 2235.93 (St. Dev. = 1524.69) workload seconds and the H-R participants had a mean value of 2101.70 (St. Dev. = 1331.11) workload seconds. There was no significant difference between the two conditions. This result shows the same trend as the work completed analysis for the Collaborative evaluation.

### III.3.2 Reaction and Response Time Analysis

The Guided evaluation H-R condition results are compared to the Collaborative evaluation H-R results. The results were compared using Mann-Whitney U tests and the two one-sided test (TOST) method of assessing statistical equivalence. The medians and IQRs statistics are provided in Table III.14.

The difference between primary task reaction time results in the Collaborative evaluation were significantly longer than from the Guided evaluation, U = 33659.5, $p < 0.001$. The comparison for secondary task reaction time demonstrated that the two distributions were statistically similar ($\delta$ = 1s: U = 11016.5, $p < 0.01$; -$\delta$ = -1s: U = 28671.5, $p < 0.01$).

Primary task response times were statistically similar, $\delta$ = 1s: U = 758441, $p < 0.01$; -$\delta$ = -1s: U = 1203251, $p < 0.01$. The secondary task response time comparison revealed that the Collaborative evaluation secondary task response time was significantly longer than the Guided evaluation, U = 12806.5, $p < 0.01$.

The similarity analysis for two of the four metrics, primary task reaction time and secondary task response time, demonstrated that they are not significantly similar across evaluations. The secondary task reaction times and the primary task response times were similar across evaluations.

### III.3.3 Discussion

The Guided evaluation focused on a relationship between a human and a robot, which was predicated on a relationship similar to that of a master-slave relationship. The robot was an instructor, providing step-by-step triage instructions to the participants, which did not permit the flexible and peer-based relationship that existed in the Collaborative scenario. Despite this major difference in the dynamics between the teammates, similar results were found for both evaluations.

The hypothesis of this comparison was that the differing relationship style between the two teaming scenarios will lead to differing workload and performance trends, but it did not. This result indicates that the teaming styles may have similar effects on the human participant's mental workload and performance. The participant and teammate do not share the knowledge about how to complete the triage task at the beginning of the task. During the task, the participant's understanding of the task grows, but he or she still relies on the human or robot teammate for guidance and the teammate remains the supervisor of the task.

The remaining collaboration relationship requirements (see Chapter II) were met, including a set of shared goals (i.e., completing the victim triage) and reliance on each other to complete the task (i.e., the participant required instruction and the partner required the participant to perform the task) (Bratman, 1992). The master-slave roles do not change, and shared knowledge builds between the participant and teammate.

The original Guided evaluation analysis raised a concern as to whether or not mental workload was lower in the H-R condition due to not having a physically collocated partner. The H-H condition required participants to communicate with their human partner via walkie-talkie. This concern was shown to not affect major trends in mental workload, since the Collaborative evaluation also demonstrated lower H-R condition mental workload levels. Data from both evaluations support the participants' ability to attain a similar performance level for assigned tasks, regardless of condition. This result is supported in the Guided evalua-

tion by the secondary task question responses, the reported victim ages, and the reported victim respiration rates. The Collaborative evaluation results also supported that performance levels were similar across the two conditions.

The lower in situ subjective workload ratings and NASA-TLX scores in the Collaboration evaluation's H-R condition are very similar to the Guided evaluation's in situ subjective workload ratings and NASA-TLX scores. Participants perceived lower mental workload in the H-R condition despite a lack of training in robotics or first response. Physiological results alone remain unclear, but the negative correlation between total subjective mental workload ratings and normalized respiration rate can be useful when analyzing data; a sudden drop in respiration rate can indicate an increase in mental workload. Overall, when comparing workload data across the two evaluations, it is apparent that a) the H-R condition mental workload ratings are lower, b) task performance is not different between the two conditions, and c) H-R teams take longer to complete the given task.

Reaction time and response time were also investigated across the Guided and Collaborative evaluations. Measuring the four presented metrics (i.e., primary task reaction time, primary task response time, secondary task reaction time, secondary task response time) in non-laboratory environments raise points about their generalizability and limitations.

Primary task reaction time requires a measurable stimulus onset time with a recognizable reaction point. The Guided evaluation's primary task relied on verbal stimuli prompts and mostly verbal reactions. As a result, the video coders were able to identify easily the stimulus onset time and a corresponding reaction occurrence. The metric, as defined, was successfully collected because the verbal stimulus onset times and resultant reaction times were clearly present in video recordings. The Collaborative evaluation's primary task reaction time was less clear. Identifying the stimulus onset required noticing an item in the environment via the video. Two potential limitations of implementing this metric for the Collaborative evaluation are: the point-of-view camera, and the lack of a task prompt as a stimulus onset time. The point-of-view camera lacks eye tracking, which may prevent video coders from determining an exact stimulus onset time when the out-of-place item entered the participants' true field-of-view. The Collaborative evaluations' measured primary task reaction time's stimulus onset time provided an estimate of the time for participants to notice an item, not necessarily the exact moment the participant first saw the item. These limitations prevented the primary task reaction time from being a true reaction time measure for the Collaborative evaluation. The metric was successful in the explicit task prompt Guided evaluation, which suggests that, either an eye-tracking device for visual tacit cues, or explicit cues (verbal or otherwise) are necessary when tasks do not have a clearly measurable stimulus onset time.

Secondary task reaction time represented the most recognizable reaction time metric, as it involved a very

clear stimulus onset (i.e., end of a verbal question) and a clear associated reaction (i.e., first verbal utterance) in both evaluations. This metric successfully captured what it was defined to capture in both evaluations. A limitation of this metric may be the type of secondary task used (i.e., verbal questions).

Primary task response time was measured similarly in both evaluations. Verbal questions and prompts were the stimulus onsets, and participants' associated appropriate responses were measured in both evaluations. The Guided evaluation paired the primary task reaction and response times by using the same stimulus onsets for both measurements. The Collaborative evaluation measured different aspects of the primary task for reaction time and response time, which is a limitation of the evaluation. Response time cannot be used to pinpoint decision-making time with respect to reaction time, if reaction and response times do not measure the same aspect of the primary task. An additional limitation is the use of response time to encompass all types of primary task questions in both evaluations. Limiting the measurement of primary task response time to a specific question type may improve the ability to model the response time, and may provide more specific insight into decision-making processes.

Secondary task response time included any verbal utterances that participants made before providing an appropriate response to a question. Both evaluations demonstrated successful measurement of secondary task response time. The time between the reaction time and response time potentially contains, for example, time during which participants' think silently, holds conversation unrelated to the question, question clarifications or repeats are provided, or encounter distractions from another task. Secondary task response time is related to demonstrating the spare mental resources available in relation to the primary task. A potential metric limitation is that the response time does not necessarily incorporate the context of what is happening specifically before the participant provides the appropriate response.

Additionally, a physical workload comparison was not specifically presented, but both the Guided and Collaborative evaluations presented very similar results. The relationship between higher subjectively-measured physical workload and lower objectively-measured physical workload in the H-H condition may relate not only to the longer time spent on task by H-R teams, but also to the lack of social pressures felt by H-H participants. H-R participants, especially in the less structured Collaborative task, may have felt more emboldened to search farther or walk around while the robot was talking, whereas a participant with a human partner did not. Sociability with robots is not an aspect of interest for this dissertation, but these social aspects may play a role in physical workload measurement.

Chapter III.4 discusses further limitations of the general evaluation designs and robot embodiment in both evaluations, as well as a general discussion of the results of both evaluations and the analysis of overall and mental workload, physical workload, and reaction and response time.

### III.4 General Discussion

The Guided and Collaborative scenarios were each designed to evaluate workload (i.e., overall, mental and physical) and task performance in human-human and human-robot teams. Human performance modeling was also evaluated for application to human-robot peer-based teams. This approach was taken because accurate human performance modeling can be used in H-R teams for planning future team configurations and task assignments.

The coupling between modeling and evaluations represents a step toward using modeling predictions onboard a robot that is a member of an H-R team. Future teams will equip robot teammates to monitor the state of their human counterparts using real-time tools (e.g., physiological sensors or cameras). The robot will be able to record task times, for example, or estimate workload levels using these tools. Model predictions can be developed for the task types the team will be assigned and the robot can compare actual performance and workload to predicted performance (e.g., completion time and workload spikes expected during tasks). Robot teammates can adapt their behavior by, for example, accommodating unexpected performance drops by slowing the task's pace or taking over a task from the human, if the human's workload is too high.

A robot can monitor the human's workload and performance levels in a variety of ways, with multiple sensors, inquiries, and observations. In general, the assessment of a range of mental workload and task performance metrics for this domain is important. Subjective measures alone are insufficient, and objective measures each offer individual pros and cons. Assessing which metrics from other domains can be applied to the development of robots for such H-R teams is critical, with potentially mobile H-R team situations that may not permit robots to directly (e.g., visually) observe human partners.

This chapter analyzed overall workload and mental workload, physical workload, and reaction and response time metrics. First, multiple measures of overall workload and mental workload were analyzed. All of these overall and mental workload metrics, and their advantages and disadvantages, are presented in Table III.36. Some of the disadvantages cannot be avoided.

The physiological measurements created high sensor noise during the Guided and Collaborative evaluations. The specific BioHarness chest strap sensor used was subject to movement about the sensing area, slightly below the participants' breastbone. The participants fit themselves with the Velcro-secured strap of a general size (e.g., small, medium or large), but the strap did not easily stay in place.

The new BioHarness chest strap sensor has a design that appears to remedy these problems. The sensor itself sits along the left side of the body, rather than in the front, and the chest band has an additional shoulder strap that secures it in place around the body. Keeping the sensor secure will aid in receiving a stronger signal. An additional issue with physiological measures is that they are influenced by physical body movements, such

as walking. It is difficult to differentiate between mental workload artifacts in heart rate and heart rate changes caused by physical movement. The tasks in both scenarios were not extremely physical, but walking may have been a factor in the inconclusive results. The disadvantages of using physiological measures do not erase the advantages; physiological sensors offer a good means for a robot to monitor the human's internal state, while working on a team task. A sensor like a chest strap heart rate monitor is wireless, relatively inexpensive, and the new design will be less obtrusive to the user.

Task density is useful to compare the amount of work accomplished in a given amount of time, including examining the overall workload that the participants experienced. The disadvantage of this measure, however, is determining what precisely constitutes a task in a more loosely-structured task (e.g., a conversation in a collaborative team or a vigilance-based task). The developed work completed metric is based on the definition of workload as the amount of work to accomplish relative to the amount of time to accomplish it (Wickens et al., 2003). Work completed factors out time and compares the amount of work completed during a subtask by multiplying subjectively rated workload and time taken to complete the subtask. Work completed offers the advantage of factoring out time but possesses the disadvantage of having arbitrary units. Arbitrary units are a disadvantage, because they are only meaningful for the specific calculation.

Secondary tasks provide information regarding the spare mental capacity of participants (Gawron, 2008). Participants in both the Guided and Collaborative evaluations correctly answered almost all of the questions, which may indicate a high level of spare mental capacity. This result can also reflect a secondary task that is too simple. A balance must be struck between a secondary task that is relevant and challenging but does not negatively impact the primary task.

Subjective metrics have the disadvantage of requiring an interruption of the task in order to administer questions. Participants also are not precisely reliable due to the lag time between when they experienced each moment during the task and when they eventually report the rating, even when the ratings are collected on a fairly frequent basis. Subjective measures do offer unique feedback regarding the participants' "perceived load" that other types of workload metrics cannot (Moray, 1982). Knowledge regarding how an individual is mentally interpreting what he or she experiences is valuable as well; however, equipping a robot to administer verbal subjective questions or the NASA-TLX survey during a deployed team mission is not practical given that administering such questionnaires will interrupt real-world tasks.

Physical workload is another important area of research in human-robot peer-based teams. As deployment times increase and robots are introduced into critical fields, such as military missions and first response, it becomes necessary to understand the impact of physical workload on the team. Additionally, potential mitigations for physical workload via use of human-robot systems are important to motivate the development of new robotic technology. Previously, physical workload may not have been perceived as important to the

Table III.36: Summary of evaluated workload metrics, with data collection method and associated advantages and disadvantages.

| **Metric** | **Data Collection Method** | **Advantages** | **Disadvantages** |
|---|---|---|---|
| Physiological Measures (heart rate, respiration rate, heart rate variability) | Chest-strap sensor | Portable unobtrusive sensor, no task interruption, potential for future live measurement by robot, validated for other domains, objective metric | Sensor noise, signal is influenced by physical movement, inconclusive findings between evaluations |
| Task Density | Ratio of subtask activities to subtask time | Provides an estimated workload ratio for comparison, objective metric | Requires determination of subtasks completed during subtask time |
| Work Completed | Product of in situ subjective workload ratings and subtask time | Demonstrates an estimate of how much was accomplished by factoring out the time taken, objective metric | Not validated metric, arbitrary units (workload seconds) |
| Secondary Task Questions | Verbal prompt, periodically during evaluation, recognition and recall | Questions can be customized to fit the scenario, brief, technic validated for other domains, objective metric | Potentially distracts from primary task, evaluation findings showed that measure was not sensitive or primary task was too easy, subjective metric |
| In Situ Subjective Workload Ratings | Verbal prompt, periodically during evaluation | Easily collected responses, offer insight into "perceived workload" | Responses influenced by most recent moments of task and multitasking, measurement must occur between tasks or subtasks, subjective metric |
| NASA-TLX | Survey, post-evaluation | Widely used/validated workload measurement tool, offer insight into "perceived workload" | Computer or paper is necessary for survey administration, measurement must occur between tasks or subtasks, response takes multiple minutes, subjective metric |

human-robot interaction community at large because many tasks or missions did not require long durations or assumed a user interacting with a robot via a graphical user interface. As the human-robot interaction field incorporates more peer-based interaction and teamwork, the response to human performance from physical demands cannot be overlooked.

An overall goal of the Collaborative evaluation was to investigate human performance for humans working with a collocated robot or human partner for shared tasks requiring joint decision-making. The Collaborative evaluation's investigation into physical workload also showed similar results to those produced by the Guided evaluation, even though the Guided evaluation's results were potentially influenced by the lack of a collocated human partner in the H-H condition. Initial analysis indicated that overall workload was lower in the H-R condition it was determined that physical workload, an individual component of overall workload, does not follow the same trend, indicating that physical workload may be worth monitoring separately. Specifically, overall workload may not provide an accurate assessment of all team task activities involving physical movement. Physical workload is an important component to consider when designing human-robot teams and systems for optimized performance. The physical impacts of a task can significantly limit the amount of time a team can be deployed and the number of tasks a team can accomplish.

Additionally, working with a robot may generally affect reaction time and response time. For example, response time increased by listening to a robot's voice; a synthetic voice takes 0.2 seconds longer to parse than human speech (Duffy and Pisoni, 1992). The Collaborative team response time models predicted successfully the 0.2s differential for the robot's synthetic speech based on a micromodel of human speech and extrapolating the time for the robot's expected speech speed.

Robots may not be equipped to time task interruptions as well as a humans. The more rigid Guided evaluation's task structure resulted in fewer multitasking opportunities. The Collaborative evaluation's robot partner was unable to perceive the participant directly; the robot supervisor was also unable to do so. The H-R participants were sometimes actively searching for items when presented with a secondary task question. Rather than stop the primary task, participants continued the search, while responding to the secondary task. Such multitasking has been shown to increase reaction time (Salvucci and Taatgen, 2008). H-H condition participants, in contrast, were typically not interrupted when actively searching. Interruption lags represent the length of time between an alert and the start of a secondary task, and the resumption lag captures the length of time between the end of that secondary task and the first action to resume the primary task from before the interruption (Trafton and Monk, 2007). The resumption lag represents the time necessary to collect one's thoughts after an interruption has completed, which may have been integrated into the primary task reaction time. Specifically, H-R participants may react to a primary task, get interrupted by the robot, handle the interruption, and collect his or her thoughts before responding to the primary task. Resumption lag has been

reported to be on the order of approximately 3s (Altmann and Trafton, 2004), which may partially explain the Collaborative evaluation's longer H-R condition primary reaction times.

This dissertation is a step toward creating robots that can predict and monitor human performance and workload during an H-R collaborative task. The Guided and Collaborative scenario investigations offer human performance model assessments, workload metric considerations, data comparisons between H-H and H-R teams, and comparisons between master-slave and collaborative teams.

### III.4.1 Limitations

There are two types of limitations associated with this work: the robot itself and experimental design limitations.

#### III.4.1.1 Robot Implementation and Embodiment

Future robot teammates in these collaborative H-R teams will be deployed with capabilities for monitoring the status of the human teammate and adapting interactions and behaviors accordingly. The design of the robot's outer appearance and voice may also affect the team's dynamic and needs to be considered within the context of H-R collaborative teams. The robot's actual role and level of agency may or may not correspond with the human teammate's interpretation of the robot's appearance.

The robot used in both evaluations was not a humanoid. Speaking with the robot can result in a collaboration time lag due to robot's ill-timed interruptions (Trafton et al., 2003). Conversation with the robot during the two evaluations suffered from awkward moments when the robot interrupted, cutting off its human partner mid-sentence or beginning to talk at a time when the human partner was performing a task involving concentration. The primary task reaction time metric appears to capture this limitation between conditions and may reflect shorter reaction times when the robot is aware of the participant's state and the potential impact of interruptions. Interacting with a robot that has more human-like qualities (e.g., eye contact and normal conversation pauses) may help to smooth the conversational flow.

The nature of the semi-autonomous (i.e., Wizard of Oz) robot implementation may be a limitation that affects participants' reaction and response times. The Wizard of Oz robot setup requires a human experimenter to listen to participant's verbal responses, interpret them and initiate the robot's preprogrammed (in most cases) response. This configuration incorporates the human experimenter's reaction and response times into the conversation between the human participant and the robot, which can impact the participant's expectations of conversation timing, resulting in the participants' responding more slowly.

Additionally, humanoid robots have qualities to emulate human interaction, such as active vision in the form of movable camera perception systems, and have been created mostly as tools, surrogates, and compan-

ions for people (Brooks, 2002). Social interactions with a robot require intentionality, and humanoid robots are able to provide intentionality cues via their human-like traits, such as facial expressions and movements (e.g., Kismet) (Breazeal and Scassellati, 1999). Collaborative interactions can be improved by including nonverbal communication, such as having the robot partner recognize head nodding and respond by nodding back (Sidner et al., 2006). Employing mutual gaze and gazing at objects that are the subjects of conversation can improve the conversation (Mutlu et al., 2009; Sidner et al., 2004). Taking appropriate turns in verbal dialogue by using gaze, verbal cues, body language, and robot learning also makes an H-R conversation easier for the human partner (Chao and Thomaz, 2010).

Human-like conversation qualities do not necessarily correspond to a humanoid robot, but in order for the robot to create eye contact, it must be capable of head or eye movements. More than likely, a non-humanoid robot may need to emulate the positive effects of eye contact via other actions, such as turning its body to face the human when he or she is speaking or toward the object that he or she is speaking about. The robot used in the Guided and Collaborative evaluations did not employ such gestures.

Hinds et al. (2004) evaluated the effect of robot appearance on feelings of human responsibility during H-R collaboration and found that working with a machine-like robot increased feelings of personal responsibility over working with a humanoid robot. The authors recommend a more machine-like robot when humans must feel task responsibility and seek solutions to problems, rather than trust the robot to find a solution for something that it cannot. Humanoid robots were suggested to be more advantageous in situations where the human was not required to be as diligent. Additionally, the participants spoke to the machine-like robots more often than they did to the humanoid robots. The authors postulated that the increase in conversation with the machine-like robots was a result of participants compensating for a lack of perceived common ground in order to explain what they were doing in more detail.

Humanoid robots provide a way for humans to rely on the robot more than a machine-like robot (Hinds et al., 2004). This fact may not always be desirable in H-R team collaboration, because the human partner can lower his or her diligence level and feelings of responsibility. A machine-like robot with a few human-like attributes may be a compromise that works to leverage the advantages of both sides of the robot appearance spectrum.

#### III.4.1.2 Experimental Design

This analysis has three primary experimental design limitations that must be acknowledged. First, workload has not been examined in the context being a moderator or mediator between assigned task partner and task performance for the presented evaluations. If workload were a mediator of performance, then the assignment of either the human or robot partner influences the resulting workload level, which, in turn, impacts the task

performance level (James and Brett, 1984). Workload may also be a moderator of task performance, which implies that task partner impacts task performance regardless of workload level, but changes in workload level can change the effect that task partner has on task performance. Future work will assess the interaction effects of workload and task performance.

Second, this work analyzes H-R interaction in a specific first response domain with two team members. Generalizing the findings to different domains may be possible if the relationship between the human and robot is collaborative, the tasks are mobile, and interaction is largely verbal. A third team member will potentially change the workload balance across the team. A team of two robots and one human will likely raise the workload of the human; the human can interact with the second robot while the first is finishing its own task or catching up. As a result, the human will have a higher task density; thus, the human will have a higher workload level. A team of three humans may result in lower workload by having an additional person to further divide work during times of overload.

The two investigations presented focus on only two relationships between human and robot; the Guided evaluation featured a master/slave-style relationship, where the robot was master, and the Collaborative evaluation analyzed a collaborative relationship. Workload and task performance differences may differ in different H-R interaction styles. For example, if the robot is truly more of a "master" of the human in the Guided evaluation, how will the human respond? The reverse may also be interesting; if the human possesses the knowledge and the robot is the executer of tasks in the Guided evaluation, then workload and performance levels may be altered. These questions regarding alternate relationships may be relevant for other types of H-R interaction.

### III.5 Summary

This chapter presented two evaluations: the Guided evaluation and the Collaborative evaluation. A variety of workload and task performance metrics were presented and analyzed in these evaluations. The goal of Chapter IV is to provide a clear guide for using the metrics presented in this chapter, in addition to a set of newly presented metrics, for workload and task performance assessment.

# Chapter IV

## Metrics of Workload and Task Performance for Human-Robot Peer-Based Teams

This chapter presents a unique set of workload and task performance metrics specifically selected for use in human-robot peer-based teams. The metrics are defined for use in this specific domain. The data analyzed is from the Guided and Collaborative evaluations (presented in Chapter III). Analyzing the existing data with provided metrics aided in determining which metrics were most appropriate to apply to the Time-Structured evaluation (Chapter V). A subset of the metrics were computed using video coding of data from the Guided and Collaborative evaluations. A goal of analyzing these metrics is to provide guidelines for moving away from the necessity of video coding, subjective questionnaires, and laboratory-based objective metrics and toward automated measurements on-board a robot in human-robot teams. Multiple coders computed the metric values. Reliability was calculated via inter-coder reliability. Repeatability was analyzed by comparing relationships between the new metrics and known valid measures of workload and task performance across the evaluations.

This analysis of the presented metrics serves as a guide that provides insights into appropriate use, relevant references, and both the positive and negative attributes associated with each metric when using them to represent workload and task performance. The metrics are presented with a corresponding definition, examples of use, general guidelines for use, implementation, measurement, and results from the Guided and Collaborative evaluations, concepts for measurement by a robot in human-robot teams, and general limitations (see Table IV.1). Table IV.1 provides a summary of all of the presented metrics, along with a brief description of the measurement and data collection technique.

### IV.1 Workload Metrics

Workload is an important measure when designing for optimal human performance (see Chapter II.2.1). This section details twelve workload metrics, including metrics representing overall, mental, and physical workload. Each metric is presented with a definition, general guidelines of use, a description of the implementation in the Guided and Collaborative evaluations, analysis results from the Guided and Collaborative evaluations, approaches for implementing measurement in future human-robot team, and associated limitations.

#### IV.1.1 Heart Rate

Heart rate refers to the number of heart beats per minute. Heart rate can be recorded in several relatively unobtrusive manners, such as chest strap monitor or wrist monitor. Heart rate increases with workload and

Table IV.1: Summary of all metrics in proposed analysis.

| Measurement | Metric Name | Method of Measurement | Data Collection |
|---|---|---|---|
| Workload | Heart Rate | mean during task time calculated post-hoc | physiological sensor |
| | Respiration Rate | mean during task time calculated post-hoc | physiological sensor |
| | Heart Rate Variability | mean during task time calculated post-hoc | physiological sensor |
| | Postural Load | percentage of task time spent with torso bent at more than 45° | physiological sensor |
| | Variance in posture | variance of all postures during task time | physiological sensor |
| | Vector Magnitude | mean during task time calculated post-hoc | physiological sensor |
| | Movement Count | number of occurrences of specific movement | video coding |
| | Memory Recall | number of correctly recalled items of interest during evaluation | list of recalled items provided by participant post-evaluation |
| | Task Density | ratio of tasks per unit time | computation of tasks performed per unit time |
| | Speech Rate | syllables spoken per unit time | calculation from audio coding |
| | Secondary Task Failure Rate | ratio of task failures to total task attempts | video coding |
| | In Situ Workload Ratings | subjective Likert-scaled ratings along 6 workload channels | verbal inquiry and observer recording |
| | NASA Task Load Index | subjective workload rating tool | online administration post-evaluation |
| Task Performance | Primary Task Failure Rate | ratio of task failures to total task attempts | video coding |
| | Primary task reaction time | time between onset of task-dependent stimulus and the first response | video coding |
| | Secondary task reaction time | time between end of question and first audible response | video coding |
| | Primary task response time | time between end of question or instruction and an appropriate response | video coding |
| | Secondary task response time | time between end of question and an appropriate response | video coding |
| | Subtask Time | time taken to complete a specific subtask | video coding |
| | Work Completed | normalization of workload ratings by time | computation of workload ratings multiplied by subtask time |
| | Distance Traveled | two-dimensional ground covered by participant during task | physiological sensor / pedometer |

physical activity (Castor, 2003).

**Guidelines for Use.**

Heart rate is a direct measure of cognitive workload (Castor, 2003) that can be measured with simple sensors and does not require complex software for analysis (Vicente et al., 1987; Poh et al., 2011b). Normalizing participant heart rate is a method of reducing the effect of individual differences in smaller sample sizes (e.g., Vicente et al., 1987; Mehler et al., 2009). It is common practice to normalize heart rate based on a baseline heart rate taken from the participant before the evaluation begins.

**Measurement implementation and analysis in prior evaluations.**

Heart rate was measured using the Bioharness chest strap heart rate monitor during both evaluations (Biopac Systems, 2013). The monitor recorded heart rate in offline mode on-board the sensor. Baseline heart rates were collected for each participant during training and following the evaluation trial. The baseline heart rates were used from training to calculate a normalized heart rate. The results associated with the Guided and Collaborative evaluations, including a comparison across the evaluations, are presented in Chapter III.

**Limitations.**

Heart rate is very sensitive to emotional arousal and physical activity. Heart rate may not accurately represent cognitive workload in scenarios where the participants are physically active. Sensor quality and placement can also influence the quality of data recorded. Chest strap bands must be the appropriate size for the wearer. They can also move around during activity, causing sensor error.

**Implications for future human-robot teams.**

A heart rate sensor can stream data in real-time to a robot. The robot can interpret changes in heart rate to assess the workload status of the human teammate. High heart rate can indicate high levels of workload and physical activity.

### IV.1.2 Respiration Rate

Respiration rate is the number of breaths taken per minute and can be measured via stethoscope or chest strap monitors.

**General Guidelines for Use.**

Respiration rate is a direct measure of workload that is easy to record using a chest strap monitor and does not typically require complex software for analysis. Respiration rate has been shown to decrease as cognitive

workload increases (Roscoe, 1992; Keller et al., 2001).

**Measurement implementation and analysis in prior evaluations.**

Respiration rate was measured using the Bioharness chest strap heart rate monitor during both the Guided and Collaborative evaluations (Biopac Systems, 2013). The monitor recorded respiration rate in offline mode on-board the sensor. Respiration rate baseline values were recorded for a period of one minute during the training period and during the post-evaluation questionnaires. Normalized respiration rate was calculated using these baseline values. The results associated with the Guided and Collaborative evaluations, including a comparison across the evaluations, are presented in Chapter III.

**Limitations.**

Physically active tasks can cause increases in respiration rate and prevent measurement of cognitive workload (e.g., getting out of breath when running). Respiration rate may most accurately represent workload during tasks with limited physical activity. Mobile human-robot teams may not choose respiration rate as a measure of workload given this limitation. Chest strap monitors can move during physical activity, resulting in incorrect measurements. Additionally, chest straps that are too loose on the wearer will not provide accurate readings.

**Implications for future human-robot teams.**

Many chest strap sensors capable of measuring respiration rate are also able to transmit data over a wireless connection in real time. A robot can assess changes in respiration rate and determine the workload status of a human team member, given the assumption that elevated physical activity is not a factor in the tasks.

### IV.1.3 Heart Rate Variability

Heart rate variability is a measurement of the variation in the beat-to-beat interval of heart rate. Heart rate variability is used as a metric for workload and emotional arousal (Vicente et al., 1987) and typically decreases as mental workload increases (Aasman et al., 1987).

**General Guidelines for Use.**

Heart rate variability is a direct measure of workload. Heart rate variability cannot be computed using heart rate information alone. A measurement device must have the capability to record the heart's inter-beat variation.

Some sensors offer the ability to live stream heart rate variability data. The Bioharness monitor's associated software, for example, automatically computes live heart rate variability by extracting the frequency of

the beat-to-beat intervals and calculating a power density calculation. If the data is recorded on the device, a post-recording power density computation must be performed on the recorded inter-beat variation data.

**Measurement implementation and analysis in prior evaluations.**

Heart rate variability was measured using the Bioharness chest strap heart rate monitor during both the Guided and Collaborative evaluations (Biopac Systems, 2013). The monitor recorded heart rate variability in offline mode on-board the sensor; thus the power calculation to compute heart rate variability values were performed following the data extraction from the monitor. The results associated with the Guided and Collaborative evaluations, including a comparison across the evaluations, are presented in Chapter III.

**Limitations.**

Heart rate variability is sensitive to noise. Heavy physical activity can influence heart rate and heart rate variability. Additionally, incorrect sensor placement can negatively influence data accuracy.

**Implications for future human-robot teams.**

A heart rate variability sensor can stream data in real-time to a robot. The robot can interpret changes in heart rate variability to assess changes in workload. A possible behavior modification for the robot includes taking over a task for a human partner if workload overload is detected.

#### IV.1.4 Postural Load

Postural load is defined as the time during which a participant's trunk is flexed more than 45° (Paul et al., 1999). Postural load represents physical workload. Higher postural load indicates higher levels of physical workload.

**General Guidelines for Use.**

Postural load is a direct measure of physical workload. Participants in the Guided evaluation were tasked with crouching over the assessed victims, and postural load determined how long participants were significantly bent over the victim. Postural load was not measured in the Collaborative evaluation due to the lack of crouching behavior inherent in the task. Postural load is relevant for tasks that involve bending the trunk.

It is necessary to know the sensor's sampling rate (e.g., 100 Hertz) when calculating postural load to compute the total number of samples in the time period of interest (e.g., 1000 samples within 10 seconds). All posture samples during the period of interest are given a score of 1 if they are over 45° and a 0 if they are not. The sum of scores indicates the number of samples where the participant was bent over (e.g., 200

samples of 1000 total samples). The sum of scores is divided by the total samples to compute the percentage of time spent bent over more than 45° during the time period of interest (e.g., 200 ÷ 1000, or 20%).

**Measurement implementation and analysis in prior evaluations.**

Postural load was measured using the Bioharness chest strap heart rate monitor during the Guided evaluation (Biopac Systems, 2013). The monitor recorded the participants' raw posture value in offline mode on-board the sensor and postural load was calculated following the evaluation. The sampling rate was 250 Hertz. Raw posture values were collected during the Collaborative evaluation, but postural load was not calculated because the involved tasks did not require much bending of the participant's trunk. The results associated with the Guided evaluation are presented in Chapter III.

**Limitations.**

Postural load is calculated as a percentage of time. This measurement technique requires a known time period of interest for framing the postural load value. For example, a participant may be tasked with bending over to place a box on the floor. The postural load will jump to very high levels if the time of interest is the time period during the box placement. However, if the task time of interest includes the time that the participant finds the box, picks it up, places it quickly on the floor and then stands up to wait for further instruction, the postural load may be much lower.

**Implications for future human-robot teams.**

A robot can begin calculating postural load following the onset of a task given a sensor that can live-stream posture data. The robot can track whether a human teammate is performing a crouching task by tracking postural load. The robot may be able to distinguish between postures, or when the human teammate has finished a task requiring crouching, when the postural load begins to dip below a threshold. The robot can then follow up with the teammate on the task's status.

### IV.1.5 Variance in Posture

Variance in posture is the variance of changes in body posture (Harriott et al., 2013). This metric was adapted from the concept of posture sway (Lasley et al., 1991). Higher values of variance in posture represent more changes from mean posture to alternate positions, which is measured in units of degrees squared.

**General Guidelines for Use.**

Variance in posture is a direct measure of physical workload that can be calculated using data from a triaxial accelerometer positioned on the body. Options for the sensor placement are task-dependent (e.g., around the

chest to assess amount of bending at the trunk, on the wrist to assess arm movements). The variance is taken of all posture values during a specified time period. Estimates of physical workload are most valuable when the task requires physical activity.

**Measurement implementation and analysis in prior evaluations.**

Variance in posture was measured using the Bioharness chest strap heart rate monitor during both the Guided and Collaborative evaluations (Biopac Systems, 2013). The monitor recorded raw posture values in offline mode on-board the sensor and variance in posture was calculated following the evaluation. The results associated with the Guided evaluations, including a comparison across the evaluations, are presented in Chapter III.

**Limitations.**

Variance in posture is limited by the placement of the triaxial accelerometer sensor. Sensor placement limits the scope of knowledge regarding amount of physical activity. Additionally, a task with a high physical workload demand may result in low variance in posture (e.g., squatting, holding a heavy object while standing). Variance in posture is really only useful for assessing gross body movements.

**Implications for future human-robot teams.**

A robot can assess physical workload by computing variance in posture for a recent time period. The robot may be able to enact specific behaviors when a participant is inactive for a long period of time, which may include checking on the participant regarding injury, sleep, or loss of sensor. Additionally, the robot may enact specific behaviors when the level of physical activity is too high, such as ending a task early or assisting the human teammate.

### IV.1.6 Vector Magnitude

Vector magnitude has been used to represent physical movement (Steele et al., 2000) and refers to the magnitude, or mathematical size, of the walking vector. High levels of vector magnitude indicate high levels of walking movement, a factor in physical workload.

Vector magnitude is measured in vector magnitude units (VMU) and is calculated by capturing accelerometer data (A) in three dimensions (i.e., vertical, x; sagittal, z; lateral, y) that is measured in gravitational force (g). Vector magnitude is integrated over a period of approximately 1 second by computing the value for VMU in Equation IV.1 (Johnstone et al., 2012).

$$VMU = \sqrt{{A_x}^2 + {A_y}^2 + {A_z}^2} \tag{IV.1}$$

**General Guidelines for Use.**

Vector magnitude is a direct measure of physical workload. Postural load and variance in posture are measures representing body movements without factoring in movement about the evaluation area, but vector magnitude represents movement about an area. Vector magnitude is directly measured using physiological sensors and has been validated as a good measure of physical activity; therefore, vector magnitude will be an appropriate metric for tasks that induce physical activity.

**Measurement implementation and analysis in prior evaluations.**

Vector magnitude was measured using the Bioharness chest strap heart rate monitor during both the Guided and Collaborative evaluations (Biopac Systems, 2013). Vector magnitude was recorded in offline mode on-board the sensor. Vector magnitude is determined in the Bioharness software to be the combination of acceleration data from the three axes of movement integrated during each 1.008 second time interval. The results associated with the Guided and Collaborative evaluations, including a comparison across the evaluations, are presented in Chapter III.

**Limitations.**

Vector magnitude is only relevant for tasks that include walking movement. If physical activity is not relevant to the task, the vector magnitude measurement may be unnecessary. Additionally, the metric will include behavior not associated with the task, such as pacing, which may not be desired.

**Implications for future human-robot teams.**

The human's physical workload can be monitored by a robot during physically active tasks. Vector magnitude provides information on the amount of walking (or running) performed by a human teammate. A robot can monitor how long the human teammate has sustained high levels of vector magnitude and offer breaks at appropriate times. If vector magnitude data drastically decreases, the robot can check with the human teammate regarding injury or fatigue.

### IV.1.7 Movement Count

The movement count metric is an observational count of the number of times a participant enacts a specific physical action during a period of time. Higher movement count implies higher physical activity.

**General Guidelines for Use.**

The movement count is a direct measure of physical workload. Movement count can be used to determine how often an important movement of interest occurs, or how often extraneous movement occurs. Movement

count can be used to count a variety of movement types, (e.g., arm gestures, head nods, crouching, turning). It is necessary to specify the exact movement specifications before collecting movement count data.

**Measurement implementation and analysis in prior evaluations.**

Movement count was measured in the Guided evaluation. Movement count summed the total number of times participants either crouched down next to a victim or stood up from a crouched position. Movement count was collected via video coding. Movement count was not collected or analyzed in the Collaborative evaluation, because bending and crouching was not an integral aspect of the primary task. The results associated with the Guided evaluation are presented in Chapter III.

**Limitations.**

Limitations of movement count include defining the characteristics of the movement of interest. Participants can be trained to raise their arm straight in the air to indicate particular information. Some participants may, instead, raise their arm out to the side or bent at the elbow. A human evaluator may recognize the intent of the gesture, but it is important to ensure that the robot can distinguish between movements and discern correct movements to count.

**Implications for future human-robot teams.**

A robot can count the occurrences of a specific movement using visual sensors and prior knowledge of visual characteristics of the movement of interest. Many robotic systems can be equipped with gesture recognition capabilities. For example, sensors currently exist, such as the Xbox Kinect, which enable body recognition within an acceptable distance from the sensor in indoor environments. Upon the instance of a movement or gesture of interest, the robot can stop or start a specific behavior. The robot may not be able to spend time and resources (i.e., sensor, computational) measuring counts of specific gestures when it is necessary to perform its own tasks.

### IV.1.8 Working Memory Recall

Working memory recall refers to the ability of a participant to correctly remember items of focus from the evaluation and is calculated by computing the ratio of correctly remembered items to total possible items. Working memory is a multi-component system of short-term memory storage-activated memory along with a focus of attention on it and central executive processes to manipulate this stored information (Cowan, 2008). Working memory takes activated stored memories into short-term memory. Once the stored memories are in short-term memory, they can be used and manipulated.

Working memory remains active as it is used, and workload increases as working memory demands increase (Berka et al., 2007). Working memory is more active when workload is higher; therefore, working memory recall will be higher with higher workload, until the point at which the human becomes overloaded. Higher demand primary tasks load the working memory central executive and limit automatic processing of novel visual stimuli (Spinks et al., 2004). This finding indicates that as workload increases and demand on the working memory increases, fewer novel visual stimuli can be identified.

The method of testing memory recall typically involves requiring participants to write down or fill in information from the task following the end of the evaluation. Working memory recall was tested to measure display effectiveness for novice and expert users (Vicente, 1992). Participants were tested immediately following the evaluation by providing participants a screen to fill in values for thirty-four variables monitored during the evaluation. Higher memory recall was associated with increased levels of subject expertise and better display effectiveness. Memory recall was also tested for participants memorizing a list of vocabulary words (Gais et al., 2006). Participants memorized a list of 24 vocabulary German-English word pairs for 10 minutes and returned 24 hours later or longer to write down the list of German words associated with the provided English words. The study demonstrated low levels of memory loss, even when the time period between memorization and recall was 48 hours.

**General Guidelines for Use.**

Working memory is an indirect measure of workload. Working memory recall levels require that participants recall as many specific items of interest as possible immediately following a task. The ground truth of which responses are correct must be determined. Participants can be given a list to explicitly memorize before the task, or the memory recall can be of items encountered or values monitored during the task without explicit request of memorization. This choice is task-dependent where participants are asked to write down as many recalled items, memorized items, or monitored variables as possible. Working memory level is computed by dividing the number of correctly recalled items by the total number of items available for recall.

**Measurement implementation and analysis in prior evaluations.**

Participants in both the Guided and Collaborative evaluations were tasked during the debriefing sessions with recalling items of significance from the evaluation. The results associated with the Guided and Collaborative evaluations are presented in Chapter III.

**Limitations.**

Requiring participants to memorize a list or pay attention to items during the evaluation may distract energy and attention from the primary task. Additionally, the choice of what the participants will recall may shape the results. Items that were not the primary task focus may not be remembered as well as items that were.

**Implications for future human-robot teams.**

Following the end of a task, the robot can ask a human teammate to recall information from the task or a memorized list. The robot will need speech recognition software and knowledge of the correct items or values on the list to be recalled. The robot can compute the working memory level and assess the status of the participant before beginning a new task. Low levels of memory recall can indicate exhaustion or workload overload and may indicate that the teammate needs a break.

### IV.1.9 Task Density

Task density is an calculation of the number of tasks initiated during a specific period of time (Weinger et al., 1994).

**General Guidelines for Use.**

Task density is useful to compare the amount of work accomplished in a given amount of time, including examining the overall workload that the participants experienced. It is useful to provide an objective, observational calculation of workload to supplement subjective or physiological measurements.

**Measurement implementation and analysis in prior evaluations.**

Task density was computed for the Collaborative model and the Collaborative evaluation. The subtask activities were considered searching an investigation area. The search was represented as a product of the size of the area and the number of items found in that area. The subtask time was the length of time taken to investigate the area, which is the metric subtask time (see Chapter IV.2.5). The measurement technique used to analyze task density is described in Chapter III.2.3.1 and used Equation III.1. The results associated with the Collaborative evaluation are presented in Chapter III.

**Limitations.**

Keeping the measurements consistent and comparable throughout a mission can be difficult if the team is switching between a variety of tasks or the missions make it difficult to measure how many tasks are initiated (e.g., vigilance, mental computations). For example, in the Collaborative evaluation, the task density calculation was based on searching an area, and in order to include the search as a task, the calculation included

area size. Task density is similar to work completed (see Chapter IV.2.7) in that it is based on workload calculations.

**Implications for future human-robot teams.**

Task density has the advantage of computing a workload value without using subjective ratings. The robot does not have to pause to ask their human partner for subjective ratings of workload demand throughout the task, or as often. The robot can track the number of tasks completed and how long it takes to complete the task. This measurement can work to compare workload, as long as the measurements of work and time are consistent in order to be compared.

### IV.1.10 Speech Rate

Speech rate is the articulation and pause rate of verbal communication (Baldwin, 2012). Individual differences in speech rate are vast; speech rate affects the listener's speech processing and mental workload. Speech rate can also reflect the speaker's internal state, including workload and stress levels.

Speech rate can be measured using syllables per second to exclude pauses between words, or using words per minute to include pauses between words (Baldwin, 2012; Strurm and Seery, 2007). Typical slow speech is around 136 to 144 words per minute (3.36 to 4.19 syllables per second) and fast speech is around 172 words per minute (about 5.29 - 5.90 syllables per second) (Venkatagiri, 1999).

Speech rate has been measured to affect the listener. Higher speech rates increase the difficulty of speech processing (Nygaard and Pisoni, 1995). Additionally, speech rate has demonstrated internal state changes in the speaker. High stress and high workload tasks (e.g., air traffic control) result in faster speech rate (Taylor et al., 1994). Speech production rates have been shown to generally increase with mental workload demand (Brenner et al., 1994).

**General Guidelines for Use.**

A comprehensive speech rate will be very difficult and time consuming to compute manually, but it is possible and has been computed. Audio coders can time each speech utterance, compute the number of words, or syllables, spoken during the utterance and compute the associated words, or syllables, per unit time.

Speech recognition software may assist in computing speech speed by automating the word counting process. Programs such as Dragon (Nuance, 2014) and iSpeech (iSpeech, 2013) can convert speech to text, and the number of words can be counted and divided by the length of time of the utterance. This process may be successful but requires significant effort. Other constraints may need to be met, including securing an environment with little background noise, training the software to the specific speaker's voice, and/or ensuring

the speaker's clear enunciation. These requirements are not often met in mobile human-robot interaction outside of laboratory environments.

**Measurement implementation and analysis in prior evaluations.**

The measurement of speech rate in the Guided and Collaborative evaluations was not performed due to the current limitations of speech rate measurement software. Existing speech recognition and speech-to-text tools were investigated, but no tool was able to successfully parse the audio from the recordings of the Collaborative or Guided evaluations. The development of a customized tool using a toolkit, such as Carnegie Mellon University's Sphinx tool (Skantze and Schlangen, 2009), required too extensive an undertaking at this stage in the research; thus, speech rate results have not been collected from the Guided and Collaborative evaluations.

**Limitations.**

Speech rate cannot be precisely measured in real time by a human teammate or observer. Additionally, the differentiation between words may be more difficult with background noise and faster speech rate. Current automated speech recognition systems (i.e., those that may be used post-hoc or mounted on a robot) are not error free. Many systems are improved by using a training set of the human speaker's voice, which may not be possible. A speaker's voice under normal conditions may also sound different than when the speaker is under high workload or stressful conditions during a team task, which may worsen the quality of speech recognition. Speech rate may be affected by the error in speech recognition more than changes in workload.

**Implications for future human-robot teams.**

Theoretically, using a speech recognition software system, a robot can measure speech rate in real time. This achievement assumes that the robot can clearly identify breaks between words. Additionally, the speech recognition system likely requires a microphone on the human. Microphone systems exist, such as bone microphones, which permit the capture of speech without ambient noise. A robot may make use of the bone microphone output to capture the human's speech in a real-world environment. Additionally, a robot may potentially measure syllable rate if parsing the actual speech into words is more difficult than parsing the beats between syllables.

#### IV.1.11 Secondary Task Failure Rate

Secondary task failure rate refers to the ratio of incorrect responses to total prompts of a secondary task. Secondary tasks include constant monitoring tasks (e.g., driving simulator, verbal recital of a distractor noise or word) or discrete prompted tasks, such as card sorting, memory recall, or mental math problems (Gawron,

2008). Secondary task failure rate for constant monitoring tasks is the ratio of times that the secondary task completely stops or fails to meet an upkeep threshold, while failure for a discrete prompted task is failure to respond correctly.

The secondary task is a task performance metric designed as an index of workload (Eggemeier and Wilson, 1991) and can be used to represent workload for two reasons: 1) simulating parts of the real-life task that is missing and 2) determining the time and work an operator can spare while still performing the primary task (Wierwille et al., 1977). Secondary task performance represents spare mental capacity, which is the difference between the human's total workload capacity and the capacity needed to perform the primary task, (Williges and Wierwille, 1979). An increase in available spare mental capacity from the primary task implies a decrease in workload resulting from the primary task. The more workload a primary task elicits, the lower the expected performance on a secondary task will be; therefore, an increase in secondary task failure rate implies an increase in workload level (Gawron, 2008; Mehler et al., 2009).

**General Guidelines for Use.**

Secondary task failure rate is a direct measure of workload. The definition of a task failure depends on the type of secondary task. Tasks, such as memorization and recall tasks, mental math, word recognition, all require a verbal response to a prompt. The response given by the participant can be compared to the desired correct response and determined to be a success or failure. Monitoring tasks, including reciting distractors, require determining a failure threshold. For instance, in a distractor reciting task (i.e., a common secondary task is to require participants to repeat a phoneme, such as "bah" throughout a tactile or visual primary task), the participant may fail if the rate of recitation falls below a threshold speed or stops. The failure rate can be determined by calculating the number of times the participant falls below the threshold if the experimenter provides a correction. Otherwise, the failure rate can be calculated by computing the ratio of time failing to total time performing the secondary task.

The secondary task failure rate can be determined for a discrete task by the experimenter or an observer during the task by recording the participant response and judging whether it was correct or incorrect. Similarly, coded video or audio recordings can be used to review participant responses after the evaluation. If the task is continuous and computer- or simulation-based (e.g., driving simulator), the failure rate can be determined through the system recording when interactions fall out of range of the set correct threshold. If the continuous task is performed via visible actions or audible responses, the analysis of failure rate can be performed using coded audio or video recordings.

**Measurement implementation and analysis in prior evaluations.**

Secondary task failure rate was computed for the Guided and Collaborative evaluations. The measurement techniques used to analyze secondary task failure rate and the associated results for each evaluation are described in Chapter III.

**Limitations.**

Higher secondary task failure rate for memory tasks can be related to the question difficulty, rather than the primary task's associated workload. It is important to consider the relationship between the complexity or familiarity of the word to be recalled or recognized, potential cognitive priming by repeating the word, and the order of questions asked during the evaluation. Randomizing the order of secondary task questions or analyzing the task difficulty may resolve these issues.

**Implications for future human-robot teams.**

Discrete secondary tasks with an audible response may be judged as incorrect or correct by a robot with on-board speech recognition software. The robot can record the participant's response and judge against knowledge of ground truth whether the response is correct or incorrect. If the speech recognition software is not advanced, or there is the possibility of intrusive background noise, the robot can transcribe the response or record the audio of the response for a human to judge correctness at a later time.

Continuous audible tasks can be judged using speech recognition software as well. The rate of repeating a phoneme can be determined, and failure can be recognized when the rate falls below the set threshold. Visually-based secondary tasks (e.g., card sorting tasks) will require complex visual processing software and may not be feasible depending on the primary task, robot embodiment, and task environment. Lighting conditions and participant position or angle can greatly influence the ability of the robot to recognize correct or incorrect actions.

### IV.1.12 In Situ Workload Ratings

The in situ workload ratings task participants with providing Likert scale values from 1 (little or no demand) to 5 (extreme demand) representing current workload demand along six channels (i.e., the Auditory, Visual, Cognitive, Speech, Tactile and Motor channels). The subjective ratings are requested at the completion of a task or subtask. This method of measuring subjective workload is adapted from the Multiple Research Questionnaire (Boles and Adair, 2001). High in situ workload ratings indicate high levels of workload.

**General Guidelines for Use.**

In situ workload ratings are a direct measure of mental workload. The evaluator must decide how often to administer the rating questions. Tasks with clear beginnings and ends (e.g., victim assessments in the Guided evaluation, investigation areas in the Collaborative evaluation) allow participants to switch from performing the task to rating workload channels with little task intrusion. Asking for in situ workload ratings too often may overwhelm the participant and prevent clear separation of tasks' associated workload values in the participants' ratings. Collecting in situ workload rating values only once or twice during an evaluation with many subtasks may not provide discernible differences between tasks or conditions.

**Measurement implementation and analysis in prior evaluations.**

The in situ workload questions were asked by the participant's responder partner (i.e., human or robot) in both the Guided and Collaborative evaluations. The questions were explained to participants during training. Responses were collected following each victim assessment in the Guided evaluation and following the search of each investigation area in the Collaborative evaluation. The values of each of the six channels were summed for each assessment to form total in situ workload rating values. The results associated with the Guided and Collaborative evaluations, including a comparison across the evaluations, are presented in Chapter III.

**Limitations.**

The limitations of the subjective ratings include the lack of continuous measure, the dependency on choice in administration time, and the need for training time to explain the rating system. Another limitation is the nature of the subjective rating; this measure relies on the participant judgment. Participants may not be conscious in subtle changes to their own workload, or may be unable to tie workload changes to a particular workload channel. Additionally, in an actual first response or other mission-critical task, in situ subjective workload questions may not make sense in relation to the primary task at hand. Interrupting actual triage tasks, for example, is not realistic, given the pressures of the task. These limitations prevent the in situ subjective workload ratings from being an ideal workload measurement in the field.

**Implications for future human-robot teams.**

A robot can ask a human teammate to provide in situ workload ratings, and spoken values from the participant can be parsed by the robot and logged. A robot can evaluate the human teammate's state using the human's judgment of his or her current workload. Speech processing capabilities on the robot are required for verbal measurement.

### IV.1.13 NASA Task Load Index

The NASA Task Load Index is a validated subjective workload measurement tool (Hart and Staveland, 1988) that defines workload as a weighted mean of subjective ratings along six channels of workload demand: mental demands, physical demands, temporal demands, own performance, effort, and frustration (see Chapter II.2.1). High Task Load Index values indicate high levels of workload.

**General Guidelines for Use.**

NASA Task Load Index responses are a direct measure of workload that can be administered via pen and paper or computer interface. Participants respond to a series of questions regarding the level of workload experienced on each channel and the relative weight of each channel. The tool is typically administered following a task during the evaluation or following the entire evaluation.

**Measurement implementation and analysis in prior evaluations.**

The NASA Task Load Index tool was administered via a computer-based survey tool following the completion of the Guided and Collaborative evaluations (Sharek, 2009). The survey tool saved each participant's raw data along each channel in addition to a weighted value based on channel choice comparisons. The results associated with the Guided and Collaborative evaluations, including a comparison across the evaluations, are presented in Chapter III.

**Limitations.**

A limitation of this tool is its current tie to written administration. Auditory adaptations of the tool can be investigated and may be more appropriate for administration during the evaluation, following the completion of each subtask. Additionally, the limitations of the tool include its lack of continuous measurement, the time taken to complete the survey, and the subjectivity of the measurement. Participants may not be able to recall specific workload experienced during a task if they provide the rating after finishing the task.

**Implications for future human-robot teams.**

The NASA Task Load Index is performed using a visual interface. Participants mark, along a horizontal scale, the amount of workload experienced in each channel. The series of channel comparisons also occurs via visual comparison. An audio version of the NASA Task Load Index can be created to allow a robot to administer the assessment verbally.

### IV.2 Task Performance Metrics

Task performance is an important aspect of assessing the team's accomplishments (see Chapter II.2.2). This section details eight task performance metrics, including metrics representing task failure rates, timing aspects (i.e., reaction time, response time, subtask time), and work completed. Each metric is presented with a definition, general guidelines of use, a description of the implementation in the Guided and Collaborative evaluations, analysis results from the Guided and Collaborative evaluations, approaches for implementing measurement in future human-robot team, and associated limitations.

#### IV.2.1 Primary Task Reaction Time

Primary task reaction time is the time taken to immediately act upon an assigned primary task without allowing for an extended thought. Reaction time has been measured to represent the speed of mental processing of the onset of a stimulus in a primary task and was extensively described in Chapter II.2.2.1. Computing the reaction time to secondary tasks is common (Gawron, 2008).

**General Guidelines for Use.**

The primary task reaction time measurement requires the ability to time the difference between the exact onset of a task stimulus and a participant's first reaction. This timing can be accomplished in real-time if the task onset and reactions are based with a computer or system that can perform precise timing. An observer cannot note task reaction times during the evaluation using a stopwatch, because the observer will incorporate his or her own reaction times into the measurement. If audio or video recordings are available and the participant is not interacting with a device that can directly record task reaction time, video or audio coding is most appropriate. Clear instructions must be made available for the coders outlining exactly when the stimulus onset and reaction is marked. For example, the onset of a stimulus of light flashing may be either the moment when the light comes on or when the light goes off. The video coding instructions must clearly indicate which time is measured.

**Measurement implementation and analysis in prior evaluations.**

Primary task reaction time was measured in both the Guided and Collaborative evaluations using video coding. Primary task reaction time was collected via video coding for the H-R condition in the Guided evaluation, and both the H-H and H-R conditions in the Collaborative evaluations. The H-H condition in the Guided evaluation was not video coded due to lack of sound recording from the human responder partner.

The video coders recorded the onset time of the primary task stimulus and the time of the first reaction to that stimulus from the participant. The primary task stimuli in the Guided evaluation included requests to

perform triage-related tasks and questions related to the triage process. Primary task stimuli in the Collaborative evaluation included requests related to the discovery of items in the search area. The primary task reaction time was calculated by subtracting the time of a specific reaction to a primary task stimulus from the specific time of that primary task stimulus onset.

The Guided evaluation's primary task stimulus onsets were the last moment of a task prompt or question. The reaction was marked by the beginning of an action or a verbal utterance. The Collaborative evaluation's task stimulus onsets were the moment when an item to be investigated entered the participant's view, seen by the video recording made by the point-of-view camera. A reaction was considered to be a verbal utterance, the use of a laser pointer to identify the item, touching the item, fixating the view directly on the item, or beginning to take a photograph of an item. The results associated with the Guided and Collaborative evaluations are presented in Chapter III.

**Limitations.**

The limitations of measuring primary task reaction time include video frame rate, human video coder judgment, or difficulties in automatic system measurement. Video frame rate refers to the number of video frames recorded per second. The frame rate dictates the smallest time measurement that the video recording contains. The frame rate of some cameras is small enough to register the smallest differences in reaction time, while some small, less expensive cameras have lower frame rates and lower sensitivity. Human coder judgment is the most limiting factor in measuring task reaction time. Using video coding as the reaction time measurement technique requires human coders to determine the starting and ending times of the reaction subjectively. It is important to have more than one coder determine reaction times and ensure that the inter-coder reliability is acceptable. Automatic system measurement restricts the task to controlled environments (e.g., laboratory settings) or controlled reactions. This restriction may not be a problem for many tasks, but in the case of human-robot interaction, ensuring that the robot can register the onset of stimulus time and time of stimulus reaction without ambiguity may require advanced software and/or task restrictions. Additionally, background noise can interfere with any assessment of audio reactions. Poor lighting conditions can interfere with the precise assessment of video.

**Implications for future human-robot teams.**

Primary task reaction time may be possible to measure on the robot in real-time. Tasks requiring an audible response may be most desirable for recognizing onset of stimuli and reactions. The robot may be able to record the stimulus onset time (e.g., the robot asking a question) and through voice recognition software, the robot can recognize simply that an audible reaction has been made. It does not need to understand the

speech in order to record the stimulus onset time or reaction time. Visible reactions (e.g., hand movement, head nodding) may require that a camera track the participant's movements and ensure that the participant is visible, which may be a more complex task.

### IV.2.2 Secondary Task Reaction Time

Secondary task reaction time is the time taken to immediately act upon a secondary task without allowing for an extended thought. Secondary task reaction time provides additional information to determine whether or not the response is correct. This metric suggests the number of mental steps taken to react to the task given the current demands of the primary task.

**General Guidelines for Use.**

It is recommended to choose a secondary task with a clear stimulus onset time and measurable response time, in addition to the guidelines outlined in Chapter IV.2.1. For example, a verbal cue and verbal response may be more unambiguously measured than the time that a participant glances at an object, given the specific task requirements and available sensors.

**Measurement implementation and analysis in prior evaluations.**

The secondary task reaction time was collected using video coding and calculated using the same method as for primary task reaction time, described in Chapter IV.2.1. The task stimuli were related to the secondary tasks. The Guided evaluation secondary task stimuli were the last moments of the posed recall secondary task questions concerning a memorized list of first responder names. The Collaborative evaluation secondary task stimuli included each instance of the last moment of the posed On List and Danger Level questions concerning the memorized list of chemical names and associated numerical danger levels. The reaction onset time in both evaluations occurred when a participant reacted to a question. The results associated with the Guided and Collaborative evaluations are presented in Chapter III.

**Limitations.**

Video coding is not an ideal mode of collection of secondary task reaction time data and is limited by the speed of analysis, granularity of recording, and potential accuracy of the data. Ideally, new ways of measuring secondary task reaction time will emerge for human-robot peer-based teams.

**Implications for future human-robot teams.**

Similar to the discussion of primary task reaction time measurement for future human-robot teams, the nature of the task will play a part in the measurement of secondary task reaction time. Tasks with unambiguous

stimuli and reactions will be most ideal for adoption for robot measurement.

### IV.2.3 Primary Task Response Time

Primary task response time is the time taken for a participant to execute a correct and appropriate response to an explicit prompt. Primary task response time includes primary task reaction time and the time taken to consider an appropriate response.

Primary task response time is similar to choice reaction time, because the participant chooses between available responses. However, task response time weighs responses from differing tasks, all with specific sets of appropriate responses. The time is determined by taking the difference between the last moment of the task onset and the first moment of a correct response. For example, if participants are verbally asked a question, the problem onset time occurs at the last moment of the question and the moment of the correct response occurs when the participant first begins to speak the answer to the question. Preliminary responses (e.g., saying "um" before saying the answer) and inappropriate responses (e.g., saying something unrelated in response to a specific question) are ignored, because the response time is determined using the time taken to reach an appropriate response.

Primary task response time was found to be a sensitive measure of workload in aviation tasks (Wierwille et al., 1985). The authors defined meditational reaction time to be the time from problem presentation to a correct response. Meditational reaction time can distinguish between different mental load levels. Higher workload tasks resulted in an increase to meditational reaction time. The authors deemed this metric to be an excellent method of assessing relative mental workload. The use of the term "reaction time" in meditational reaction time is not appropriate for this dissertation, since it includes time taken to reach a correct response from the clear onset of a given problem. Instead, this metric is referred to as primary task response time to uphold a distinction between reaction and response times.

#### General Guidelines for Use.

The primary task response time must have a clearly identified onset time and time of response. Verbal questions can provide clear onset time and time of response; recording the end of the last moment of the question and the first moment of the response provides the task response time. It is possible to compute task response times using visual task onset times and times of responses (e.g., participant touching a specific item once a light turns on), but determining the onset time and time of response may be more difficult than with purely auditory questions and responses, or a combination of the two (e.g., participant asked verbally to touch a specific item).

Given verbal questions and responses, the easiest way to determine task response time is through audio or

video coding the onset time and response times. Clear requirements must be given to coders for determining onset time (i.e., moment of the beginning or end of the question?) and time of response (i.e., ignore words, such as "um" if they occur before the response). It is recommended to use more than one coder, and the inter-coder reliability must be at an acceptable level.

**Measurement implementation and analysis in prior evaluations.**

Primary task response time was measured in the Guided and Collaborative evaluations. Measurements were recorded via video coding. Primary task response time was collected via video coding for the H-R condition in the Guided evaluation and both the H-H and H-R conditions in the Collaborative evaluations. The H-H condition in the Guided evaluation was not video coded due to lack of sound recording from the human responder partner. The H-R condition task response time results from the Guided evaluation are provided for reference.

The stimulus onsets for primary task response time measurements were identical to the stimulus onsets for primary task reaction time. The time of response was recorded when the participants provided an appropriate response to the question asked or performed the task requested by his or her partner. For example, if the responder partner posed the question, "How many breaths did the victim take?" and the participant responded with "um... what was the question? How many breaths? Oh. He took 40." The stimulus onset time is recorded at the moment when the responder finishes uttering the word "take." The time of response is recorded at the moment the participant uttered the word "He."

Primary task response time is computed as primary task reaction time plus any additional time to contemplate an appropriate response to the primary task stimulus. The additional time may equal zero. The results associated with the Guided and Collaborative evaluations are presented in Chapter III.

**Limitations.**

The limitations of determining task response time are based on the human judgment necessary to determine when an "appropriate" response is made, rather than any response. This limitation requires very clear definitions of the onset time and time of response for video coders. Additionally, task response time can only be measured for tasks with a clear onset time and clear time of response. If these two instances cannot be specified, the task response time may not be a relevant measure.

**Implications for future human-robot teams.**

It may be difficult for a robot to judge the onset of an "appropriate" task response, because robots do not have the innate ability to make judgment calls, as humans do. For example, each task response may be classified

by the robot as "appropriate" or "not appropriate" by rules learned in advance of the task. This limiting factor may prevent task response time from being a simple measurement to implement on a robot as a different metric than reaction time. Measurement will occur in the same way, but with additional task-dependent constraints to help the robot determine the difference between reaction and response times.

### IV.2.4 Secondary Task Response Time

Secondary task response time is the time taken for a participant to execute a correct and appropriate response to a secondary task. Secondary task response time includes the secondary task reaction time and the time taken to consider an appropriate response.

**General Guidelines for Use.**

As discussed in the guidelines for primary task response time (see Chapter IV.2.3), it is critical to identify a secondary task with an unambiguous task stimuli and measurable response time. A data collection technique (e.g., video recording, audio feed, button presses) must also be available to provide the feedback regarding appropriate secondary task responses.

**Measurement implementation and analysis in prior evaluations.**

Secondary task response time was measured in the Guided and Collaborative evaluations. Measurements were recorded via video coding. Secondary task response time was collected via video coding for the H-R condition in the Guided evaluation and both the H-H and H-R conditions in the Collaborative evaluations. The H-H condition in the Guided evaluation was not video coded due to lack of sound recording from the human responder partner. The H-R condition task response time results from the Guided evaluation are provided for reference.

The stimulus onsets for secondary task response time measurements were identical to the stimulus onsets for secondary task reaction time. The time of response was recorded when the participants provided an appropriate response to the question asked or performed the task requested by his or her partner. For example, if the responder partner posed the question, "Was Chlorine on your list of chemicals?" and the participant responded with "Chlorine. Let me think, let me think. No, it wasn't." The stimulus onset time is recorded at the moment when the responder finishes uttering the word "chemicals." The time of response is recorded at the moment the participant uttered, "no."

Secondary task response time is computed as secondary task reaction time plus any additional time to contemplate an appropriate response to the task stimulus. The additional time may equal zero. The results associated with the Guided and Collaborative evaluations are presented in Chapter III.

**Limitations.**

The largest limiting factors for measuring secondary task response time are 1) the subjectivity of measuring an "appropriate" response and 2) using video coding for collection of secondary task response time analysis. Overcoming these limitations requires highly sophisticated natural language processing, which is an unsolvable problem at this time.

**Implications for future human-robot teams.**

Due to the fact that secondary task response time measurement requires the recognition of an "appropriate" response to an assigned task, a robot teammate may need to know what responses to look for if it will be able to measure secondary task response time in real-time. Certain appropriate task responses may be more recognizable than others (e.g., gesture, keyword), but some tasks may include ambiguous responses. If potential responses are unknown or not pre-programmed, it may be infeasible to measure secondary task response time on-board a robot.

### IV.2.5 Subtask Time

Subtask time is the time taken to complete a single or distinct set of atomic tasks that comprise an entire mission.

**General Guidelines for Use.**

Subtask time measures how long it takes for an individual or team to complete actions of interest. This measurement can represent task performance and can, in turn, be used to ground assessments of workload. Timing a task is necessary to calculate workload. Steinfeld et al. (2006) presented common human-robot interaction metrics including subjective and objective mental workload metrics. For example, mental workload can be measured as the number of interventions (i.e., unplanned robot interactions) per unit of time. Alternatively, mental workload can be assessed as the ratio of operator engaged task time to robot execution time or fan out (Goodrich and Olsen, 2003). Subtask time must be measured in order to compute these metrics.

**Measurement implementation and analysis in prior evaluations.**

Subtask time was measured in both the Guided and the Collaborative evaluations. The start times and end times of each of the primary tasks (i.e., triage on each victim in the Guided evaluation and the search of each investigation area in the Collaborative evaluation) were recorded by video coders. The results associated with the Guided and Collaborative evaluations, including a comparison across the evaluations, are presented in Chapter III.

**Limitations.**

Timing measurements require knowledge of when to start and stop recording time. This type of measurement implies that a human or robot judges when the onset or stop behavior occurs during video coding or observation, leaving room for error. Automated timing can occur, depending on the task type and available sensors (e.g., a human interacts with a system that records when the human enters an area based on sensor readings).

**Implications for future human-robot teams.**

Robots are able to set timers to track the length of tasks by recording start and end times and logging the difference between them. The robot's assessment of whether subtask time is on track with acceptable levels will influence a behavior change; the robot will perhaps pursue follow-up measures to ensure thoroughness if the team finished significantly faster than expected, or may assess whether the human is still within an acceptable workload, fatigue, and/or stress range if the task took extremely longer than expected.

### IV.2.6 Primary Task Failure Rate

Primary task failure rate refers to the percentage of times the primary task was not properly performed out of the total number of task attempts. Primary task failure rate represents the completion quality of the primary task objective, or overall performance. Higher levels of primary task failure rate reflects poorer task performance.

Primary task failure rate has been determined to be a task performance measure that reflects workload levels. Wierwille et al. (1985) assessed primary task failure rate, termed meditational error rate, in a simulated flight task where the participants were tasked with maintaining a given altitude, airspeed and heading. Participants were asked questions as part of the primary task regarding their altitude, airspeed and heading. Meditational error rate is defined as the number of incorrectly answered and unanswered task prompts divided by total number of prompts. Meditational error rate was shown to represent changes in workload more effectively than respiration rate, pupil diameter, and heart rate.

Primary task failure rate has also been investigated more recently in predicting human task failure rates during flight tasks (Kunlun et al., 2011). The predicted task failure rate probability was validated by comparing the estimated failure rate, based on task analysis, to a failure rate calculated from past flight records. The failure rate was calculated by dividing the number of actual task failures by the number of opportunities to initiate the task.

**General Guidelines for Use.**

Primary task failure rate is a task performance measure. Using this measure requires that the failure of the primary task is clearly defined. If the primary task is driving in between obstacles, for example, failure can be determined when the participant collides with an obstacle. The number of collisions divided by the total number of obstacle avoidance attempts provides the primary task failure rate.

Primary task failure rate may be derived by determining the task's objective and a measurable way to identify failure of that objective. If the mentioned driving task were based upon speed rather than avoiding collisions, a failure can be determined, for example, by the time between obstacles. If the time between obstacles is slower than the desired time, that primary task is considered a failure. Depending on the nature of the task failure (i.e., tasks using a computer, robot, or system interface, verbal responses, visible physical actions) and available equipment (e.g., audio recording equipment, video recording equipment, availability of trained observers, system capabilities) the primary task failure rate can be computed in a variety of ways. A coder can observe each failure and each attempt through video or audio recordings after the evaluation has been completed, or through automatic system recordings (e.g., having the system save the number of collisions and total obstacles).

The primary task failure rate calculation requires determining the measurable objectives of the primary task. The evaluation may include more than one type of primary task component (e.g., meeting a time requirement, finding a specific number of tokens, giving correct responses to questions), and it can be useful to measure individual primary task failure rates for each primary task component, as well as a total primary task failure rate for the entirety of the primary task. Primary task components are determined by assessing the essential goals of the primary task and what skills they require. Different primary tasks have been generically termed decision tasks (both spatial and verbal), tracking tasks, detection tasks, memory tasks (recall and recognition), attention tasks, effort tasks, and integration tasks, for example (Kessel and Wickens, 1982; Wickens and Liu, 1988; Wickens and Andre, 1990; Gawron, 2008). These generic primary task component terms can be used in order to analyze primary task failure rate between evaluations.

**Measurement implementation and analysis in prior evaluations.**

Primary task failure rate was measured in both the Guided and the Collaborative evaluations. Each evaluation contained four principal components of the primary task with a measurable task failure rate. The primary task components each had a distinct task element: attention, effort, assessment, and detail-tracking. The results associated with the Guided and Collaborative evaluations, including a comparison across the evaluations, are presented in Chapter III.

**Limitations.**

A high primary task failure rate can possibly represent workload underload, rather than high levels of workload. It is important to have a baseline prediction of expected workload levels during the evaluation. If the highest task failure rate is during tasks with the lowest expected workload, it may be due to workload underload. Additionally, a high task failure rate may be attributed to a participant lacking an understanding of the task. The driving example may need to clearly state if the task requires not colliding with obstacles or completing each task in a certain timeframe. If the participant does not understand the task requirements, a high primary task failure rate can represent the lack of understanding, not high workload.

**Implications for future human-robot teams.**

Measurement of primary task failure rate can be achieved in future human-robot teams. If, as in the Guided evaluation, the task requires assessing the correctness of verbal responses from the participant, a robot partner with speech recognition software can capture and parse the participant's response and compare it to the correct response, recording it as a failure, or not, in real-time. This method will require the robot to possess knowledge of the correct response to a question, assessment questions that have only one correct response, and a high quality speech recognition system to recognize the human's responses. Tasks that require visual assessment of correctness, (e.g., the proper assembly of an object), will require the robot to use image processing software to visually assess the shape of the object. However, this will also necessitate that the robot be able to determine the placement of the object and its viewing angle of the object, which may not be possible in a real-world scenario.

Another example of moving the assessment of primary task failure rate aboard a robot partner is a path traversal task. If a human-robot team is deployed to travel a given route to a destination, the objective of a primary task can be to remain on the most efficient path to a goal for the entire time on task. The robot will have a map of the area and compute the distance between the desired path and current path. If the distance crosses an acceptability threshold, the time where the distance between the desired and actual paths is above a threshold will be recorded. The ratio of time off the path and total time will be computed as the primary task failure rate. The robot can easily record the total task time, and also record the time spent off the desired path in order to compute this ratio.

### IV.2.7 Work Completed

Work completed computes an estimate of the task load a human accomplishes during a task. Work completed is a measurement that factors out the time taken to complete a task from the computed workload values and is a task performance metric.

**General Guidelines for Use.**

Workload is the amount of work completed in the amount of time available to complete it (Wickens et al., 2003). Based on this definition, an equation can be formed to calculate workload and solve for the work completed. The model predictions of workload or measured workload values are multiplied by modeled or recorded task completion time.

**Measurement implementation and analysis in prior evaluations.**

Work completed was calculated for the Collaborative evaluation using the modeled workload results and timing values. Additionally, work completed was calculated using in situ workload ratings and measured subtask time results. Work completed was calculated by multiplying the total in situ workload ratings and the subtask times for each investigation area and the measurement techniques are explained in Chapter III.2.3.1. Work completed is measured using Equation III.2. The results associated with the Collaborative evaluation are presented in Chapter III.

**Limitations.**

Work completed is not a validated metric on its own; it is based on the metric of workload. Work completed offers the advantage of factoring out time from workload, but possesses the disadvantage of having arbitrary units. Arbitrary units are a disadvantage because they are only meaningful for the specific calculation.

**Implications for future human-robot teams.**

Work completed calculations can be used in human-robot teams in order to assess whether two missions elicit similar demands on a human. Additionally, a robot can use expected levels of work completed for a task based on a model and compare the expected values to computed real-time work completed values. The robot can measure task time completed and workload levels in order to compute the real-time work completed value and assess whether the current work completed value is reasonable or too low based on the modeled value for the current task.

### IV.2.8 Distance Traveled

Distance traveled is measured by the span of physical space covered by a participant during a task. Distance traveled has been measured in human-computer interaction and virtual reality experiments to evaluate task performance (Czerwinski et al., 2002; Riecke et al., 2010; Riggs et al., 2014).

**General Guidelines for Use.**

Distance traveled can be measured in a variety of ways, such as virtual distance traversed in an interface (Czerwinski et al., 2002; Riggs et al., 2014), distance measured by a pedometer (Witmer et al., 1996), or the distance as measured by the global positioning system (GPS) or geographic information systems (GIS) (Duncan and Mummery, 2007).

**Measurement implementation and analysis in prior evaluations.**

Distance traveled was not successfully measured in the Guided evaluation. Tri-axial accelerometer data was collected from the Bioharness chest-strap sensor in an attempt to gather the distance traveled in each two-dimensional direction; however, when the data was analyzed, it was discovered that only the peak acceleration for each direction, and overall, was recorded for 1.008-second epoch, rather than the mean velocity or acceleration. Peak acceleration data was analyzed, but the choice was made not to move forward with this analysis in the Collaborative evaluation.

The precise calculation of distance traveled was not accurate with peak acceleration information alone; thus, a new sensor was purchased for the Collaborative evaluation to measure distance traveled. Vector magnitude (see Chapter IV.1.6) is a measurement that uses acceleration information to calculate an approximation of gross movement over time. Distance traveled was measured in the Collaborative evaluation using a Garmin footpod pedometer designed for measuring distance traveled while exercising. The pedometer requires wearing a watch and attaching a small sensor to the laces of the shoe. The results associated with the Collaborative evaluation are presented in Chapter III.

**Limitations.**

Distance traveled is only relevant for tasks where participants are moving around an area. GPS-measurement devices do not have the resolution and accuracy to measure small distances and GPS signals are not always available in urban (i.e., around skyscrapers), indoor, or remote areas. Sensors that measure gross distance traveled via steps taken, for example, may offer a more reliable measurement than measuring the distance traveled by tracking the human's location in the environment via GPS.

**Implications for future human-robot teams.**

Real-time transmission of distance traveled information to a robot may provide useful information for the team. The robot may make estimations of the human's location based on the distance traveled. The addition of a direction of travel and logging the human's time of movements can help to estimate the location of a human on an internal map for the robot. Systems like GPS do not have the resolution and accuracy to

measure small distances, and are not available indoors or in remote locations; thus, having other options for tracking the human's location by distance traveled can aid the robot in finding its partner or relaying location information about him or her to other teammates.

Additionally, the robot may assess whether or not the human is traveling the correct distance for a task (e.g., searching an environment). If the human partner did not travel enough distance to finish the task, for example, the robot partner may be able to prompt the human to finish the task.

### IV.3 Discussion

This chapter presented a variety of objective and subjective metrics that have been analyzed using data collected in the Guided and Collaborative evaluations. All of the metrics have benefits and drawbacks that are exhibited when adapted from other domains and utilized for human-robot interaction. Analysis of the metrics in this Chapter aided in the design of the Time-Structured evaluation (see Chapter V).

The physiological measures (i.e., heart rate, respiration rate, heart rate variability, vector magnitude, variance in posture, postural load) are influenced by physical activity. Vector magnitude, variance in posture, and postural load are used for measuring physical workload. Heart rate, respiration rate, and heart rate variability demonstrated inconsistent results in the Guided and Collaborative evaluations measuring mental workload. One potential reason is that neither evaluation induced extreme levels of high mental workload. The second potential reason is that physical workload prevented the accurate measurement of mental workload by the physiological measures; however, the physical activity levels in both evaluations were not strenuous. The Time-Structured evaluation investigated a correlation-comparison of high and low levels of physical workload with high and low levels of mental workload in order to assess changes in the physiological measures. The Time-Structured evaluation's assessment technique provides further information regarding the use of physiological measures of mental workload in mobile human-robot interaction.

One major limitation of using the Bioharness chest strap monitor to record physiological data is the nature of wearing a chest strap under the clothes. The sensor must be placed correctly against the chest underneath all other clothing and protective equipment. The sensor can easily become dislodged while moving around or donning gear without the human's knowledge. Experimenters in the Guided and Collaborative evaluations did not know if the sensor had moved out of place during the evaluation, because of its location under the clothes, and the fact that the sensor was not in live-streaming mode. Additionally, chest straps are not comfortable to wear, and participants voiced complaints. The Bioharness chest strap sensor is also significantly more expensive than athletic activity monitors.

Additionally, the Guided and Collaborative evaluations both featured attempts to measure distance traveled; the Guided evaluation analysis used the Bioharness acceleration data and the Collaborative evaluation

analysis used a Garmin GPS foot pod pedometer. Neither analysis was successful. Research was performed following the Guided and Collaborative evaluations (summer 2014) to find commercially available sensors to record heart rate and distance traveled without the use of a chest strap sensor for heart rate or GPS measurement for distance.

Commercially-available sensors most often do not provide the resolution necessary to record accurate heart rate variability (i.e., sampling rate of 250 Hertz or greater); thus the goal was to find a sensor to measure heart rate (i.e., often reported at 60 Hertz). Raw heart rate is a metric that can be easily recorded, but many sensors do not provide continuous measures; they provide moving averages, which are not useful for fine-grained analysis. Many sensors also only provide heart rate readings when a fingertip is placed on a small sensor (e.g., Phyode w/Me (Rooti Labs, 2014)), or if the participant is physically active (e.g., Basis Steel (BASIS, 2015)). An additional concern for several heart rate sensors was the ability to access the data recorded by the monitor, as several monitors do not have Bluetooth, ANT+, or an easy method of synchronizing the monitor to obtain heart rate. The sensors were narrowed down to the Mio Alpha (Mio, 2014) and the Scosche Rhythm+ (Scosche Industries, 2014). The Scosche Rhythm+ was purchased due to its lower price point.

The Fitbit Zip (Fitbit, 2013) was chosen to measure steps taken because of its simplicity, low price point, Bluetooth syncing to mobile devices, and the ability to access data via the Fitbit website. Other sensors for distance traveled included data that was unnecessary for this investigation, such as sleep trackers, ability to send notifications from smartphones, or complex food-tracking applications (e.g., Nike Fuelband (Nike, 2015), Jawbone Up (Jawbone, 2013)). The Scosche Rhythm+ also provides GPS-based distance traveled, which was recorded. As mentioned, GPS-provided location data is unreliable indoors, so the data was not the primary source of distance information.

Subjective measures (i.e., in situ workload ratings, NASA Task Load Index survey) indicated similar trends to one another; however, the manipulated levels of workload in the evaluations were not extreme (i.e., triage level, investigation index). The Time-Structured evaluation was designed to pose stark differences between low and high workload tasks.

Primary task failure rate analysis demonstrated that the same four primary task components were present in the Guided and Collaborative evaluations: attention, effort, assessment, and detail-tracking. Similar trends were found in the results of the failure rates between conditions in both evaluations. Participants paired with the robot partner had significantly higher primary task failure rate for effort component tasks in both evaluations (i.e., checking for injuries on the victims, taking photographs of suspicious items). Participants may fail more at these tasks because there were no direct consequences for their actions (i.e., due to the robot's lack of ability to notice failures), an attempt to speed up the overall task because human-robot evaluations

took a longer time, or because human-robot condition participants experienced lower levels of workload and expended lower levels of effort as a result. The Time-Structured evaluation investigated effort in assigned tasks in relation to primary task failure rate.

Secondary task failure rate showed no effect of condition. The failure rate was extremely low in the simple secondary task of recognizing first responder names in the Guided evaluation and was higher in the Collaborative evaluation with a more complex task of a two part question of recognizing a chemical name and recalling an associated danger level. The Time-Structured evaluation posed a more difficult secondary task and included different difficulty levels between low and high workload tasks in order to gauge workload and task performance differences between conditions.

Working memory recall was slightly better in the H-R condition, as demonstrated by the Collaborative evaluation results, and the trend in the Guided evaluation results. Vicente and Wang (1998) explained that even after brief exposures to a domain, higher memory recall of a task is associated with higher levels of domain expertise. This theory implies that the human-robot participants may be gathering higher levels of domain expertise than the human-human condition participants, due to their partner assignment. This result may be due to the fact that the human-robot participants spent more time in the domain than the human-human condition participants, or it may be that having a robot partner increases self-reliance for building knowledge of the environment. The longer duration that the human-robot participants spent in the domain has influenced the interpretation of many of the results of the Guided and Collaborative evaluations and stressed the necessity for imposing a time limit for tasks. The Time-Structured evaluation assigned timed tasks in order to address these concerns.

Moving away from video coding will require on-line sensors on the human and on-board sensor perception and processing on the robot. The presented research does not aim to advance sensor and robot technology in order to advance the live-sensing technology field.

The aim of this dissertation is to discern more information about trends in workload and task performance in human-robot peer-based teams, and identify appropriate metrics for measuring workload and task performance in mobile human-robot peer-based teams. There are benefits and drawbacks to working with a robot and the precise impact on workload and task performance has not yet been identified. The Time-Structured evaluation used the presented metrics in order to form a more complete picture of how the human is affected by a robot partner. This knowledge can be used in the human-robot interaction domain for developing future robot teammates and predicting human-robot team performance.

# Chapter V

## Time-Structured Evaluation

The Time-Structured evaluation evaluated workload and task performance metrics in human-robot peer-based teams in direct comparison with human-human teams. A subset of the metrics discussed in Chapter IV were collected and analyzed in order to assess workload and task performance.

The evaluation scenario required training the participants as an emergency responder, and was centered on the steps necessary to organize information, search for hazardous materials, and collect samples of hazardous materials. The four tasks were ordered in a manner to represent the general order of response steps to a disaster event and were focused on tasks in which robots are potential peers for first responders (Humphrey and Adams, 2011).

Participants were paired with either a human or robot partner in order to complete four tasks. Each task was completed in either a low workload or high workload state. There was a fifteen-minute time limit for each task, regardless of partner (i.e., human or robot) or workload state (i.e., high or low). Each task was timed in order to create time pressure and demonstrate how much of the task each team was able to achieve in a set amount of time.

The experimental hypotheses are:

- $H^1_{TS}$ : The low workload tasks will elicit lower levels of workload than high workload tasks.
- $H^2_{TS}$ : The human-robot team workload will be lower than the human-human team workload due to slower task. completion.
- $H^3_{TS}$ : Human-robot teams will have lower task performance.
- $H^4_{TS}$ : Workload levels will measurably impact task performance.

These hypotheses are based on the results of the previous evaluations. $H^1_{TS}$ is the primary hypothesis and represents verification that the workload measures are sensitive to workload changes. Metrics that measure changes in workload due to experimental manipulation from the Time-Structured evaluation's design demonstrate the ability for use in human-robot peer-based teams. Workload metrics that are unable to detect the difference between low and high workload tasks do not hold promise for use in mobile human-robot peer-based teams.

Table V.1: Randomly assigned combination of low and high workload tasks completed by each participant.

| Participant | Task 1 | Task 2 | Task 3 | Task 4 |
|---|---|---|---|---|
| 1 | Low | Low | Low | Low |
| 2 | Low | High | Low | High |
| 3 | High | Low | Low | Low |
| 4 | High | High | High | High |
| 5 | High | High | High | Low |
| 6 | Low | Low | High | Low |
| 7 | Low | Low | High | High |
| 8 | High | High | Low | High |
| 9 | Low | High | High | High |
| 10 | Low | High | High | Low |
| 11 | High | Low | High | Low |
| 12 | High | Low | High | High |
| 13 | Low | High | Low | Low |
| 14 | High | High | Low | Low |
| 15 | Low | Low | Low | High |
| 16 | High | Low | Low | High |

## V.1 Experimental Method

### V.1.1 Design

The design is a repeated measures design with the responder partner and the tasks as within-subjects elements. The participants completed four tasks with both a human partner (H-H) and a robot partner (H-R). Two sessions were completed, one with each partner and the team completing all four tasks. The first task, the photo searching task, involved determining whether photographs contained hazardous or suspicious items. The second task, the item search task, involved searching a hallway for hazardous or suspicious items. The third task was the solid contaminant sampling task and required participants to follow a hazard collection protocol for solid samples. The fourth task was the liquid contaminant sampling task and participants followed a similar hazard collection protocol to the third task, but adapted for high toxicity liquid hazards. The tasks were performed in the same order each time for each participant and team pairing. Each participant had a randomized combination of low and high task workload levels, shown by participant in Table V.1. The evaluation partner presentation was counterbalanced; thus, half of the participants completed the tasks first with the human partner and next with the robot partner, while the other half of participants completed the tasks with the robot partner first and with the human partner during their second session.

### V.1.2 Environment

The evaluation took place in an academic building on Vanderbilt University's campus. All tasks were close to one another on the same floor of a single building. Training took place in a small office with minimal distractions. The photo search task also occurred in the same room. The item search task occurred in the

hallway where people unrelated to the evaluation were able to walk through the environment. Sound traveled into the area from nearby offices, classrooms, and laboratories. The solid contaminant sampling task occurred in an engineering laboratory isolated from foot traffic. The room contained engineering equipment, lab benches, tables and tools. The liquid contaminant sampling task occurred in a virtual reality laboratory that has an open space with two large tables on which the task area was focused.

### V.1.3 Apparatus

During the H-H condition, a second experimenter played the role of the first responder and human teammate. A script dictated verbal interactions between the human participant and human experimenter. The experimenter carried a tablet computer throughout the evaluation, which provided the evaluation script and simulated additional equipment and sensors for the task. The same male experimenter was partnered with all participants.

The Pioneer 3-DX robot's navigation was controlled by an experimenter using line of sight teleoperation and a web-cam streaming video to the experimenter's laptop, unbeknownst to the participants. Participants were instructed that the robot moved and spoke autonomously. The robot spoke using a digital voice through speakers mounted on its platform. Participants donned a wireless microphone headset to amplify their voices when communicating with the robot. The same experimenter that controlled the robot heard participant questions and responses, and used this knowledge to advance the robot's pre-programmed speech script. The robot's speech script was identical to the human experimenter's verbal script in the H-H condition. When participants asked questions, the experimenter chose from a set of pre-programmed responses, repeated the robot's last statement, or wrote in a custom response.

The four tasks were limited to fifteen minutes. Breaks of approximately five minutes occurred between tasks in order to allow for equipment setup, robot battery changes, and participant restroom breaks. Five minutes is a common time period used as a break between tasks using physiological measures (Lei and Roetting, 2011; Tomaka et al., 1997). Reimer et al. (2009) found that within 5 minutes of the highest workload task, heart rate dropped significantly to within 0.6 beats per minute of baseline values.

### V.1.4 Participants

Sixteen participants completed the evaluation. Four male and four female participants were paired with the human partner first and the robot partner second (human partner first group) and five male and three female participants were paired with the robot first and the human partner second (robot partner first group).

The mean age of the human partner first group was 22.63 (St. Dev. = 6.16) with a range from 18 to 39, and Group 2's mean age was 21.89 (St. Dev. = 5.04) with a range from 18 to 34. Participants rated their mean

search and rescue experience and experience with robots on a Likert scale from 1 (little or no experience) to 9 (very experienced). The human partner first group rated their search and rescue experience as a median of 1 with a range from 1 to 9, while the robot partner first group rated their search and rescue experience with a median of 4.5 with a range from 1 to 7. The human partner first group rated their robotics experience as a median of 1 with a range from 1 to 9, while the robot partner first group rated their robotics experience as a median of 3 with a range from 1 to 8. There was no significant difference between the two groups for either rating.

Participants also rated their agreement with statements regarding feeling inhibited with strangers, preference for working alone, and tendency for angering others. These social behavior questions were measured in order to form a basis for participants' teamwork with his or her partner and were based on the most successful questions from a verbal sociability questionnaire (Asendorpf and Meier, 1993). Agreement was rated on a Likert scale from 1 (totally disagree) to 9 (totally agree). The human partner first group rated their inhibition with strangers with a median of 6 (range from 3 to 6), preference to work alone with a median of 6 (range from 4 to 8), and tendency to make others angry with a median of 2.5 (range of 1 to 5). The robot partner first group rated their inhibition with strangers with a median of 5 (range from 3 to 6), preference to work alone with a median of 5 (range from 2 to 8), and tendency to make others angry with a median of 2 (range of 1 to 5). There were no significant differences for any of these statements between the groups.

The human partner first group slept a median of 6.5 hours the night before their first session, with a range of 3 to 8 hours. The robot partner first group slept a median of 8 hours the night before their first session, with a range of 7 to 10 hours. The robot partner first group slept significantly more hours the night before the first session, as demonstrated by a Wilcox test, $U = 8.5$, $p = 0.012$. The human partner first group was awake for a median of 7.125 hours (range of 1 to 12.5 hours) before the first session, while the robot partner first group was awake for a median of 7.625 hours (range of 0.25 to 12.45 hours). There was no significant difference between the two groups in the number of hours awake prior to the first session.

The human partner first group slept a median of 6 hours the night before their second session, with a range of 3 to 9 hours. The robot partner first group slept a median of 7.5 hours the night before their second session, with a range of 3 to 9 hours. There was no significant difference between the two groups for number of hours slept prior to the second session. The human partner first group was awake for a median of 8 hours (range of 4.75 to 13 hours) before the first session, while the robot partner first group was awake for a median of 8.125 hours (range of 1.83 to 9.75 hours). There was no significant difference between the two groups for number of hours awake prior to the second session.

### V.1.5 Metrics

The objective metrics included physiological responses (i.e., heart rate, Chapter IV.1.1; respiration rate, Chapter IV.1.2; heart rate variability, Chapter IV.1.3; vector magnitude, Chapter IV.1.6); steps taken; timing measures (i.e., subtask time, Chapter IV.2.5; primary task response time, Chapter IV.2.3; primary task reaction time, Chapter IV.2.1); primary task failure rate, Chapter IV.2.6; secondary task failure rate, Chapter IV.1.11; a memory recall task, Chapter IV.1.8; and the calculation of task density, Chapter IV.1.9.

The participants wore the Bioharness chest strap monitor in order to collect physiological responses (i.e., heart rate, respiration rate, heart rate variability, vector magnitude). Additionally, the Scosche Rhythm+, a commercially available sensor, was worn to collect heart rate information and distance traveled. A Fitbit Zip recorded the number of steps taken by each participant during the evaluation.

The timing measures (i.e., subtask time, primary task response time, primary task reaction time) were collected via video coding. The video was recorded from a Looxcie camera worn to capture the participant's point-of-view.

Primary task failure rate was calculated based on participant performance for each assigned task. The primary tasks completed were analyzed in order to determine the components completed. The four evaluation tasks included an effort, attention, detail-tracking, and assessment component (see Chapter IV). The primary task required participants to find and report suspicious items in the hallway during the item search task (Attention component), take photographs of said suspicious items when requested during the item search task (Effort component), follow strict protocols during the solid and liquid contaminant sampling tasks without making mistakes (Assessment component), and examine photographs for suspicious details for reporting back to the incident command in the photo search task (Detail-tracking task). An additional component determined how often participants failed to complete the assigned tasks in the given time frame (Achievement component). Failures were determined when the participant was unable to correctly perform an assigned step. The failure rate was calculated as the number of failures divided by the total number of assigned steps for the task. Chapter V.1.6 provides information regarding the individual steps for each evaluation task.

A failure rate was also computed for the assigned secondary task. Participants were assigned to monitor a walkie-talkie for incoming messages from the Incident Commander for their team, Team 10. An example of a message was, "Incident Command to Team 10: there is a suspicious person running south on Anderson Road." The participant was responsible for recognizing that the message was directed at Team 10, and repeating it to his or her partner at the time the communication arrived. A failure was identified when the participant did not report the message to his or her partner, and a partial failure was identified when the participant was able to respond to the Team 10 message, but relayed the message content incorrectly. During low workload

tasks, eight total messages were relayed over the walkie-talkie; two messages were for Team 10. During high workload tasks, 24 total messages were relayed over the walkie-talkie, six of which were directed to Team 10.

The memory recall task required participants to write down every Team 10 message that he or she remembered from all four tasks during the evaluation period. The participants wrote every Team 10 message recalled following the completion of all four tasks during each session. The memory recall score was calculated by summing the number of correctly remembered Team 10 messages. A correctly remembered message received one point and a partially remembered phrase received one half point.

Task density, as described in Chapter IV.1.9, was calculated by computing the number of subtasks completed for a specific task (e.g., number of solid contaminant samples taken during the fifteen minute task period) divided by the active subtask time for that task (e.g., time taken to individually complete each sampling action, and does not include the idle time between sample collections). Task density was somewhat controlled via the experimental design, because the number of initiated tasks (e.g., samples to take, secondary tasks requiring responses) were limited and paced by the experimenter. Participants varied in efficiency and speed; thus, the density of the tasks completed during the fifteen-minute task time span was not identical for each participant. The task density was measured by analyzing the portion of fifteen minute task time spent actively accomplishing the primary task. The task density ratio was computed by dividing the number of primary tasks completed during the task by the total active time (i.e., summed subtask time for the specific task). The time waiting for the next sample (i.e., idle time) was not considered. Higher task density indicates higher workload.

The subjective metrics included the in situ subjective workload ratings, the NASA Task Load Index workload questionnaire (see Chapter IV.1.13), and a post-trial questionnaire. The in situ subjective workload ratings were elicited following each of the four tasks (the rating measurements are described in Chapter IV.1.12). The post-trial questionnaire focused on the participant's trust in his or her partner, confidence in his or her ability, and the team interaction. The statements were identical to those posed in the Collaborative evaluation (see Chapter III), and participants rated agreement on a Likert scale from 1 (strongly disagree) to 9 (strongly agree). One additional statement was added, "I believe I could have completed this task on my own as well or better than I did with my teammate." A final preference questionnaire was added at the completion of the second session, in order to assess which partner each participant preferred. Participants were asked to rate agreement with the following statement, "I preferred working with one partner over the other." Following the rating, participants identified the preferred partner (i.e., human partner, robot partner, no clear preference). Participants were also free to provide open-ended comments and suggestions.

### V.1.6 Procedure

Participants completed the two evaluation sessions on different days. The distribution of timing between participants' sessions were, on average, 13.25 days (St. Dev. = 10.77). A subset of second sessions were delayed due to participant cancellations and scheduling restrictions.

Upon arrival for the first session, participants completed a consent form and demographic questionnaire. The participants received an evaluation task briefing and were shown a 3 minute 40 second training video. Participants donned the Bioharness heart rate monitor, the Looxcie camera, the Shure microphone headset, a walkie talkie with ear piece, the Fitbit activity monitor, the Scosche Rhythm+ heart rate monitor, and a reflective vest to identify them as evaluation participants to non-participants in the vicinity. Participants were introduced to the responder partner, either the human or robot partner and began the first task, the photo search task.

#### Photo Search Task

The photo search task required identifying suspicious items in photographs of an area taken by a surveillance team. The participant was told that a team of robots previously entered the building and photographed rooms and areas that may contain victims, hazardous chemicals, suspicious items, or nothing to investigate. The participant was presented with an Google Nexus 7 tablet computer running the Android mobile operating system. The tablet was used in order to view, search, and edit the photographs. The incident commander was responsible for sending folders of photographs to the participant's tablet, and notes from the investigation team to the responder, throughout the task period.

The environment was a small office space that included two adjacent tables and two chairs. During the H-H team tasks, the human partner sat at the table on the right, while the participant sat at the table to the left. During the H-R team tasks, the robot drove to a spot near the table on the right.

Participants held the tablet during the task, and were also able to rest his or her arms on the table, or set the tablet down as he or she preferred. There were no constraints in place on how the participant held the tablet, and no stand was used to prop the tablet on the table. The participant was able to swivel the chair to face the partner, but no movement outside of controlling the tablet, was required.

The participant's tablet computer received the transmitted folders of photographs by a document sharing service, Box. Periodically, new folders containing three photographs each appeared for the participant and his or her partner to investigate. Once the participant reviewed the photograph and identified something suspicious, it was his or her job to edit the photograph in the photo-editing application, Aviary. The participant was responsible for circling the suspicious item or adding a note to describe why the item must be investigated by a follow-up team. An instruction sheet provided with information regarding how to use Box and perform

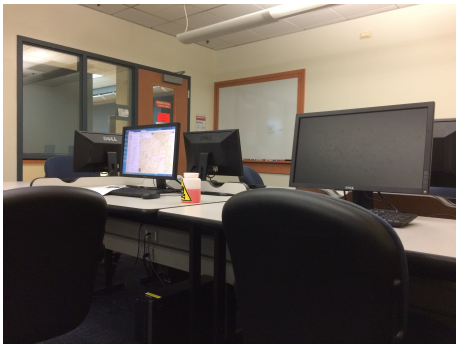


Figure V.1: An example photograph examined, searched, and edited during the photo search task.

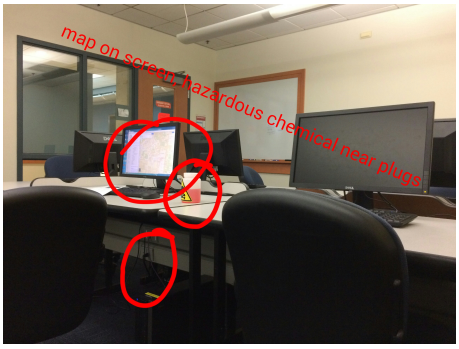



Figure V.2: Photograph edited by a participant during the photo search task.

the photo-editing steps in the Aviary application.

The participant performed a short training session using test photographs prior to the beginning of the fifteen minute task period and used the instruction sheet to learn how to use the tablet, Box application, and photo-editing system without time pressure. The participant was able to take as long as he or she desired to train on the three test photos. Participants were also permitted to ask questions during the task.

Once the fifteen minute task time began, two sets of three photographs were searched in the low workload condition; four sets of three photographs were searched in the high workload condition. The three photographs in each folder showed different angled perspectives of a room. The rooms included areas of the building in which the evaluation took place, including study areas, classrooms, a computer laboratory, and the two laboratories in which evaluation tasks were completed. The folders of photographs were presented in the same order for all participants within a workload condition, but different and comparable sets of photographs were used for the two evaluation sessions for each participant. An example photograph that participants examined is provided in Figure V.1, and an example of the photograph post-participant annotations is provided in Figure V.2.

This task required a low level of physical activity, because the participant wore no protective gear and was stationary and seated in a chair. The team collaborated on deciding whether a follow-up team needed to be sent to that room.

A remote evaluator uploaded the sets of photographs to the folder visible by the participant at predefined times during the fifteen minute task. During both workload conditions, the first photo set was uploaded immediately before the task began. During the low workload condition, the second set of photographs was uploaded seven minutes and thirty seconds into the task, while in the high workload condition, the four folders arrived with three minutes and forty-five seconds in between. Figure V.3 presents the timing for the photo search task, divided by workload condition.

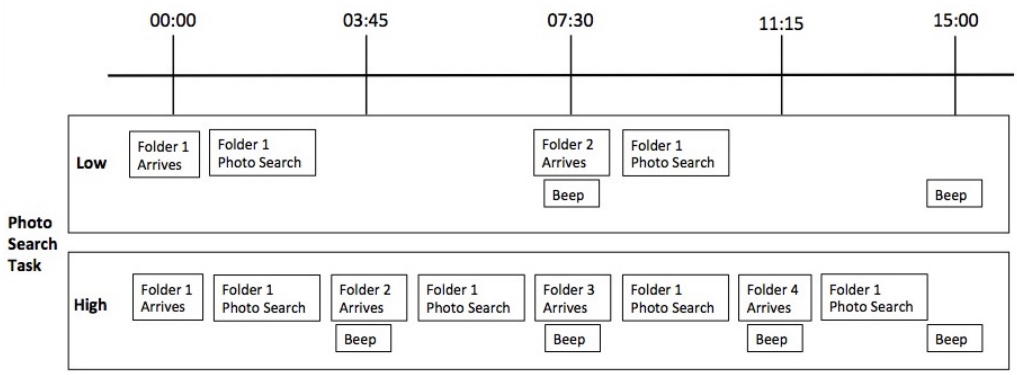



Figure V.3: The time line of the photo search task, by workload condition.

Some participants took longer to finish examining the folders than others; if a particular participant was not finished with Folder 1 by the time Folder 2 arrived, they simply opened Folder 2 when they finished Folder 1. Similarly, if there was extra time after the team finished evaluating Folder 1 and before Folder 2 arrived, for example, the responder explained that the incident commander uploads folders as soon as photographs become available. An audible beep tone was used (the same tone in all tasks) to indicate to the team when a new folder arrived. Teams completed the photo search task in a varying length of time; thus the beeps indicated time pressure to the teams that were completing the task more slowly. The teams who finished before the next folder of photographs became available were able to know precisely when the new folder was sent by incident command and were able to start the investigation immediately.

**Item Search Task**

The second task, the item search task, required conducting an exhaustive search of an environment for potentially hazardous items, gathering environmental context, and gathering air samples, with the goal of identifying as many hazardous items, while searching as much of the assigned area within the time limit. The participant's role was to physically locate the items and photograph the item. The participants wore equipment to simulate personal protective equipment, including safety gloves, goggles, a dust mask, and a weighted backpack, weighing 10 pounds. The dust mask and backpack represented a first responder's rebreather apparatus recirculating clean breathing air in a contaminated environment (see Figure V.4). A participant is seen in Figure V.5 wearing the weighted backpack and dust mask, as well as the goggles and gloves, while performing the item search task.

The environment for the item search task was a hallway of an academic building. Faculty, graduate student, and administrative offices, as well as engineering laboratory classrooms were present in this area. A fire extinguisher, several bulletin boards, and three trash cans were also present.

Figure V.4: A first responder wearing a rebreather apparatus (Domestic Preparedness, 2007).

Figure V.5: Participant in H-R team performing the item search task, while wearing weighted backpack and mask, simulating a rebreather apparatus.

This task was a higher-level physical activity task, because it involved walking around a hallway and the participant wore the weighted backpack. The participants were responsible for searching areas above the robot's sensors' field of view. The human or robot partner's role was to scan for hazards near the ground, collect air samples, and alert the participant if any hazards or high air samples were detected. The team collaborated by discussing whether items detected were suspicious.

Four items were investigated in the low workload condition, while eight items were investigated in the high workload condition. Each of the items investigated in the first session were identical, regardless of the participant's partner. A similar set of items, placed in different locations in the same environment, was used for all participants during the second session, independent of the assigned partner. The first session items are depicted in Figure V.6. The star indicates that the item was only used in high workload tasks, while the items without the star were used for both the low and high conditions. The items used for the second session are depicted in Figure V.7 with the same use of the star indication as in Figure V.6.

During the investigation, the participant and his or her partner traveled through the environment. Either teammate was able to determine whether the team needed to stop to investigate an item, though the responder teammate only did so when the participant missed an item that required investigation. The time line for the item search task is presented in Figure V.8.

During the task, the responder was responsible for pausing the search in order to send current search information to incident command and waiting for a message in response indicating approval to continue with the search. Incident command's messages were indicated with an audible beep at pre-defined times during the task (i.e., at 3:45, 7:30, 11:15, and 15:00, see Figure V.8). As teams completed the search and investigation of

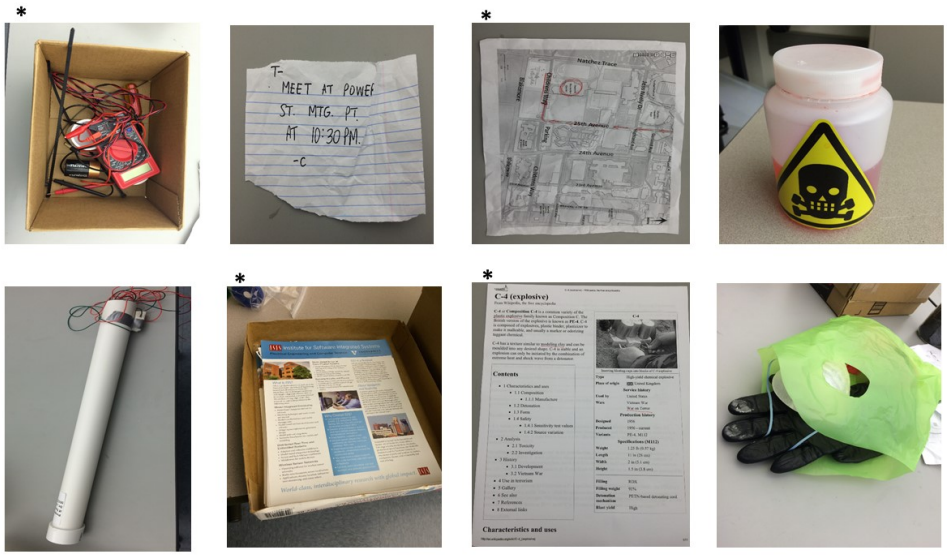



Figure V.6: The eight items used for the first session of the item search task. Starred items (*) were not used in the low workload condition. From top left, clockwise: cardboard box filled with wires and suspicious material, cryptic note, suspiciously marked map of Vanderbilt campus, hazardous liquid, bag with gloves and dust mask, papers regarding C4 explosive use, box of advertisements (not suspicious), pipe bomb.

items in varying time lengths, the beep created time pressure for the teams completing the task more slowly. The teams who finished before the next message from incident command arrived were able to know precisely when the new notification was sent by incident command.

**Solid Contaminant Sampling Task**

The solid contaminant sampling task was the third task and required collecting samples from potentially hazardous solids in the room. The participant donned safety gloves and goggles, and this task had a lower physical activity level, because the items were in close proximity and the participant did not wear the weighted backpack. The participant's role was to collect samples from the solids stored in various containers using a sterile collection kit and following guidelines provided by the human or robot responder partner. These guidelines included detailed procedures that required strict compliance for maintaining safe and sterile sampling procedures. These evaluation steps were based on published government standards for the bulk sample collection of visible powders and suspected biological agents (ASTM International, 2010b,a). The responder partner indicated which solid to sample based on a message from the incident commander that arrived via the tablet computer.

The participant also provided information and context about the hazard. The human or robot responder partner's role was to provide hazard collection guidelines and also request information regarding the context

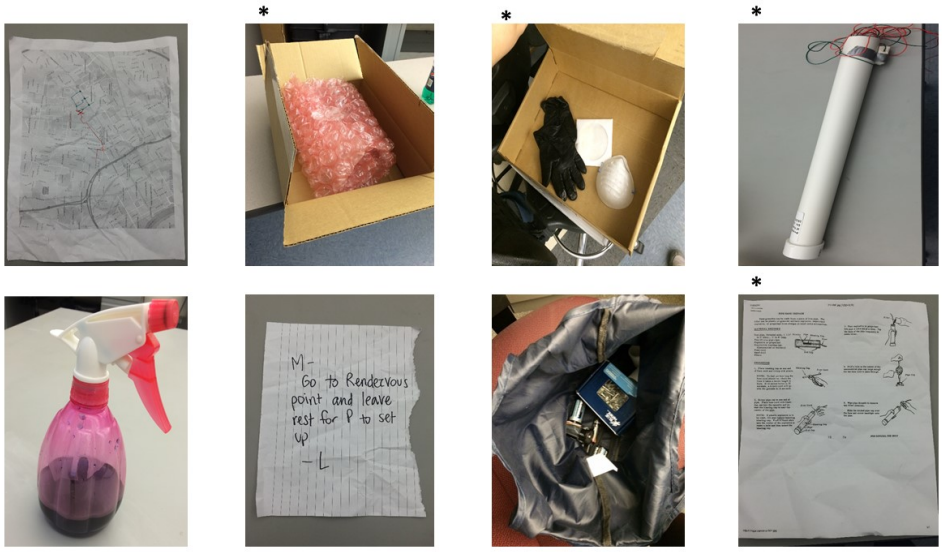



Figure V.7: The eight items used for the second session of the item search task. Starred items (*) were not used in the low workload condition. From top left, clockwise: suspiciously marked map of Nashville, bubble wrap (not suspicious), box with gloves suspicious envelope with white powder, pipe bomb, papers instructing the fabrication of pipe bombs, bag filled with batteries and nails, cryptic note, suspicious liquid in spray bottle.

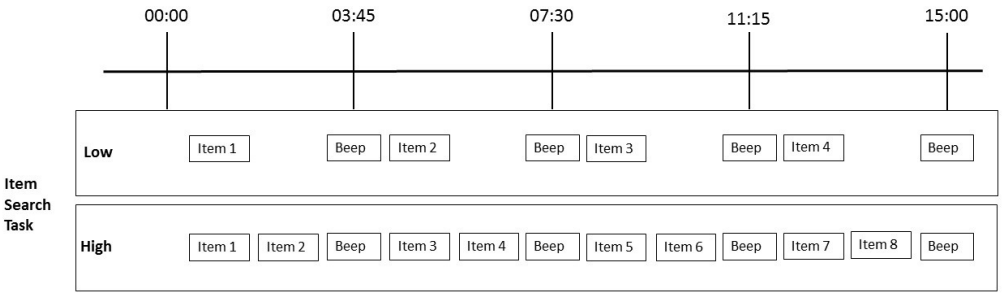



Figure V.8: The time line of the item search task, by workload condition.

Figure V.9: Participant and human responder partner performing the solid contaminant sampling task.

Figure V.10: Participant and robot responder partner performing the solid contaminant sampling task.

and details of each potential hazard. Figures V.9 and V.10 depict a participant completing task three with the human partner and robot partner, respectively.

The team entered the room with hazards visible on a table. The hazards were containers (e.g., clear plastic storage container, glass jar, film canister) filled with unknown, colored solids. The solids were composed of colored sand and baby powder. Participants used sampling kits, transported in a large gardening wagon, in order to collect small samples from a subset of the present hazards in the order requested by the incident commander. Each kit contained two sandwich-sized zip-lock plastic bags, one four-ounce glass sample jar, one stainless steel scoopula, and one alcohol wipe. The kit was placed in a gallon-sized storage bag and wrapped in a diaper to maintain sterility and protect the kit from breakage. The wagon also contained mailing labels to seal the bags and one permanent marker for writing the time on the seal. The wagon was stationed at the entrance to the room in the same place during each evaluation session. The participants were free to move the wagon.

Two samples were assigned in the low workload condition, and four samples were assigned in the high workload condition. Figure V.11 provides a guide to all of the steps completed when collecting each of solid contaminant sample. The messages to the responder regarding the samples to collect were marked by an audible beep, indicating that an additional sample was required. Teams completed solid contaminant sampling with varying time durations. The beep added time pressure to teams completing the task more slowly, in a similar manner to the other tasks. The teams who finished before the next message arrived were able to know precisely when the new sample request was sent by incident command. The time line of the solid contaminant sampling task is available in Figure V.12.

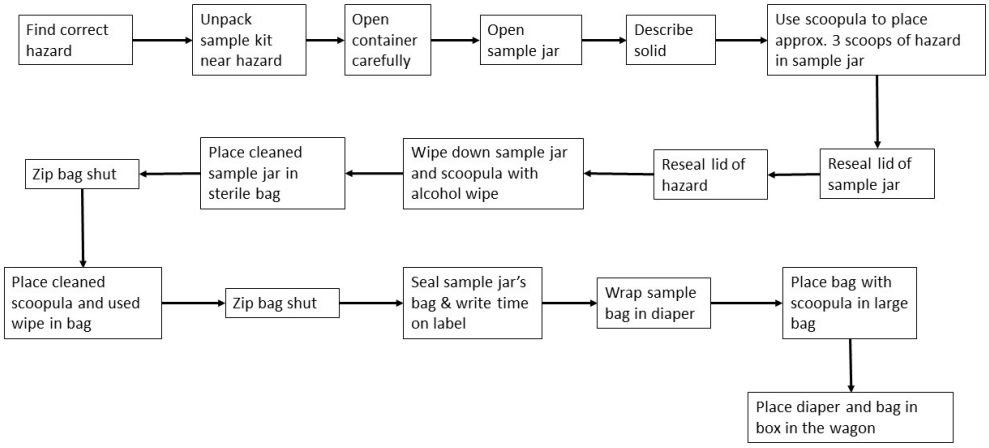



Figure V.11: The steps completed for each solid contaminant sample collected in the solid contaminant sampling task.

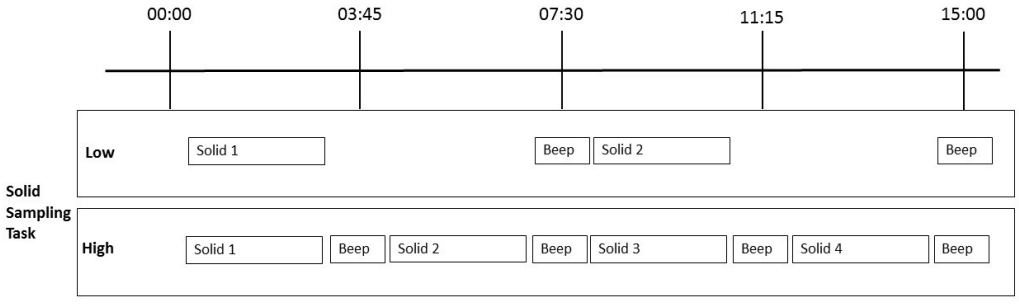



Figure V.12: The time line of the solid contaminant sampling task, by workload condition.

**Liquid Contaminant Sampling Task**

The liquid contaminant sampling task required collecting hazardous liquid samples. Hazardous material sample collection procedures dictate conducting the collection from the least hazardous materials to the most hazardous materials (ASTM International, 2010b,a). This evaluation simulates training; thus, the participants moved onto a task with a more difficult sampling technique and sampled liquids, while wearing safety gloves, goggles, a dust mask, and a weighted backpack. The additional equipment simulated the cumbersome personal protective gear that emergency responders wear (see Figures V.4 and V.5). This task had a higher level of physical activity because it required walking around a larger laboratory space between the samples, and wearing the backpack. The gloves and mask added extra physical workload by increasing the task difficulty, such as opening plastic bags and taking the samples.

The sampling steps were very similar to the solid contaminant sampling task. A highly structured protocol, based on government requirements, was used to ensure sterile and safe collection of toxic liquid hazards. The responder partner provided guidelines and gathered information from the participant, who performed the sampling.

Participants entered the room with two tables each with different containers (e.g., sports water bottle, glass jar, Pepsi bottle) containing colored liquids. The liquids were water dyed with various hues of food coloring. There were nine containers set out for all tasks in the same configuration for each evaluation during both sessions. The first session assigned a different subset of the nine containers, than used during the second session.

The kits contained two sandwich-sized zip-lock plastic bags, one four-ounce glass sample jar, one plastic pipette, one drop cloth cut from plastic sheeting (approximately 2 feet by 1 foot), and one alcohol wipe. The drop cloth was placed under the sampling area to catch potential spillage. The kits were stored in a gallon-sized plastic bag and wrapped in a diaper to maintain sterility and protection from breakage. The kits were also stored in the gardening wagon stationed near the room's entrance for each evaluation session. The wagon also held the mailing labels for sealing and a marker for labeling the sample. Participants were able to move the wagon throughout the room at their discretion.

The specific steps required for each liquid contaminant sample collection are presented in Figure V.13. The participant's role was similar to that in the solid contaminant sampling task; the human or robot partner guided the participant on how the liquid sampling technique differs from the solid contaminant sampling, and how to collect each liquid contaminant sample. Two samples were assigned in the low workload condition, while four samples were assigned in the high workload condition. The teams completed liquid contaminant sampling in varying time lengths. Beeps indicated when a new sample request was sent from incident com-

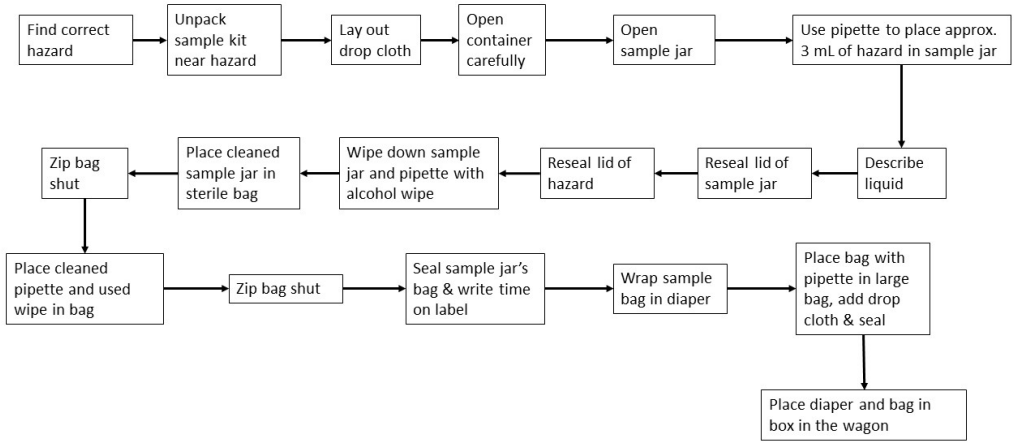



Figure V.13: The steps completed for each liquid contaminant sample collected in the liquid contaminant sampling task.

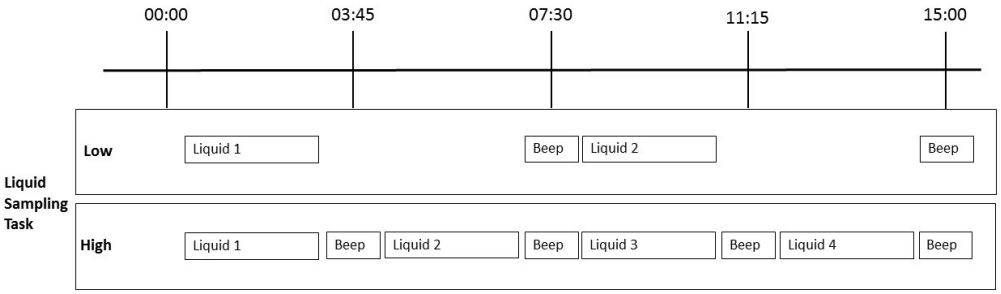



Figure V.14: The time line of the liquid contaminant sampling task, by workload condition.

mand and added time pressure to teams completing the task more slowly. The teams who finished the sample before the beeps sounded were able to know precisely when the new sample request was sent by incident command. The task time line is presented in Figure V.14.

**After All Four Tasks**

Following the completion of each task, the participant's responder partner asked the participant to provide in situ workload ratings for each demand channel (see Chapter V.1.5). Upon completing all four tasks from the first session, the participant completed the memory recall task, the NASA Task Load Index, and a post-trial questionnaire.

The participants returned for the second session and completed a questionnaire regarding their prior

night's sleep, donned the equipment, and completed the evaluation in the identical task order as the first session, independent of responder partnering - human or robot. Following the evaluation, the participants completed the memory recall task, the NASA-Task Load Index, and the post-trial questionnaire, in addition to a final partner preference questionnaire.

## V.2 Modeling Description

### V.2.1 Time-Structured Scenario Model

The Time-Structured scenario model was developed in IMPRINT Pro prior to conducting the evaluation. Each of the four tasks (i.e., photo search, item search, solid sampling, and liquid contaminant sampling tasks) were modeled with each partner type (i.e., H-H and H-R teams) and each workload condition (i.e., low and high workload). The modeled tasks were limited to fifteen minutes in order to match the evaluation's timing constraint.

The subtasks were modeled individually as IMPRINT Pro functions. The task timings were determined by estimating task times via preliminary pilot tests with experimenters as participants and using IMPRINT Pro's built-in micromodels of human behavior for tasks, such as speech. The secondary tasks were added via IMPRINT Pro's scheduled task feature. Workload for each individual task was determined using IMPRINT Pro's workload assignment guidelines. Workload channel conflicts were included in order to increase workload when more than one simultaneous task requires conflicting channels of demand (e.g., writing, while speaking a different sentence; listening for a message with high volumes of background noise). The workload channel conflicts were calculated using the IMPRINT Pro defaults.

The model results were collected from ten model runs, in order to collect a mean and standard deviation of the workload and subtask times. The models were analyzed by comparing the subtask time and workload levels between workload conditions and team partner types. Subtask time was represented by the length of time taken to perform the individual task (e.g., performing the solid contaminant sampling steps for one sample).

#### V.2.1.1 Time-Structured Model Results

The mean and standard deviations are provided in Table V.2. Kruskal-Wallis tests assessed the main effect of workload level and partner on the modeled results. The results indicated that the high workload models had significantly higher workload than the low workload models ($\chi^2(1) = 100.93, p < 0.001$), and that there was no main effect of partner on modeled workload. A Kruskal-Wallis test indicated that subtask time was longer in the H-R model, $\chi^2(1) = 4.34, p = 0.037$, and subtask time was significantly longer for the low workload tasks, $\chi^2(1) = 11.10, p < 0.001$.

Table V.2: Time-Structured Evaluation Model Results

| Metric | Analysis | | H-H | | H-R | |
|---|---|---|---|---|---|---|
| | | | Mean | St. Dev. | Mean | St. Dev. |
| **Modeled Workload** | **Low** | Photo Search | 15.68 | 0.12 | 16.34 | 0.13 |
| | | Item Search | 15.09 | 0.18 | 15.22 | 0.17 |
| | | Solid Contaminant Sampling | 16.05 | 0.34 | 16.72 | 0.36 |
| | | Liquid Contaminant Sampling | 24.59 | 0.93 | 25.67 | 0.82 |
| | **High** | Photo Search | 25.42 | 1.33 | 26.71 | 1.42 |
| | | Solid Search | 29.97 | 0.80 | 30.96 | 0.78 |
| | | Solid Contaminant Sampling | 33.58 | 1.35 | 21.62 | 0.90 |
| | | Liquid Contaminant Sampling | 46.01 | 0.93 | 30.58 | 0.83 |
| **Modeled Subtask Time** | **Low** | Photo Search | 284.33 | 19.68 | 299.50 | 13.50 |
| | | Item Search | 66.67 | 17.71 | 70.75 | 20.61 |
| | | Solid Contaminant Sampling | 288.50 | 14.01 | 307.41 | 14.50 |
| | | Liquid Contaminant Sampling | 290.17 | 12.32 | 310.33 | 27.63 |
| | **High** | Photo Search | 221.33 | 41.48 | 232.02 | 45.09 |
| | | Item Search | 81.04 | 17.71 | 89.34 | 18.97 |
| | | Solid Contaminant Sampling | 255.08 | 30.70 | 277.27 | 35.55 |
| | | Liquid Contaminant Sampling | 251.83 | 29.94 | 275.54 | 35.74 |

The modeled results indicate that in the first two tasks, H-R workload tends to be higher than H-H workload. This result is due to the fact that modeled H-R teams have significantly longer subtask time and have less low workload down time than H-H participants, which increases the overall task workload. During the sampling tasks, the results indicate a trend in lower workload for the H-R teams, because the modeled H-R teams took longer to complete the sampling steps. The models predict that H-R teams will be unable to complete all four samples in the high workload tasks, unlike the H-H teams.

### V.3 Results

The Time-Structured evaluation results are analyzed and presented in the same order as the metrics presented in Chapter IV. A subset of metrics were collected via video and audio coding (i.e., timing metrics, task failure rates). Videos were collected from the participant-worn point-of-view cameras during each task, but due to technology failure, 45 of the 128 tasks were not fully recorded on video; thus audio coding was used when necessary, recorded from the headset microphone worn by the participant. Video coding was performed on the 83 full videos. The participant, the partner, and the secondary task prompts can be heard in both the audio and video recordings. A primary coder collected data from all 128 task recordings, and a second coder collected data for only the solid contaminant sampling task, as it is common practice for a second coder to verify the primary coder's data via a subset of the video and audio files (e.g., Friedman et al., 2006; Ford et al., 2010). The inter-coder reliability had a Cohen's kappa value of 0.74.

A subset of the results were analyzed using Kruskal-Wallis tests; one assumption of the Kruskal-Wallis

test requires independence of samples. This assumption was not met due to the repeated-measures design of the Time-Structured evaluation; however, the Kruskal-Wallis test was used to assess the main effects of independent variables. Much of the data was unable to be transformed to a normal distribution for use of the ANOVA test due to ordinal measure. The Kruskal-Wallis test, a statistical test that assessed main effects without assuming normality, was chosen.

### V.3.1 Physiological Workload Metrics

The physiological workload metrics include heart rate, respiration rate, heart rate variability, postural load, variance in posture and vector magnitude. The latter three metrics measured physical workload. Each physiological metric was analyzed by assigned workload level (i.e., low, high), partner (i.e., human, robot), and session (i.e., first, second). The descriptive statistics for each test are available in Table V.3. The distributions were not normal; thus, non-parametric statistical analysis was used. Each physiological workload measure was also correlated to the in situ subjective workload ratings (see Chapter V.3.3.1).

#### V.3.1.1 Heart Rate

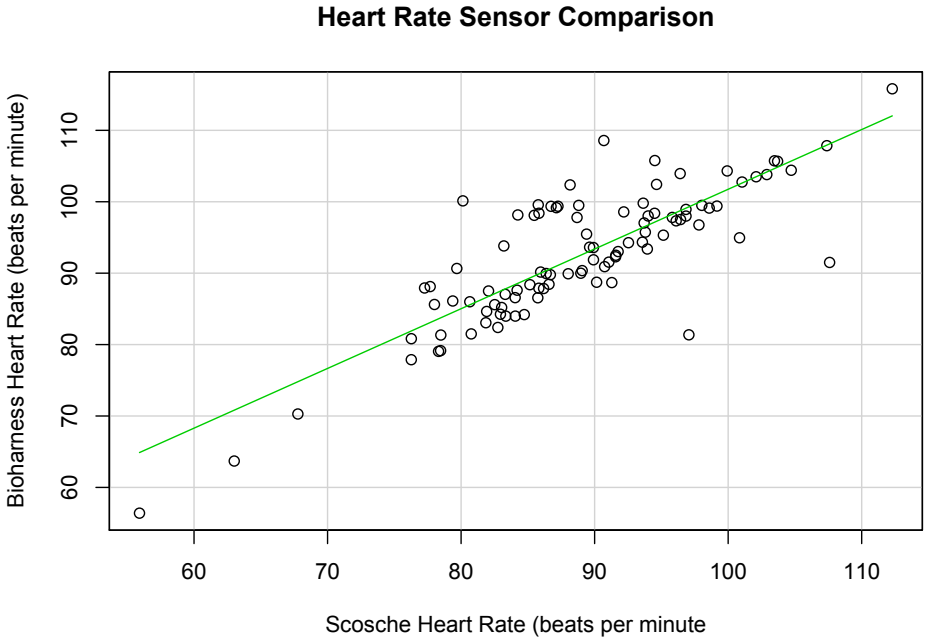
Heart rate was analyzed in the Time-Structured evaluation to represent workload and was recorded using the Bioharness chest strap monitor and the Scosche Rhythm+ athletic activity monitor. Both the Bioharness and Scosche heart rate data is presented in Table V.3 and was analyzed by partner, workload level, and session. There were no significant main effects of partner, session, or workload level on heart rate measured by either sensor.

The use of the Scosche fitness activity monitor was evaluated by performing a Pearson's correlation between the Bioharness-collected heart rate data and the Scosche arm band heart rate data. Seventeen data points were removed, because the difference between the Bioharness and Scosche was more than 20 beats per minute. There was a statistically significant correlation ($r(114) = 0.25, p = 0.008$) and a stronger correlation after the removal of the seventeen outliers, $r(95) = 0.82, p < 0.001$. Figure V.15 displays the heart rate collected for each participant from both monitors, after the removal of the seventeen outlier data points. Each point in the graph represents the data from both sensors for one participant during one task (all four tasks' data is included). The green line indicates the positive correlation between the Scosche sensor values and the Bioharness sensor values. The figure indicates that the Scosche values tended to be lower, as evidenced by the means in Table V.3. A Mann-Whitney one-sided test indicated that the Scosche values were significantly lower, $U = 3518, p = 0.001$. The heart rate data was not significantly correlated to the in situ workload ratings.

Table V.3: Descriptive statistics in the Time-Structured evaluation for each physiological measure of workload by workload level, partner, and session.

| Metric | Analysis | | Mean | St. Dev. |
|---|---|---|---|---|
| **Heart Rate (beats per minute) - Bioharness** | **By Partner** | **H-H** | 94.73 | 10.02 |
| | | **H-R** | 96.85 | 14.79 |
| | **By Workload Level** | **Low** | 94.66 | 10.01 |
| | | **High** | 96.77 | 14.54 |
| | **By Session** | **First** | 97.17 | 13.65 |
| | | **Second** | 94.45 | 11.49 |
| **Heart Rate (beats per minute) - Scosche** | **By Partner** | **H-H** | 89.04 | 9.11 |
| | | **H-R** | 88.74 | 9.31 |
| | **By Workload Level** | **Low** | 89.23 | 7.55 |
| | | **High** | 88.57 | 10.61 |
| | **By Session** | **First** | 89.32 | 9.29 |
| | | **Second** | 88.52 | 9.10 |
| **Respiration Rate (breaths per minute)** | **By Partner** | **H-H** | 19.44 | 2.95 |
| | | **H-R** | 19.56 | 2.90 |
| | **By Workload Level** | **Low** | 19.09 | 3.02 |
| | | **High** | 19.90 | 2.77 |
| | **By Session** | **First** | 19.04 | 2.65 |
| | | **Second** | 19.95 | 3.11 |
| **Heart Rate Variability ($ms^2$)** | **By Partner** | **H-H** | 140764.5 | 173223.2 |
| | | **H-R** | 119547.1 | 189648.4 |
| | **By Workload Level** | **Low** | 121416.5 | 172759.9 |
| | | **High** | 138924.8 | 190001.3 |
| | **By Session** | **First** | 147607.3 | 193901.3 |
| | | **Second** | 113149.8 | 167421.2 |
| **Variance in Posture ($degrees^2$)** | **By Partner** | **H-H** | 258.69 | 246.31 |
| | | **H-R** | 277.76 | 281.38 |
| | **By Workload Level** | **Low** | 218.27 | 198.80 |
| | | **High** | 317.23 | 308.11 |
| | **By Session** | **First** | 284.19 | 263.34 |
| | | **Second** | 252.36 | 264.60 |
| **Vector Magnitude ($VMUx10^3$)** | **By Partner** | **H-H** | 60.26 | 30.31 |
| | | **H-R** | 63.99 | 31.54 |
| | **By Workload Level** | **Low** | 56.62 | 26.98 |
| | | **High** | 67.51 | 33.59 |
| | **By Session** | **First** | 64.21 | 30.30 |
| | | **Second** | 60.04 | 31.49 |

Figure V.15: Comparison between the heart rate collected via Scosche Rhythm+ activity monitor and the Bioharness chest strap.




#### V.3.1.2 Respiration Rate

Respiration rate was collected via Bioharness chest strap and the descriptive statistics are available in Table V.3. There was no significant main effect of partner and nearly significant main effects of both session ($\chi^2(1) = 2.89, p = 0.089$) and workload level ($\chi^2(1) = 2.99, p = 0.084$). These results indicate a trend that participants had higher respiration rates during the second participation session, and higher respiration rate in the high workload level tasks. The Pearson's correlation between respiration rate and in situ workload ratings was not significant.

#### V.3.1.3 Heart Rate Variability

Low frequency heart rate variability was collected via Bioharness chest strap monitor and the descriptive statistics are presented in Table V.3. There were no significant main effects of partner, session, or workload level. There was a significant correlation between low frequency heart rate variability and in situ subjective workload ratings, $r(123) = 0.39, p < 0.001$. As workload increased, heart rate variability increased.

#### V.3.1.4 Postural Load

Postural load was not analyzed in the Time-Structured evaluation because crouching and bending down was not required during the tasks. Additionally, variance in posture was demonstrated in the Collaborative eval-

uation to represent physical movement in tasks where crouching and bending is not an essential duty of the task.

#### V.3.1.5 Variance in Posture

Variance in posture was collected using raw posture values recorded from the Bioharness chest strap monitor, and descriptive statistics are available in Table V.3. There were no significant main effects of partner, session, or workload level. There was a significant correlation between variance in posture and in situ subjective workload ratings, $r(123) = 0.18, p = 0.04$. An increase in physical movement was reflected in higher workload ratings.

#### V.3.1.6 Vector Magnitude

Vector magnitude is presented in vector magnitude units (VMU) multiplied by $10^3$, for easier viewing and comparison with the Guided and Collaborative evaluations. Descriptive statistics are presented in Table V.3. There were no significant main effects of partner or session. A Kruskal-Wallis test indicated that assigned high workload level tasks resulted in significantly higher vector magnitude, $\chi^2(1) = 7.06, p = 0.008$. There was also a significant Pearson's correlation between vector magnitude and in situ workload ratings, $r(123) = 0.24, p = 0.01$.

The assigned tasks in the Time-Structured evaluation involved physical activity for all tasks except for the photo search task. Vector magnitude was analyzed in a two-way Kruskal-Wallis test to examine the interaction effect of task and workload level in order to assess whether the photo search task elicited lower vector magnitude. Descriptive statistics are presented in Table V.4. The results indicated a significant main effect of task ($\chi^2(1) = 63.46, p < 0.001$) and a significant interaction effect of task and workload level ($\chi^2(1) = 61.26, p < 0.001$). Post-hoc pairwise Wilcoxon tests with a Bonferroni correction indicated that the low workload photo search task had significantly lower vector magnitude than the low workload item search task ($p < 0.001$), solid contaminant sampling task ($p < 0.001$), and liquid contaminant sampling task ($p < 0.001$). The high workload photo search task had lower vector magnitude than the high workload item search task ($p < 0.001$), solid contaminant sampling task ($p < 0.001$), and liquid contaminant sampling task ($p < 0.001$).

Each of the other physiological measures were tested for correlation to vector magnitude using Pearson's correlations, in order to evaluated whether physical activity impacted measurement results. Heart rate ($r(116) = 0.504, p < 0.001$), respiration rate ($r(116) = 0.278, p = 0.002$), and heart rate variability ($r(116) = 0.239, p = 0.009$) each significantly correlated to vector magnitude. This result indicates that as vector magnitude increases, each of these measures increases as well, whether or not the task mental workload increases.

Table V.4: Vector magnitude by task and workload level in the Time-Structured evaluation.

| Task | Workload Level | |
|---|---|---|
| | **Low** | **High** |
| Photo Search | 16.23 (11.24) | 16.69 (11.11) |
| Item Search | 69.76 (19.39) | 81.47 (14.68) |
| Solid Contaminant Sample | 72.47 (13.57) | 81.14 (16.36) |
| Liquid Contaminant Sample | 66.87 (15.66) | 91.60 (17.48) |

Additionally, posture variance significantly correlated to vector magnitude ($r(116) = 0.505, p < 0.001$). Both measures relate to physical workload; thus, this result is expected.

#### V.3.1.7 Regression Analysis

Regression Analysis was performed in order to determine whether the physiological measures of workload reflected changes rated subjectively (i.e., in situ workload ratings), or physical workload (i.e., vector magnitude). Heart rate, respiration rate, and heart rate variability correlated to vector magnitude. Heart rate variability correlated significantly to in situ workload ratings. A series of multiple linear regressions tested whether in situ workload ratings or vector magnitude impacted heart rate, respiration rate, and heart rate variability. The first test indicated that vector magnitude significantly predicted heart rate, ($R^2 = 0.24, F(2,115) = 19.72, p < 0.001; \beta = 0.51, p < 0.001$) and that in situ workload ratings did not have a significant effect. A second regression indicated that vector magnitude significantly predicted respiration rate ($R^2 = 0.05, F(2,122) = 4.33, p0.02; \beta = 0.24, p < 0.01$). The in situ workload ratings did not have a significant effect on respiration rate. The third regression analysis indicated that in situ workload ratings significantly predicted low frequency heart rate variability ($R^2 = 0.16, F(2,122) = 12.82, p < 0.001; \beta = 0.36, p < 0.001$). There was no significant predictive effect of vector magnitude on heart rate variability. These results indicate that heart rate and respiration rate are influenced by physical activity, while heart rate variability is a more stable metric of mental workload during mobile tasks.

### V.3.2 Other Objective Workload Metrics

The objective workload metrics include movement count (not analyzed), working memory, task density, speech rate (not analyzed), and secondary task failure rate. The distributions of the objective workload metrics were not normally distributed; thus non-parametric tests were used.

#### V.3.2.1 Movement Count

Movement count was not analyzed in the Time-Structured evaluation, as a specific movement of interest was not identified for the assigned tasks. Physical workload was analyzed using other metrics, including variance

in posture, and vector magnitude.

#### V.3.2.2 Working Memory

Working memory was recorded in total phrases written by each participant, as well as a score given to each participant for the quality of the phrases recalled. The median number of phrases recalled was 4.5 for participants in H-H teams (range = 2-9) as well as H-R teams (range = 1-9). The median number of recalled phrases was 5 in the first session (range = 2-9) and 4 (range = 1-9) in the second session. There was no main effect of partner or session on number of phrases recalled.

The median memory score given to participants in H-H teams was 3 (range = 1.5-7.5) and 2.5 (range = 1-7) was the median score given to H-R team participants. Evaluated by session, participants received a median score of 3 (range = 1.5-7.5) in the first session and a median of 2.75 (range = 1-7) in the second session. There were no significant main effects of partner or session on the working memory score.

A correlation was tested between the total number of secondary task messages for Team 10 heard during the evaluation and the number of phrases recalled, $r(30) = 0.37, p = 0.04$. As the number of messages heard during the evaluation increased, the number of phrases recalled increased. Additionally, the correlation between memory score and total number of Team 10 messages heard was nearly significant, $r(30) = 0.30, p = 0.09$. The increase in workload tended to improve participant memory recall.

#### V.3.2.3 Task Density

The presented analysis of task density considers the primary task only. The means and standard deviations for task density by partner, workload level, and session are available in Table V.5. Kruskal-Wallis tests indicated that participants experienced higher task density in their second experimental session ($\chi^2(1) = 4.911, p = 0.027$), and task density was significantly higher in high workload tasks ($\chi^2(1) = 6.392, p = 0.011$). Analysis indicated that H-H teams had nearly significantly higher task density than H-R teams, $\chi^2(1) = 3.001, p = 0.083$.

These results indicate that high workload increased participants' experienced task density. Teams performed faster and more smoothly (e.g., lower primary task failure rate) in the second session; thus the higher task density result in the second session is in line with other results. Additionally, the H-H teams generally moved slightly faster through the subtasks, task density was higher in H-H teams.

#### V.3.2.4 Speech Rate

Speech rate was not measured in the Time-Structured evaluation due to the constraints discussed in Chapter IV.

Table V.5: Median and range for in situ workload ratings in the Time-Structured evaluation by partner, workload level, and session.

| Metric | Analysis | | Mean | St. Dev. |
|---|---|---|---|---|
| Task Density | By Partner | H-H | 0.628 | 1.136 |
| | | H-R | 0.366 | 0.249 |
| | By Workload Level | Low | 0.380 | 0.318 |
| | | High | 0.613 | 1.122 |
| | By Session | First | 0.454 | 0.488 |
| | | Second | 0.539 | 1.071 |

Table V.6: Median and inter-quartile range for secondary task failure rate in the Time-Structured evaluation by partner, workload level, and session.

| Metric | Analysis | | Median | Range |
|---|---|---|---|---|
| Secondary Task Failure Rate | By Partner | H-H | 0.33 | 0.00-0.83 |
| | | H-R | 0.31 | 0.83-1.00 |
| | By Workload Level | Low | 0.17 | 0.00-1.00 |
| | | High | 0.39 | 0.00-1.00 |
| | By Session | First | 0.39 | 0.00-1.00 |
| | | Second | 0.28 | 0.00-0.83 |

#### V.3.2.5 Secondary Task Failure Rate

The secondary task failure rate was analyzed by partner, session, and workload level using Kruskal-Wallis tests (descriptive statistics available in Table V.6). There was no main effect of session or partner on secondary task failure rate. There was a significant main effect of workload level, indicating that a higher failure rate was present in high workload tasks, $\chi^2(1) = 12.353, p < 0.001$. Secondary task failure rate reflected the experimental workload manipulation.

### V.3.3 Subjective Workload Metrics

The subjective measures of workload include the in situ workload ratings and the NASA Task Load Index responses. The responses are not normally distributed for either questionnaire type; thus, non-parametric analysis is used.

#### V.3.3.1 In Situ Workload Ratings

The in situ workload ratings were analyzed by session, partner, and assigned workload level. The descriptive statistics are presented in Table V.7. There were no significant main effects of any of these three factors. The median total in situ workload rating in the photo searching task was 15 (range = 6-24), was 19.5 (range = 8-25) in the item searching task, 18 (range = 8-25) in the solid contaminant sampling task and 17.5 (range = 9-27) in the liquid contaminant sampling task. The total in situ workload ratings were also analyzed by task, and there was a significant main effect ($\chi^2(126) = 2.56, p = 0.012$). Post-hoc pairwise Wilcoxon tests

with Bonferonni corrections indicate that workload was rated significantly higher in the solid contaminant sampling task ($p = 0.033$) and nearly significantly higher in the item searching task ($p = 0.103$) and the liquid contaminant sampling task ($p = 0.096$) than in the photo searching task. Overall, workload in the photo searching task was rated lowest.

$H_1^{TS}$ involved evaluating the sensitivity of workload measures to changes in task workload; thus, it is important to examine why the in situ workload ratings were not sensitive to the workload level differences, given that the models predicted such sensitivity. Individual responses from participants were examined. The responses were ordered and participants were identified that rated high workload tasks with a score of 14 or under for more than four of eight tasks or a participant that rated a low workload task over 20 for four or more of eight tasks. These characteristics were defined as extreme unmatched ratings and resulted in three participants being classified as outliers. One participant was removed due to rating five high workload tasks of 14 or less, and two participants were outliers by rating four low workload tasks each over 20. These characteristics were identified as being extreme because they reflect providing low workload scores with an average individual channel rating of 3.5 or above (i.e., above the middle score of 3), and providing a high workload task with individual channel ratings of 2.5 or less (i.e., less than the middle score of 3). Consistently (i.e., for four or more tasks of eight) rating tasks incorrectly reflected that the specific person did not use the rating scale in the same way as the rest of the participants, due to a misunderstanding of rating channels or a misperception of workload.

After omitting the responses from the three identified participants, a Kruskal-Wallis test indicated that there was a significant main effect of workload, $\chi^2(1) = 4.07, p = 0.044$. The updated descriptive statistics are available in Table V.7. This result indicates that the majority of participants (thirteen participants) were sensitive to workload task changes, measured by the in situ subjective workload ratings.

#### V.3.3.2 Model Comparison with In Situ Workload Ratings

The model results were compared to the in situ subjective workload ratings and the subtask time evaluation results. Table V.8 presents the comparison between the H-H evaluation results and modeled workload. Table V.9 provides the same comparison for the H-R evaluation results and modeled H-R workload. Each workload channel collected via the situ subjective workload ratings were converted to the same scale modeled in IMPRINT Pro and summed to create a total workload value for comparison. The models' 95% confidence intervals were calculated in order to assess whether in situ workload ratings were within the range, which demonstrates that the models predicted the participants' workload accurately. The models did not predict the participant-rated workload within a 95% confidence interval for any task or workload level for either partner. The modeled results for both partners were within one standard deviation of the in situ workload ratings for

Table V.7: Median and range for in situ workload ratings in the Time-Structured evaluation by partner, workload level, and session.

| Metric | Analysis | | Median | Range |
|---|---|---|---|---|
| **Total In Situ Workload Ratings** | **By Partner** | **H-H** | 17 | 6-26 |
| | | **H-R** | 18 | 7-27 |
| | **By Workload Level** | **Low** | 17 | 6-25 |
| | | **High** | 18 | 9-27 |
| | **By Session** | **First** | 18 | 6-25 |
| | | **Second** | 17 | 7-27 |
| **Total In Situ Workload Ratings - 3 Outliers Removed** | **By Partner** | **H-H** | 17 | 6-25 |
| | | **H-R** | 17.5 | 7-27 |
| | **By Workload Level** | **Low** | 16 | 6-25 |
| | | **High** | 18 | 9-27 |
| | **By Session** | **First** | 17.5 | 6-25 |
| | | **Second** | 17 | 7-27 |

low workload tasks. Additionally, the H-R model's solid contaminant sampling task was also within one standard deviation of the in situ workload rating results. The high workload tasks' models overestimated the high workload level in situ workload ratings. Participants provided ratings on a similar level to the low workload task ratings.

Subtask time was also analyzed, in order to ensure that the models were accurate for timing. Tables V.10 and V.11 present the H-H evaluation results in comparison to the H-H modeled subtask time, and the H-R evaluation results in comparison to modeled H-R subtask time, respectively. The H-H results demonstrate that the item search task model was within the 95% confidence interval for both the high workload and low workload conditions. The H-R subtask time evaluation results indicated that the measured subtask times for the photo search, solid sampling, and liquid contaminant sampling tasks in the high workload condition were all within the model's 95% confidence interval. Each modeled subtask time was within one standard deviation of the evaluation results.

#### V.3.3.3 NASA Task Load Index

The NASA Task Load Index was examined along each of its channels: mental, physical, temporal, performance, effort, and frustration. The total score was also analyzed. The unweighted scores were used in this analysis, and the results were normally distributed. Each channel and the overall score were analyzed by partner (i.e., human or robot).

The overall mean score for H-H team participants was 50.06 (St. Dev. = 13.28) and was 53.38 (St. Dev. = 12.10) for H-R team participants. The means of the NASA Task Load Index responses are presented in Figure V.16. The mean frustration score for the H-H teams was 29.25 (St. Dev. = 20.82) and was 47.06

Table V.8: Descriptive statistics for the H-H condition In Situ Workload Ratings and the 95% confidence interval for the H-H model.

| | **Photo Search Low** | **Item Search Low** | **Solid Contaminant Sampling Low** | **Liquid Contaminant Sampling Low** | **Photo Search High** | **Item Search High** | **Solid Contaminant Sampling High** | **Liquid Contaminant Sampling High** |
|---|---|---|---|---|---|---|---|---|
| **H-H Model Mean** | 15.68 | 15.09 | 16.05 | 24.59 | 25.42 | 29.97 | 33.58 | 46.01 |
| **H-H Model St. Dev.** | 0.12 | 0.18 | 0.34 | 0.93 | 1.33 | 0.80 | 0.90 | 0.93 |
| **H-H Model 95% C.I. Low** | 15.61 | 14.98 | 15.84 | 24.01 | 24.60 | 29.47 | 33.02 | 45.43 |
| **H-H Model 95% C.I. High** | 15.75 | 15.20 | 16.26 | 25.17 | 26.24 | 30.47 | 34.14 | 46.59 |
| **H-H Converted In Situ Workload Mean** | 18.13 | 17.15 | 20.45 | 20.58 | 14.23 | 24.40 | 20.28 | 18.75 |
| **H-H Converted In Situ Workload St. Dev.** | 5.13 | 9.99 | 8.06 | 5.98 | 5.84 | 5.01 | 7.23 | 7.82 |
| **Within C.I.?** | No | No | No | No | No | No | No | No |
| **Model within one St. Dev. of Evaluation Results?** | **Yes** | **Yes** | **Yes** | **Yes** | No | No | No | No |

Table V.9: Descriptive statistics for the H-R condition In Situ Workload Ratings and the 95% confidence interval for the H-R model.

| | **Photo Search Low** | **Item Search Low** | **Solid Contaminant Sampling Low** | **Liquid Contaminant Sampling Low** | **Photo Search High** | **Item Search High** | **Solid Contaminant Sampling High** | **Liquid Contaminant Sampling High** |
|---|---|---|---|---|---|---|---|---|
| **H-R Model Mean** | 16.26 | 15.22 | 16.72 | 25.67 | 27.25 | 30.96 | 21.62 | 30.58 |
| **H-R Model St. Dev.** | 0.13 | 0.17 | 0.36 | 0.82 | 1.42 | 0.78 | 0.90 | 0.83 |
| **H-R Model 95% C.I. Low** | 16.18 | 15.11 | 16.50 | 25.16 | 26.37 | 30.48 | 21.06 | 30.07 |
| **H-R Model 95% C.I. High** | 16.34 | 15.33 | 16.94 | 26.18 | 28.13 | 31.44 | 22.18 | 31.09 |
| **H-R Converted In Situ Workload Mean** | 13.88 | 15.60 | 21.18 | 22.63 | 15.00 | 21.63 | 22.68 | 19.53 |
| **H-R Converted In Situ Workload St. Dev.** | 6.66 | 7.87 | 8.27 | 6.05 | 7.47 | 4.44 | 5.87 | 8.34 |
| **Within C.I.?** | No | No | No | No | No | No | No | No |
| **Model within one St. Dev. of Evaluation Results?** | **Yes** | **Yes** | **Yes** | **Yes** | No | No | **Yes** | No |

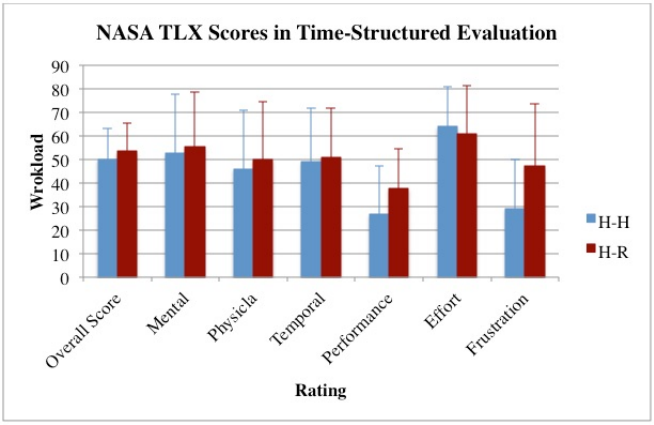
Table V.10: Descriptive statistics for the H-H condition subtask time and the 95% confidence interval for the H-H model subtask time.

| | **Photo Search Low** | **Item Search Low** | **Solid Contaminant Sampling Low** | **Liquid Contaminant Sampling Low** | **Photo Search High** | **Item Search High** | **Solid Contaminant Sampling High** | **Liquid Contaminant Sampling High** |
|---|---|---|---|---|---|---|---|---|
| **H-H Model Mean** | 284.33 | 66.67 | 288.50 | 290.17 | 221.33 | 81.04 | 255.08 | 251.83 |
| **H-H Model St. Dev.** | 19.68 | 17.71 | 14.01 | 12.32 | 41.48 | 17.71 | 30.70 | 29.94 |
| **H-H Model 95% C.I. Low** | 272.13 | 55.69 | 279.82 | 282.53 | 195.62 | 70.06 | 136.05 | 232.52 |
| **H-H Model 95% C.I. High** | 296.53 | 77.65 | 297.18 | 297.81 | 247.04 | 92.02 | 274.11 | 269.64 |
| **H-H Subtask Time Mean** | 331.25 | 61.71 | 265.38 | 263.25 | 266.75 | 89.41 | 319.31 | 293.25 |
| **H-H Converted In Situ Workload St. Dev.** | 59.87 | 21.31 | 52.20 | 44.59 | 61.98 | 25.68 | 99.58 | 54.96 |
| **Within C.I.?** | No | **Yes** | No | No | No | **Yes** | No | No |
| **Model within one St. Dev. of Evaluation Results?** | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** |

Table V.11: Descriptive statistics for the H-R condition subtask time and the 95% confidence interval for the H-R model subtask time.

| | **Photo Search Low** | **Item Search Low** | **Solid Contaminant Sampling Low** | **Liquid Contaminant Sampling Low** | **Photo Search High** | **Item Search High** | **Solid Contaminant Sampling High** | **Liquid Contaminant Sampling High** |
|---|---|---|---|---|---|---|---|---|
| **H-R Model Mean** | 299.50 | 70.75 | 307.41 | 310.33 | 232.02 | 89.34 | 277.27 | 275.54 |
| **H-R Model St. Dev.** | 13.50 | 20.61 | 14.50 | 27.63 | 45.09 | 18.97 | 35.55 | 35.74 |
| **H-R Model 95% C.I. Low** | 291.13 | 57.98 | 298.39 | 296.30 | 204.07 | 77.58 | 255.24 | 253.39 |
| **H-R Model 95% C.I. High** | 307.87 | 83.52 | 316.43 | 324.36 | 259.97 | 101.10 | 299.30 | 297.69 |
| **H-R Subtask Time Mean** | 266.75 | 89.41 | 319.31 | 293.25 | 216.77 | 70.63 | 279.82 | 262.76 |
| **H-R Converted In Situ Workload St. Dev.** | 61.88 | 25.68 | 99.58 | 54.96 | 88.81 | 21.65 | 78.95 | 81.66 |
| **Within C.I.?** | No | No | No | No | **Yes** | No | **Yes** | **Yes** |
| **Model within one St. Dev. of Evaluation Results?** | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** |

Figure V.16: Mean NASA Task Load Index responses in the Time-Structured evaluation.




(St. Dev. = 26.47) for the H-R teams. Comparisons were made using t-tests, and there were no significant comparisons, with the exception of the frustration channel, $t(30) = 2.12, p = 0.043$. Participants with robot partners were significantly more frustrated during the tasks.

### V.3.4 Timing Task Performance Metrics

The timing metrics include reaction and response time for primary and secondary tasks, and the subtask time. The timing data was collected for all available instances during video and audio file coding. Background noise and team conversation prevented audible stimulus onsets for a small subset of the tasks. All the timing metrics were not normally distributed and were analyzed using non-parametric statistics. The descriptive statistics (i.e., median and interquartile range) are presented in Table V.12. All timing metrics are analyzed by partner, workload level, and session.

#### V.3.4.1 Primary Task Reaction Time

Primary task reaction time median and range by partner, workload level, and session is presented in Table V.12. The mean secondary task reaction time in the H-H teams was 1.16 (St. Dev. = 1.27) s and was 1.49 (St. Dev. = 1.38) s in the H-R teams. H-H team primary task reaction time was demonstrated to be significantly shorter than in the H-R team tasks, $\chi^2(1) = 29.99, p < 0.001$. Kruskal-Wallis tests indicated no significant main effects of workload level or session on primary task reaction time. Participants demonstrated improved task performance via faster primary task reaction time in H-H teams.

#### V.3.4.2 Secondary Task Reaction Time

The median and range for secondary task reaction time by partner, workload level, and session is available in Table V.12. Additionally, the mean secondary task reaction time in the low workload level was 1.13

Table V.12: Median and range for reaction and response time metrics in the Time-Structured evaluation by partner, workload level, and session.

| Metric | Analysis | | Median | IQR |
|---|---|---|---|---|
| **Primary Task Reaction Time** | **By Partner** | **H-H** | 0 | 1-2 |
| | | **H-R** | 0 | 1-2 |
| | **By Workload Level** | **Low** | 0 | 1-2 |
| | | **High** | 0 | 1-2 |
| | **By Session** | **First** | 0 | 0-1 |
| | | **Second** | 0 | 1-2 |
| **Secondary Task Reaction Time** | **By Partner** | **H-H** | 1 | 1-1 |
| | | **H-R** | 1 | 0.5-1 |
| | **By Workload Level** | **Low** | 1 | 1-1 |
| | | **High** | 1 | 0-1 |
| | **By Session** | **First** | 1 | 0.5-1 |
| | | **Second** | 1 | 1-1 |
| **Primary Task Response Time** | **By Partner** | **H-H** | 2 | 1-4 |
| | | **H-R** | 1 | 1-3 |
| | **By Workload Level** | **Low** | 1 | 1-3 |
| | | **High** | 1 | 1-3 |
| | **By Session** | **First** | 1 | 1-4 |
| | | **Second** | 1 | 1-3 |
| **Secondary Task Response Time** | **By Partner** | **H-H** | 1 | 1-2 |
| | | **H-R** | 1 | 0.75-2 |
| | **By Workload Level** | **Low** | 1 | 1-1 |
| | | **High** | 1 | 1-2 |
| | **By Session** | **First** | 1 | 1-1 |
| | | **Second** | 1 | 1-2 |
| **Subtask Time** | **By Partner** | **H-H** | 136.5 | 55-234.25 |
| | | **H-R** | 185 | 78.5-268 |
| | **By Workload Level** | **Low** | 222.5 | 77.75-283 |
| | | **High** | 153 | 62-220 |
| | **By Session** | **First** | 185.5 | 67.75-282 |
| | | **Second** | 156 | 70-229 |

(St. Dev. = 0.84) s and was 0.89 (St. Dev. = 0.88) s in the high workload level. Kruskal-Wallis tests indicated no significant main effects of partner or session on secondary task reaction time. High workload secondary task reaction time was demonstrated to be significantly shorter than in the low workload tasks, $\chi^2(1) = 9.86, p = 0.002$. Participants demonstrated improved task performance via faster secondary task reaction time in high workload tasks.

#### V.3.4.3 Primary Task Response Time

The median and range for primary task response time by partner, workload level, and session is presented in Table V.12. The mean primary task response time in H-H teams was 2.76 (St. Dev. = 3.07) s and was 2.71 (St. Dev. = 3.98) s in the H-R teams. Kruskal-Wallis tests indicated no significant main effects of workload level or session on primary task response time. H-R team primary task response time was demonstrated to be significantly shorter than in the H-H team tasks, $\chi^2(1) = 4.322, p = 0.04$. Participants demonstrated improved task performance via faster primary task response time in H-R teams; H-R participants reacted significantly slower to the primary task, but they enacted the appropriate response, marked by primary task response time, more quickly than the H-H participants.

#### V.3.4.4 Secondary Task Response Time

The median and range for secondary task response time by partner, workload level, and session is presented in Table V.12. The mean secondary task response time in low workload level tasks was 1.48 (St. Dev. = 1.92) s and was 1.29 (St. Dev. = 1.55) s in high workload level tasks. Kruskal-Wallis tests indicated no significant main effects of partner or session on secondary task response time. High workload secondary task response time was nearly significantly shorter than in the low workload tasks, $\chi^2(1) = 3.70, p = 0.05$. Participants demonstrated improved task performance via faster secondary task response time in high workload tasks.

#### V.3.4.5 Subtask Time

The mean subtask time in H-H teams was 152.49 (St. Dev. = 110.61) s and was 188.23 (St. Dev. = 113.48) s in H-R teams. The mean subtask time was 197.86 s in low workload tasks (St. Dev. = 116.82) s and was 156.82 (St. Dev. = 108.91) s in high workload tasks. The mean subtask time for the first evaluation session was 183.66 (St. Dev. = 126.51) s and was 157.95 (St. Dev. = 94.59) s in the second evaluation session. Further descriptive statistics are available in Table V.12. Kruskal-Wallis tests indicated that H-R teams had significantly longer subtask times ($\chi^2(1) = 15.70, p < 0.001$), and participants had significantly longer subtask times in low workload tasks ($\chi^2(1) = 18.73, p < 0.001$). A Kruskal-Wallis test indicated no significant main effect of session on subtask time. These results indicate that participants increased task

performance (i.e., faster subtask time) with a human partner and with higher workload tasks.

### V.3.5 Other Objective Task Performance Metrics

#### V.3.5.1 Primary Task Failure Rate

Overall primary task failure rate was analyzed by partner and session. The mean overall primary task failure rate in H-H teams was 11.42% (St. Dev. = 6.06%) and was 14.16% (St. Dev. = 8.68%) in H-R teams; the mean was 14.96% (St. Dev. = 8.72%) in the first session and 10.62% (St. Dev. = 5.48%) in the second session. A Kruskal-Wallis test indicated that there was no significant main effect of partner or session on primary task failure rate, but that the difference between sessions was nearly significant, $\chi^2(1) = 2.63, p = 0.10$.

The failure rates of each of the five primary task components were analyzed by partner, workload level, and session. Figure V.17 presents the mean primary task failure rate for each component, and overall. The means and St. Dev. for each task component are presented in Table V.13. Participants struggled with Detail-tracking tasks the most. The H-R teams failed at Effort component tasks at a rate that was not significant, but a Kruskal-Wallis test indicated that the H-R teams tended to fail more often than H-H teams, $\chi^2(1) = 2.67, p = 0.10$. A Kruskal-Wallis indicated that participants failed on Assessment tasks significantly more often in their first session than in their second session, $\chi^2(1) = 7.70, p < 0.01$. There were no significant main effects of partner, workload level, or session on the Attention or Detail-Tracking component failure rates. There was a significant main effect of workload level on the Achievement task component failure rate, $\chi^2(1) = 28.89, p < 0.001$, indicating that failure rate was higher in high workload tasks. Additionally, there were no significant main effects of workload level on the Effort or Assessment component failure rate, no main effect of session on the Effort component failure rate, and no effect of partner on the Assessment component failure rate.

The results of the primary task failure rate analysis indicates that while there was no significant difference between H-H and H-R teams in general, H-R teams had significantly worse task performance specifically in effort tasks. Additionally, teams generally improved during their second session, as evidenced by Assessment component failure rate, and performance was worse in high workload tasks, as evidenced by the Achievement task component results.

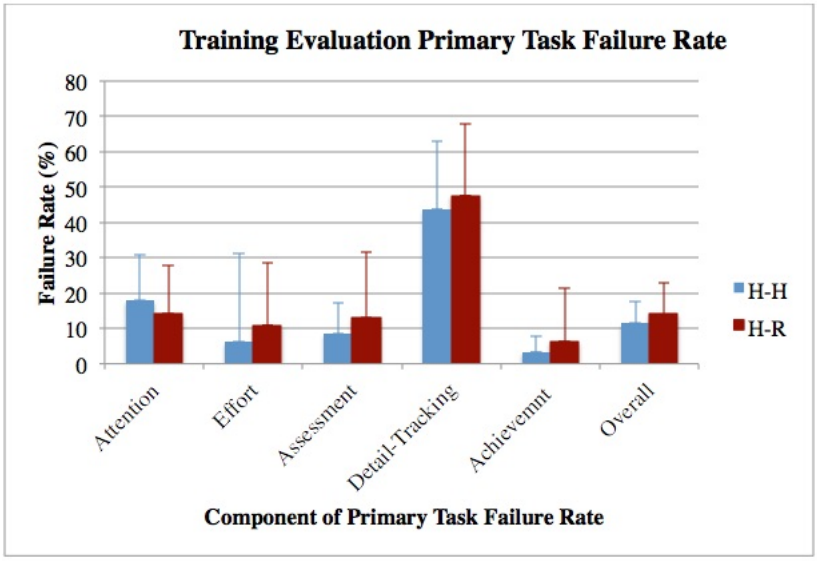
#### V.3.5.2 Work Completed

Work Completed was not analyzed in the Time-Structured evaluation because this metric is designed to analyze workload without factoring in the time it takes to complete a task. Each task in the Time-Structured evaluation was timed for fifteen minutes; thus, calculating work completed does not add insight to workload analysis.

Table V.13: Median and range for reaction and response time metrics in the Time-Structured evaluation by partner, workload level, and session.

| Metric | Analysis | | Mean | St. Dev. |
|---|---|---|---|---|
| **Attention Component Failure Rate** | **By Partner** | **H-H** | 17.97 | 12.88 |
| | | **H-R** | 14.06 | 13.60 |
| | **By Workload Level** | **Low** | 18.75 | 14.43 |
| | | **High** | 13.28 | 11.61 |
| | **By Session** | **First** | 12.50 | 12.08 |
| | | **Second** | 19.53 | 13.67 |
| **Effort Component Failure Rate** | **By Partner** | **H-H** | 6.25 | 25.00 |
| | | **H-R** | 10.94 | 17.66 |
| | **By Workload Level** | **Low** | 12.50 | 27.04 |
| | | **High** | 4.69 | 13.60 |
| | **By Session** | **First** | 13.84 | 28.39 |
| | | **Second** | 3.35 | 9.17 |
| **Assessment Component Failure Rate** | **By Partner** | **H-H** | 8.53 | 8.87 |
| | | **H-R** | 12.89 | 18.59 |
| | **By Workload Level** | **Low** | 17.97 | 29.94 |
| | | **High** | 6.64 | 8.08 |
| | **By Session** | **First** | 16.86 | 17.81 |
| | | **Second** | 4.56 | 6.05 |
| **Detail-Tracking Component Failure Rate** | **By Partner** | **H-H** | 43.75 | 19.32 |
| | | **H-R** | 47.50 | 20.20 |
| | **By Workload Level** | **Low** | 50.00 | 17.21 |
| | | **High** | 41.25 | 21.25 |
| | **By Session** | **First** | 44.17 | 18.03 |
| | | **Second** | 47.08 | 21.43 |
| **Achievement Component Failure Rate** | **By Partner** | **H-H** | 3.33 | 4.42 |
| | | **H-R** | 6.35 | 14.87 |
| | **By Workload Level** | **Low** | 0.00 | 0.00 |
| | | **High** | 9.49 | 15.12 |
| | **By Session** | **First** | 6.99 | 14.38 |
| | | **Second** | 2.69 | 5.37 |

Figure V.17: Primary task failure rate for Time-Structured evaluation.




#### V.3.5.3 Distance Traveled

Distance traveled was measured using a Fitbit pedometer and using the Scosche Rhythm+ armband fitness activity monitor. Both of these monitors are commercially available. Distance traveled mean and standard deviation is presented for both monitors by partner, session, and workload level in Table V.14. The total distance traveled for each task was analyzed for a significant effect of partner, session, and workload level. The Scosche Rhythm+ records distance traveled via GPS signal, and since the evaluation occurred indoors, the data was unreliable.

There was no main effect of workload level on the Fitbit recorded number of steps taken. There was a significant main effect of the evaluation session, $\chi^2(1) = 5.13, p = 0.024$, indicating that participants walked further in the second session than the first. Additionally, there was a nearly significant main effect of partner ($\chi^2(1) = 2.90, p = 0.089$), indicating that people tended to walk further with a robot partner. Distance traveled was correlated to vector magnitude, another measurement of gross physical movement. The plot of each participant's steps taken and vector magnitude is presented in Figure V.18. The green line represents the positive correlation between the two measures. The total number of steps taken by the participant was significantly correlated to vector magnitude, $r(112) = 0.26, p = 0.005$. The Fitbit Zip activity monitor was a good choice as a measure of distance traveled.

### V.3.6 Subjective Post-Trial Questionnaire

Analysis of the post-trial questionnaire for the Time-Structured evaluation was performed using non-parametric analysis. Median agreement responses for each of the nineteen questions are provided in Table V.15. Overall,

Figure V.18: Comparison of vector magnitude gross movement data to the distance traveled, measured by Fitbit Zip commercially available activity monitor.

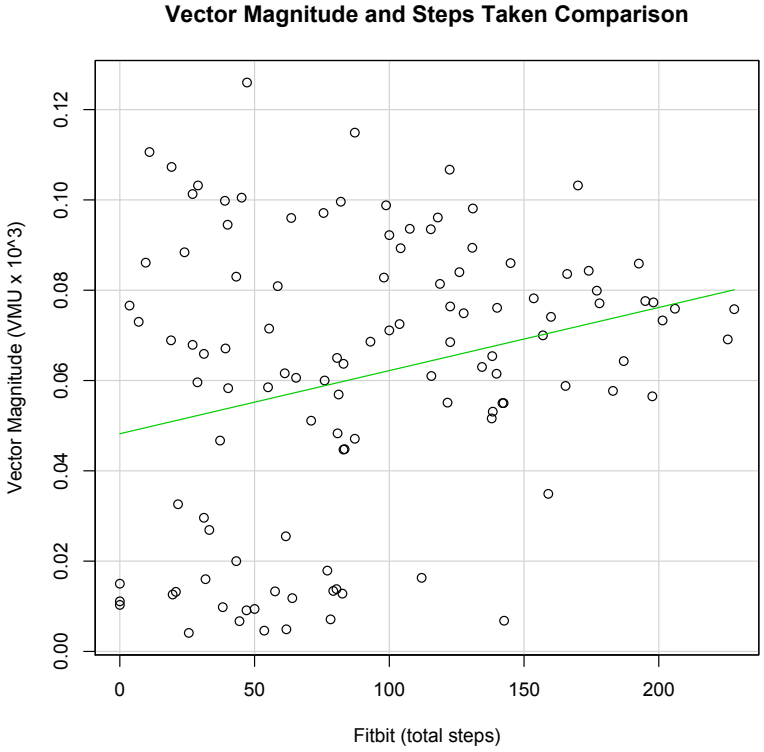



Table V.14: Descriptive statistics in the Time-Structured evaluation for distance traveled, as measured by two sensors, by workload level, partner, and session.

| Distance Traveled Sensor | Analysis | | Mean | St. Dev. |
|---|---|---|---|---|
| **Fitbit (steps)** | **By Partner** | **H-H** | 102.38 | 64.54 |
| | | **H-R** | 81.59 | 49.08 |
| | **By Workload Level** | **Low** | 92.44 | 57.18 |
| | | **High** | 91.16 | 14.66 |
| | **By Session** | **First** | 80.05 | 55.94 |
| | | **Second** | 103.97 | 57.12 |

the median response for H-H team participants was a 7 (range of 1 to 9) and was a 6 for H-R team participants (range of 1 to 9). A Kruskal-Wallis test indicated that the H-R condition resulted in significantly lower agreement with the question statements, $\chi^2(1) = 12.38, p < 0.001$.

Additionally, Kruskal-Wallis tests indicated that there was a significant main effect of question number ($\chi^2(18) = 235.78, p < 0.001$) and that there was a significant interaction effect of the question number and condition ($\chi^2(80) = 366.79, p < 0.001$); however, the follow-up pairwise Wilcoxon tests with Bonferonni corrections for family-wise error indicated no significant differences between conditions for any individual questions. The large number of questions increases the number of post-hoc comparisons, which decreases the p-value necessary for significance, and decreases the likelihood of obtaining a significant post-hoc test result. The post-trial questionnaire results can still be observed for trends between H-H and H-R teams on individual questions. The difference between partner types closest to significant was for Statement 3 ($p = 0.12$), indicating that participants partnered with a robot felt less comfortable communicating with his or her teammate.

The final preference questionnaire indicated that participants preferred working with the human teammate. Twelve participants selected the human teammate (5 participants from the human first group, 7 participants from the robot first group). Three participants indicated that they did not have a strong preference (2 from the human first group and 1 from the robot first group). One participant was in the human first group and preferred working with the robot. The median agreement with the statement, "I preferred working with one partner over another," was 7 in both groups, with a range of 4 to 8 in the human first group and a range of 1 to 9 in the robot first group.

### V.3.7 Summary

The Time-Structured evaluation assessed workload and task performance metrics via a first response training scenario with four tasks: the photo search task, the item search task, the solid contaminant sampling task, and the liquid contaminant sampling task. The results demonstrated the successful completion of tasks by human-robot teams.

Workload metrics that demonstrated a main effect of task workload level include: respiration rate, vector magnitude, task density, secondary task failure rate, and in situ workload ratings (after removal of three outliers). Heart rate variability, postural load, variance in posture, vector magnitude, and working memory demonstrated correlations to workload levels. The regression analysis indicated that heart rate variability changes were influenced by mental workload while heart rate and respiration rate were influenced by physical workload. H-H and H-R teams, overall, did not have a measurable difference in workload. The human performance model was able to predict workload within one standard deviation of evaluation means for low

Table V.15: Descriptive statistics for Time-Structured evaluation post-trial questionnaire data.

| Statement Number | Statement | H-H | | H-R | |
|---|---|---|---|---|---|
| | | Median | Range | Median | Range |
| 1 | My teammate gave me clear instructions. | 8.5 | 7-9 | 8 | 4-9 |
| 2 | I trusted my teammate. | 9 | 8-9 | 7.5 | 2-9 |
| 3 | I felt comfortable communicating with my teammate. | 9 | 7-9 | 6 | 2-9 |
| 4 | My teammate understood what I was trying to communicate. | 8.5 | 7-9 | 7 | 4-9 |
| 5 | I did a good job on the tasks I was assigned. | 8 | 3-9 | 6 | 4-9 |
| 6 | I often felt confused about my teammate's instructions. | 2 | 1-7 | 3 | 1-8 |
| 7 | I often felt confused as to the purpose of my actions. | 2.5 | 1-5 | 2.5 | 1-8 |
| 8 | I felt stressed during the scenario. | 4 | 1-7 | 3 | 1-9 |
| 9 | My teammate helped me identify possible alternatives. | 7 | 4-9 | 5.5 | 1-9 |
| 10 | I understood the problem. | 8 | 5-9 | 7.5 | 4-9 |
| 11 | I understood my teammate's concerns (or suggestions). | 8 | 5-9 | 7 | 5-9 |
| 12 | I made contribution to the decision-making process. | 7.5 | 5-9 | 7 | 5-9 |
| 13 | I felt I was in charge of the team. | 4.5 | 2-9 | 6.5 | 1-9 |
| 14 | My teammate led the team in the investigation. | 6 | 1-9 | 6.5 | 1-9 |
| 15 | I felt I had greater responsibility in the team than my teammate. | 5.5 | 2-9 | 6 | 4-9 |
| 16 | My teammate asked for help from me when it was needed. | 6 | 2-9 | 5.5 | 1-9 |
| 17 | I understand my teammate's strengths and abilities. | 8 | 3-9 | 7 | 1-9 |
| 18 | My teammate was collaborative. | 8 | 5-9 | 7.5 | 3-9 |
| 19 | I believe I could have completed this task on my own as well or better than I did with my teammate. | 3 | 1-5 | 3.5 | 1-9 |

workload tasks for H-H and H-R teams.

Task performance metrics evaluated whether H-H and H-R teams experienced differences. Timing metrics primary task reaction time, primary task response time, subtask time, distance traveled, and primary task failure rate measured differences between H-H and H-R teams. Primary task failure rate, secondary task reaction time and secondary task response time measured differences between workload levels. The human performance model was able to predict subtask time for five of the sixteen modeled tasks, and within one standard deviation of evaluation means for each task.

A comparison of the results to the previous two evaluations, the Guided and Collaborative evaluations, must be performed in order to fully understand the impact of the Time-Structured evaluation results. Chapter V.4 discusses the relationship between the results of the metrics measured in the three evaluations.

## V.4 Comparison with Guided and Collaborative Evaluations

The three evaluations each investigated human-robot peer-based teams in first response domains, but each evaluation included a different perspective. The Time-Structured evaluation placed a time limit on the four tasks required of the participants, unlike the unlimited time given in the previous two evaluations. The participants were given a structured fifteen minutes to complete a task in either a low or high workload level. Additionally, the Time-Structured evaluation required that each participant completed all tasks with both a human and a robot in a repeated-measures design, unlike in the Guided and Collaborative evaluations' between-subjects designs. The Guided and Collaborative evaluations were compared to one another in Chapter IV. This chapter will evaluate the differences between the previous two evaluations and the Time-Structured evaluation.

### V.4.1 Physiological Measures of Workload

Heart rate, respiration rate, and heart rate variability all demonstrated inconsistent results across the three evaluations. Specifically, opposing significant respiration rate findings in the Guided and Collaborative evaluations related to the same trends in mental and physical workload; thus, the larger difference in physical activity between the two evaluations may have influenced respiration rate and heart rate. This result was confirmed with a regression analysis using the Time-Structured evaluation that confirmed respiration rate's influence by physical activity.

Heart rate variability was demonstrated to be predicted by in situ workload ratings, a subjective workload rating, and not by vector magnitude, a physical workload measure. Heart rate variability demonstrated a main effect of triage level in the Guided evaluation, but no main effects in the Collaborative evaluation, due to a lack of successful manipulation of workload levels. The Time-Structured evaluation did not result in a main

effect of workload level, but did find a positive correlation between in situ workload ratings and heart rate variability.

Variance in posture and vector magnitude were demonstrated to be higher in H-R teams in the Guided and Collaborative evaluations, but were not significantly higher in the Time-Structured evaluation. The more structured nature with applied time pressure in the Time-Structured evaluation tasks may affected these results.

### V.4.2 Other Objective Metrics of Workload

Working memory results in the Guided and Time-Structured evaluations indicated no difference between conditions, but H-R condition teams had higher memory recall scores in the Collaborative evaluation. The longer time spent on task in the Collaborative evaluation may have allowed participants to remember items more easily for recall than H-H participants.

Task density results presented the same findings in the Collaborative and the Time-Structured evaluations: H-H teams had higher task density. Task density is a workload measure and this result indicates that workload may still be lower when the task time is constrained in the Time-Structured evaluation, due to the longer subtask time.

Secondary task failure rate results in the Guided and Collaborative evaluations indicated no difference between conditions. The Time-Structured evaluation was designed to be more difficult than the previous two evaluations and create more demand on the participant when completing the secondary task, in order to create a higher secondary task failure rate, generally, in order to analyze trends. The Time-Structured evaluation also indicated no difference between H-H and H-R teams; there was also significantly higher failure rate in high workload tasks, indicating that making the task more challenging was partially successful.

### V.4.3 Subjective Metrics

The in situ workload ratings indicated that H-R workload levels were significantly lower than H-H condition workload levels in both the Guided and Collaborative evaluations. Analysis of other metrics, such as work completed (see Chapter IV) indicated that the lower workload in the Guided and Collaborative evaluation H-R condition was likely due to the longer time taken for the H-R teams to complete tasks. The Time-Structured evaluation constrained task times and did not measure any difference between H-H and H-R teams with in situ workload ratings.

The NASA Task Load Index questionnaire indicated very similar results in all three evaluations. There were no significant differences in overall scores and H-R teams experienced higher frustration channel scores.

### V.4.4 Timing Measures

The measures of reaction time and response time were compared between H-H and H-R teams in the Collaborative evaluation and demonstrated longer time taken to react and respond in H-R teams. The Time-Structured evaluation results indicated that primary task reaction time was slower in H-R teams, supporting this result; however, primary task response time indicated the opposite result: H-R participants were faster to respond. This result may indicate that once participants take the longer time to recognize the primary task prompt and react for primary task reaction time, they are closer to comprehending and making an appropriate response, marking primary task response time. The primary task reaction time took

Secondary task reaction and response time were longer for H-R teams in the Collaborative evaluation, but this result was not corroborated in the Time-Structured evaluation. These results conflict due to the fact that the robot was responsible for providing secondary task prompts in the Collaborative evaluation in H-R teams, but a human experimenter prompted all secondary tasks in the Time-Structured evaluation. Recognizing the robot's voice did not play a role in secondary task reaction and response time, but the participant was required to converse with the robot and interrupt it to relay the secondary task messages in the Time-Structured evaluation, playing a role in recorded secondary task reaction and response times.

Subtask time was demonstrated in all three evaluations to take longer in H-R teams. The robot moves more slowly and speaks slower, measurably affecting the time taken to complete tasks. The H-R teams were an average of 1.11 times slower in the Guided evaluation, 1.26 times slower in the Collaborative evaluation, and 1.23 times slower in the Time-Structured evaluation when completing subtasks (mean = 1.20 times slower).

### V.4.5 Primary Task Failure Rate

Primary task failure rate results indicated a similar trend in all three evaluations: H-R participants fail more often at effort-related tasks. Additionally, overall primary task failure rate was not significantly different between H-H and H-R conditions in all three evaluations; thus, failing more at effort tasks does not indicate that the H-R team members are suffering significantly more in overall task performance.

### V.4.6 Distance Traveled

Distance traveled was not easily measured in the Guided or Collaborative evaluations, but the pedometer results in the Collaborative evaluation suggested that participants walk farther with a robot partner. These results were supported by the Time-Structured evaluation results with the Fitbit Zip.

### V.4.7 Post-Trial Questionnaire

The Guided evaluation analysis of the post-trial questionnaire included eight questions and found that participants rated the human partner higher for statements related to trust, clarity of instructions, and communication. Adding more questions to the questionnaire in the Collaborative and Time-Structured evaluations also increased the difficulty of achieving a significant result, due to family-wise error corrections; however, a nearly significant result in the Time-Structured evaluation indicates that participants felt less comfortable communicating with the robot. Additionally, the general responses on all three questionnaires follow similar trends. Participants tended to rate trust and confidence in the human higher and self-reliance with the robot higher (e.g., statements 13, 19).

### V.4.8 Summary

Overall, the three evaluations demonstrated similar trends for a majority of the metrics. A subset of the metrics (i.e., respiration rate, heart rate, primary task response time) demonstrated conflicting results. A summary table of all workload metrics is presented in Table V.16 and a summary of all task performance metrics is presented in Table V.17. The bold metrics are recommended for use in human-robot peer-based teams.

## V.5 Discussion

$H^1_{TS}$ stated that the assigned low workload tasks will elicit lower levels of workload than the assigned high workload tasks. This hypothesis was supported via the workload measures of task density, secondary task failure rate, in situ workload ratings, and vector magnitude. The experimental manipulation of workload was successful. A subset of the metrics that did not demonstrate the manipulation of workload (i.e., heart rate variability, variance in posture, working memory); however, these metrics demonstrated a significant correlation to in situ workload ratings.

Respiration rate did demonstrate an effect of workload level, but it was also demonstrated to be predicted by vector magnitude, rather than the in situ workload ratings. This result indicates that respiration rate is susceptible to influence from physical activity, which prevents its use as a mental workload measurement in mobile tasks. Heart rate also demonstrated the influence of physical activity. Heart rate variability, however, was predicted by the in situ workload ratings and not by vector magnitude; thus, heart rate variability can be used in mobile peer-based human-robot teams as a measure of mental workload.

$H^2_{TS}$ investigated whether H-R workload was lower than H-H workload due to slower task completion. $H^2_{TS}$ is supported by the recorded slower subtask times. Slower subtask time dictates calculation of workload, and making calculations, such as task density, will result as lower in the H-R teams with longer subtask times.

Table V.16: Summary of workload metrics from Guided, Collaborative, and Time-Structured evaluations. Bold metrics are recommended for measurement in human-robot peer-based teams.

| Metric | Measurement Technique Notes | Results | Recommendation? |
|---|---|---|---|
| Heart Rate | Bioharness chest strap | Related to physical activity | Not good choice for physically active tasks |
| | Scosche Rhythm+ arm band | Correlated to Bioharness results, tended to be lower values | Possible to use for measuring heart rate, but overall is not good mental workload metric for physically active tasks |
| Respiration Rate | Bioharness chest strap | Related to physical activity | Not good choice for physically active tasks |
| **Heart Rate Variability** | Bioharness chest strap | Related to mental workload; demonstrated mixed results regarding sensitivity to workload manipulation levels | Good physiological mental workload metric for active tasks |
| Variance in Posture | Bioharness chest strap | Not consistently sensitive to workload manipulations | More useful for tasks with torso movement, bending |
| **Vector Magnitude** | Bioharness chest strap | Sensitive to workload manipulation | Good representation of physical workload |
| **Working Memory** | Written, post-evaluation | Higher workload increased memory recall in Time-Structured evaluation | Useful, given that there is an appropriate time to gather memory feedback |
| **Task Density** | Number of tasks initiated divided by subtask time | Lower in H-R teams and higher in high workload tasks in Time-Structured evaluation | Recommended objective measure of workload |
| **Secondary Task Failure Rate** | Video / Audio coding | Higher failure rate in high workload tasks in Time-Structured evaluation | Recommended when use of secondary task does not intrude on primary task, or is already required |
| **In Situ Workload Ratings** | Verbal questions after each task | Lower in H-R teams in Guided and Collaborative; sensitive to workload manipulations in Guided and Time-Structured evaluations | Recommended for quick subjective workload probe |
| NASA Task Load Index | Written questionnaire post-evaluation | No differences between conditions in overall workload | Only recommended for evaluations, as it is time-consuming, written (or computer-based), and must occur following the task |

Table V.17: Summary of task performance metrics from Guided, Collaborative, and Time-Structured evaluations. Bold metrics are recommended for measurement in human-robot peer-based teams.

| Metric | Measurement Technique Notes | Results | Recommendation? |
|---|---|---|---|
| **Primary Task Reaction Time** | Video / Audio coding | Longer in H-R teams | Useful when tasks have clear onset stimuli |
| **Secondary Task Reaction Time** | Video / Audio coding | Shorter in low workload tasks in Task-Structured evaluation | Useful when tasks have clear onset stimuli |
| **Primary Task Response Time** | Video / Audio coding | Longer in H-R in Collaborative evaluation, Shorter in H-R in Time-Structured evaluation | Useful when tasks have clear onset stimuli and responses |
| **Secondary Task Response Time** | Video / Audio coding | Shorter in low workload tasks in Task-Structured evaluation | Useful when tasks have clear onset stimuli and responses |
| **Subtask Time** | Video / Audio coding | Shorter subtask time in high workload tasks in Time-Structured evaluation; longer subtask time in H-R overall | Subtask time is useful task performance metric |
| **Primary Task Failure Rate** | Video / Audio coding | Sensitivity when broken into components of task | Useful to see more detail of task failure |
| **Distance Traveled** | **Fitbit (steps taken)** | Correlated to vector magnitude | Useful to measure length of participant travel |
| | Scosche Rhythm+ arm band GPS tracker | Unreliable data generated | Not reliable method indoors |
| | Garmin pedometer | Unreliable data generated | Not reliable method indoors |
| | Peak Acceleration - Bioharness chest strap | Does not provide continuous data, just peaks, though sensitive to workload level manipulations | Not a relevant metric |

Even though the H-R teams completed subtasks more slowly, the in situ workload ratings were not sensitive to differences between task partners during this evaluation. Subjective rating scales, such as the in situ subjective workload ratings, are not ideal measurements for measuring workload for many reasons, including the inability to gather the metric during the task. Unlike the secondary task responses, which can blend into the task, the in situ workload questions are removed from the primary task. Additionally, the ratings may have been more sensitive to differences between H-H and H-R condition workload because the differences between subtask times in those two evaluations were much larger, stretching the work completed over a longer duration and creating lower workload. The average of 1.2 times slower that H-R teams completed subtasks over all three evaluations is unlikely to decrease in peer-based interaction without improvements to robot capabilities, namely speech recognition software and robot mobility. Developing peer-based teams with voice interaction depends on speech systems.

$H^3_{TS}$ assessed whether task performance was lower in H-R teams. This hypothesis was only partially supported. Subtask time and primary task reaction time indicated that H-R teams were slower, but primary task response time indicated that H-R teams were faster. The faster time to respond may indicate that the H-R participants were simplifying their speech for the robot and uttering simpler, and faster, replies. The video recordings demonstrate examples of more succinct replies uttered by H-R participants. Alternatively, the participant may be more frustrated with the robot (as indicated by NASA Task Load Index results) and attempting to speed the robot along by responding quickly. The increased responsibility for the team that the participant may have felt, as reported in the post-trial questionnaire, also may have influenced the participant to take charge of the conversation and respond more quickly.

Additionally, the H-R teams may benefit somewhat from the speed-accuracy trade-off by being forced to slow the task steps down more than a H-H team. Human-robot teams were expected to suffer in all measures of task performance due to their slower performance; however, this was not the case. The subtasks took significantly longer, but did not result in a significant difference in overall primary task failure rate (with the notable exception of the Effort component tasks). Participants may be better able to catch their own mistakes when given more time because of the robot's slower speed to think and plan actions.

$H^4_{TS}$ stated that there was an interaction between workload and task performance in the Time-Structured evaluation. This evaluation was supported by demonstrating a main effect of workload level on task performance metrics (i.e., subtask time, secondary task reaction time, secondary task response time, and the Achievement primary task failure rate component), demonstrating that workload changes affected task performance. The effect is not the same for all metrics; the timing metrics improve in high workload tasks, while Achievement failure rate increases. Participants improved their timing, but failed more frequently at the Achievement component. This result may be due to the time limits placed on the tasks, preventing

participants from finishing subtasks in the high workload tasks.

The Time-Structured evaluation demonstrated the use of the workload and task performance metrics (presented in Chapter IV) in a mobile, human-robot peer-based teaming scenario. The majority of metrics indicated that they are appropriate choices for use in this field (see Tables V.16 and V.17).

# Chapter VI

## Conclusions, Contributions, and Future Work

### VI.1 Conclusions

The focus of this dissertation was the analysis of appropriate metrics in the three evaluations for peer-based human-robot teams in the first response domain. Many domains, such as the first response domain, have specific constraints that will impact the measurement of task performance and workload. These constraints include unpredictable environmental conditions, mobile tasks, specialized protective gear, and time-critical task goals. The investigated metrics addressed these constraints and included a variety of appropriate measurement methods. The metric-associated results provided by the three evaluations and the human performance models inform how to measure task performance and workload for mobile human-robot peer-based teams.

This dissertation investigated differences in workload and task performance between human-human and human-robot peer-based teams. Three evaluations were performed: the Guided evaluation, the Collaborative evaluation, and the Time-Structured evaluation. This dissertation also developed human performance models corresponding with each evaluation using the tool, IMPRINT Pro.

The Guided scenario focused on modeling and evaluating workload in human-human and human-robot peer-based teams performing the same set of triage tasks (Harriott et al., 2011a,b, 2013). The Collaborative scenario focused on modeling and evaluating workload and task performance, specifically reaction time and response time, in human-human and human-robot teams performing a search task for hazardous items (Harriott et al., 2012b, 2013). The Guided and Collaborative evaluations provided preliminary data for investigating the applicability of potential workload and task performance metrics. Many metrics representing both workload and task performance were analyzed (e.g., heart rate variability, subjective workload ratings, secondary task response time). Overall, both evaluations indicated that workload was lower in human-robot teams, and while human-robot teams were able to accomplish task goals, they took significantly longer to do so. Human-robot teams also had slower reaction time and response time results.

The Time-Structured evaluation evaluated workload and task performance in four time-constrained tasks. The tasks focused on searching images for suspicious items, searching a hall for suspicious items, performing a hazard collection procedure for solid contaminants and performing a hazard collection procedure for liquid contaminants. The results indicated there was no difference in workload due to the assignment of partner (human or robot). This result confirmed that the lower workload seen in the human-robot teams in the

Guided and Collaborative evaluations were due to the longer time required to complete the task with the robot partner. This result motivates the need to develop robot peers with advanced speech processing capabilities. The psychological impact of working with a robot partner does not in and of itself change workload levels.

Each team type (i.e., human-human and human-robot) in each evaluation (i.e., Guided, Collaborative, and Time-Structured) was represented by a human performance model. The models provided workload estimates for the evaluations, and these results suggest that human performance modeling tools are able to predict workload level differences in human-robot teams. Human performance modeling tools are generally developed for domains that do not include human-robot peer-based teams. IMPRINT Pro, specifically, is designed for human-human and human system interaction with military scenarios, but was successfully adapted to human-robot peer-based teams in this dissertation.

The evaluation of workload and task performance metrics is an important combination of measurement techniques spanning fields, such as human-computer interaction, aviation, medicine, human factors, and robotics (Harriott and Adams, 2013). The metrics were presented with a structured analysis by definition, guidelines, specific examples based on their use in the Guided and Collaborative evaluations, limitations, and implications for future human-robot interaction. Chapter IV can be used as a reference guide for choosing relevant metrics for assessing workload and task performance in peer-based human-robot teams.

The results from the three evaluations are generalizable to other human-robot peer-based teaming situations. The trends seen in workload and task performance in both time-limited (i.e., the Time-Structured evaluation) and time-unlimited (i.e., the Guided and Collaborative evaluation) scenarios are usable to predict expectations for similar situations. Additionally, the experimental design developed to evaluate the human-human and human-robot conditions can be used in future peer-based human-robot interaction experimental designs. The given robot capabilities or assigned team tasks may be different depending on the specific scenario; however, this dissertation offered tools for assessing workload and task performance differences, assessments of differences in workload and task performance between peer-based human robot teams, and experimental designs for comparing team types for specific tasks.

Overall, this dissertation made novel contributions to the human-robot interaction field by evaluating workload and task performance differences between human-human and human-robot teams in the first response domain. The specific contributions are described in detail in Chapter VI.2, while the remaining future work is presented in Chapter VI.3.

### VI.2 Contributions

There are two primary contributions of this dissertation and two secondary contributions. The first contribution involves a unique set of metrics evaluated in non-laboratory settings for mobile, peer-based human-robot

teams in first response tasks. The associated secondary contribution includes the assessment of commercially available biometric sensors as an inexpensive alternative to scientifically validated physiological sensors, such as the Biopac Bioharness. The second primary contribution is the evaluation of differences in task performance and workload between human-human and human-robot teams. The associated secondary contribution is the development of human performance models using the tool IMPRINT Pro for both human-human and human-robot team tasks. The models were compared with experimental data and demonstrated the application of human performance modeling tools, designed only for human-human or human-system interaction, to human-robot peer-based teams.

First, this dissertation developed and evaluated a unique set of metrics for measuring workload and task performance in non-laboratory settings for human-robot peer-based teams. Such a distinct and varied set of metrics has not previously been identified and evaluated to measure workload and task performance in mobile peer-based human-robot interaction. The metrics were assessed for their various advantages and disadvantages in application to human-robot peer-based teams. Recommendations were also made for adapting the human-robot interaction to analyze specific metrics in real-time, on-board a robot, and for comparison with human performance model results.

An additional aspect of this primary contribution is the evaluation of two commercially available biometric sensors for assessing physiological information (i.e., heart rate and distance traveled), the Fitbit Zip, a Garmin Forerunner pedometer, and the Scosche Rhythm+. The results indicate that these commercial sensors show promise for capturing heart rate and movement data for mobile humans, such as those in the evaluated peer-based human-human and human-robot teams. The use of commercially available sensors offer inexpensive alternatives to scientifically validated sensors (e.g., Biopac Bioharness) for providing real-time biometric information.

Second, a primary contribution of the dissertation is the analysis of differences between human-human and human-robot teams in cognitive workload, physical workload, reaction time, and response time for teams working in non-laboratory settings. This type of evaluation and analysis, which directly and comprehensively compares workload and reaction time in human-human and human-robot teams performing the same set of tasks, is a novel contribution to the human-robot interaction field.

An additional secondary contribution is the development and analysis of three sets of human performance models of human-human and human-robot teams. The models used existing functions for human workload, reaction time, and task completion time to represent both human-human and human-robot teams. The model analysis evaluated the application of human performance modeling tools not validated for human-robot peer-based teams, to those specific team types. The models were analyzed for accuracy in representation of experienced workload in correspondence with experimental data. This information contributes to knowledge

informing accurate human performance models for peer-based human-robot teams in IMPRINT Pro, which has not yet been accomplished and serves as a basis for designing and developing future robot capabilities to predict a human partner's state.

### VI.3 Future Work

This dissertation leaves multiple avenues for future work. Future work includes topics, such as, metrics measurement on-board the robot, speech rate metrics of workload, robot performance adaptation based on a human's state, and investigating other aspects of human performance. Additionally, future work includes investigating alternate relationship types between humans and robots, expanding the size of teams, and generalizing research to other domains other than first response.

An area of future work involves developing on-board measurement of the presented metrics by the robot partner. Allowing the robot to directly measure human performance metrics is the first step toward a robot understanding their human partner's state and ability to adapt the robot's behaviors based on the perceived metric values. Additionally, by measuring task performance and workload on-board a robot, the research will move away from video coding for gathering metrics, such as reaction time measurement. This method of measurement will allow for faster and potentially more accurate analysis. Accomplishing this goal requires developing a code base for gathering information from appropriate sensors (e.g., Scosche Rhythm+, and future wearable devices) and determining associated workload or task performance values. For example, given the measured state of the human, is the human's workload or task performance lower or higher than the desired range for the specific task? This goal requires gathering the real-time state of the human on-board the robot and determining expected value ranges for each metric (e.g., heart rate variability, secondary task failure rate), that correspond to workload levels (e.g., low, medium, high), or task performance levels (e.g., low, medium, high), for a specific range of tasks. The robot can compare real-time results to the predicted results for a particular task. Finally, the robot can assess whether the workload or task performance level is expected based on current task load.

Speech-based metrics, such as speech rate or number of sentence fragments, are potential workload analysis metrics, but automatic detection of such metrics requires significant future work (Yin and Chen, 2007). Additional metrics requiring investigation include false starts (Berthold and Jameson, 1999), utterance content quality, number of syllables in an utterance (Müller et al., 2001), or hesitations (Jameson et al., 2010). Speech-based metrics were not analyzed in this dissertation, but hold promise for representing changes in workload during human-robot peer-based team tasks (see description of speech rate in Chapter IV.1.10). Humans and robots often interact using speech, and a real-time measurement of speech rate may offer insight into changing workload values. Current technology for speech recognition limited the analysis of speech rate

for this dissertation, but future work may leverage future technology with transcription or syllable recognition software.

Additionally, adapting robot behavior based on the current human state is a natural next step to follow the development of on-board metrics measurement. The robot will be able to change its actions in order to create optimized teamwork with its human partner. For example, if the human is struggling to complete multiple concurrent tasks and has low task performance levels and high workload levels, the robot may take on a task or delay the onset of an additional task. This capability is necessary to provide and extend the capabilities of robots as partners for humans in order to develop teams that exceed the capabilities of similar human-human teams.

Other aspects of human performance factors can be investigated, for example, fatigue, situation awareness, or arousal. Throughout the course of this dissertation over 500 human performance factors were reviewed, categorized, and considered for use in human-robot teams (Harriott et al., 2012c). The procedures for modeling, evaluating, and analyzing the metrics presented in this dissertation can also be applied to additional performance moderators by evaluating the differences between human-human and human-robot peer-based teams, modeling team tasks, as well as evaluating and analyzing additional performance factor metrics.

Alternate human-robot relationship types require investigation. Peer-based teams are simply one type of human-robot relationship. Scholtz (2003) outlines other types of relationships, such as the supervisor, bystander, and mechanic roles, which may impact the assignments of tasks, distribution of workload, and task performance. Additionally, it is necessary to investigate larger teams with varying compositions of robots and humans. This dissertation focused on teams of two, including a single ground-based mobile-robot. Additional team members, either humans or robots, will change the team dynamic. Investigating whether the results generalize to larger teams, or teams with different compositions (e.g., including unmanned aerial vehicles), and developing the associated models to include additional team members is a large avenue of potential research.

Finally, future work may include examining workload and task performance in human-robot teams, and the presented metrics for their measurement, for other deployment domains. Other domains, such as health care robotics, autism research, home robots, or autonomous automobile research have avenues for the exploration of task performance and workload evaluation using the presented set of metrics.

## BIBLIOGRAPHY


Aasman, J., Mulder, G., and Mulder, L. (1987). Operator effort and the measurement of heart-rate variability. *Human Factors*, 29(2):161–170.

Aasman, J., Wijers, A., Mulder, G., and Mulder, L. (1988). Mental fatigue in normal daily working routines. In Hancock, P. and Meshkati, N., editors, *Human Mental Workload*. Elsevier Science Publishers, North-Holland.

Abdulghani, A., Casson, A., and Rodriguez-Villegas, E. (2009). Quantifying the feasibility of compressive sensing in portable electroencephalography systems. In *Proceedings of the 5th International Conference on Foundations of Augmented Cognition, Neuroergonomics, and Operational Neuroscience*, pages 319–328.

Allender, L. (2000). Modeling human performance: Impacting system design, performance, and cost. In *Government and Aerospace Simulation Symposium, 2000 Advanced Simulation Technologies Conference*, pages 139–144.

Allender, L., Kelley, T. D., Salvi, L., Lockett, J., Headley, D. B., Promisel, D., Mitchell, D., Richer, C., and Feng, T. (1995). Verification, validation, and accredidation of a soldier-system modeling tool. In *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, pages 1219–1223.

Altmann, E. M. and Trafton, J. (2004). Task interruption: Resumption lag and the role of cues. In *In Proceedings of the 26th Annual Conference of the Cognitive Science Society*.

Anderson, J. and Lebiere, C. (1998). *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, NJ.

Archer, S., Gosakan, M., Shorter, P., and Lockett III., J. (2005). New capabilities of the Army's maintenance manpower modeling tool. *Journal of the International Test and Evaluation Association*, 26(1):19–26.

Arkin, R. and Moshkina, L. (2014). Affect in human-robot interaction. In Calvo, R., D'Mello, S., Gratch, J., and Kappas, A., editors, *The Oxford Handbook of Affective Computing*. Oxford Uniersity Press, Oxford.

Asendorpf, J. and Meier, G. (1993). Personality effects on children's speech in everyday life: Sociability-mediated exposure and shyness-mediated reactivity to social situations. *Journal of Personality and Social Psychology*, 64(6):1072–1083.

ASTM International (2010a). Standard guide for operational guidelines for initial response to a suspected biothreat agent. Technical Report ASTM E2770-10, American Society for Testing and Materials, West Conshohocken, PA.

ASTM International (2010b). Standard practices for bulk sample collection and swab sample collection of visible powders suspected of being biological agents from nonporous surfaces. Technical Report ASTM E2458-10, American Society for Testing and Materials, West Conshohocken, PA.

Bainbridge, W., Hart, J., Kim, E., and Scassellati, B. (2008). The effect of presence on human-robot interaction. In *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication*, pages 700–706.

Baines, T. S. and Benedetti, O. (2007). Modeling of human performance within manufacturing systems design: from theoretical towards practical framework. *Journal of Simulation*, 1(2):121–130.

Baldwin, C. L. (2012). *Auditory Cognition and Human Performance*. CRC Press, Boca Raton, FL.

Ball, J. T. (2004). A cognitively plausible model of language comprehension. In *Proceedings of the 30th Conference on Behavior Representation in Modeling Simulation*, pages 305–316.

Baron, S., Kruser, D., and B. Messick Huey (1990). *Quantitative modeling of human performance in complex, dynamic systems*. Academic Press, Washington DC.

BASIS (2015). The ultimate fitness and sleep tracker. http://www.mybasis.com/.

Benson, M., Koenig, K. L., and Schultz, C. (1996). Disaster triage: Start, then savea new method of dynamic triage for victims of a catastrophic earthquake. *Prehospital and Disaster Medicine*, 11(2):117 – 124.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V., Olmstead, R. E., Tremoulet, P. D., and Craven, P. L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, 78(Supplement 1):B231–B244.

Berthold, A. and Jameson, A. (1999). Interpreting symptoms of cognitive load in speech input. In *Proceedings of the 7th International Conference on User Modeling*, volume 407, pages 235–244. Springer, Vienna.

Best, B. J. and Lebiere, C. (2006). Cognitive agents interacting in real and virtual worlds. In Sun, R., editor, *Cognitive Modeling to Social Simulation*, pages 186–218. Cambridge University Press, New York, NY.

Biopac Systems, I. (2013). Complete systems for life science research and education. http://www.biopac.com/.

Boles, D. and Adair, L. (2001). Validity of the multiple resources questionnaire (MRQ). In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*, volume 45, pages 1795–1799.

Boles, D., Bursk, J., Phillips, J., and Perdelwitz, J. (2007). Predicting dual-task performance with the Multiple Research Questionnaire (MRQ). *Human Factors*, 49(1):32–45.

Booher, H. R. (2003). *Handbook of human systems integration*. John Wiley, Hoboken, NJ.

Bratman, M. (1992). Shared cooperative activity. *The Philosophical Review*, 101(2):327–341.

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, 42(3):167–175.

Breazeal, C. and Scassellati, B. (1999). How to build robots that make friends and influence people. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, pages 858–863.

Brenner, M., Doherty, E. T., and Shipp, T. (1994). Speech measures indicating workload demand. *Aviation, Space, and Environmental Medicine*, 65(1):21–26.

Briggs, G. and Scheutz, M. (2011). Facilitating mental modeling in collaborative human-robot interaction through adverbial cues. In *Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Brodbeck, V., Spinelli, L., Lascano, A., Wissmeier, M., Vargas, M.-L., Vuillemoz, S., Pollo, C., Schaller, K., Michel, C., and Seeck, M. (2011). Electroencephalographic source imaging: A prospective study of 152 operated epileptic patients. *Brain*, 134(10):2887–2897.

Brooks, R. (2002). Humanoid robots. *Communications of the ACM*, 45(3):33–38.

Brookshire, J., Singh, S., and Simmons, R. (2004). Preliminary results in sliding autonomy for coordinated teams. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 1, pages 706 – 711.

Bruemmer, D., Few, D., Boring, R., Marble, J., Walton, M., and Nielson, C. (2005). Shared understanding for collaborative control. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 35(4):494–504.

Card, S., Moran, T., and Newell, A. (1980). Computer text-editing: An information-processing analysis of a routine cognitive skill. *Cognitive Psychology*, 12(1):32–74.

Casper, J. and Murphy, R. (2003). Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 33(3):367–385.

Cassimatis, N. L. (2002). *Polyscheme: A cognitive architecture for integrating multiple representation and inference schemes*. PhD thesis, Massachusetts Institute of Technology.

Castor, M. (2003). *GARTEUR Handbook of mental workload measurement*. GARTEUR, Group for Aeronautical Research and Technology in Europe, Flight Mechanics Action Group, fm ag13: 164 edition.

Chadwick, R. (2006). Operating multiple semi-autonomous robots: Monitoring, responding, detecting. In *Proceedings of the Human Factors and Ergonomic Society 50th Annual Meeting*, pages 329–333.

Chao, C. and Thomaz, A. L. (2010). Turn taking for human-robot interaction. In Bohus, D., Horvitz, E., Kanda, T., Mutlu, B., and Raux, A., editors, *Proceedings of the AAAI Fall Symposium on Dialog with Robots*, number FS-10-05:2164, pages 132–134.

Chen, J., Haas, E., and Barnes, M. (2007). Human performance issues and user interface design for teleoperated robots. *IEEE Transactions on Systems, Man and Cybernetics - Part C*, 37(6):12311245.

Chernova, S. and Breazeal, C. (2010). Learning temporal plans from observation of human collaborative behavior. In Broz, F., Michalowski, M., and Mower, E., editors, *Proceedings of the AAAI Spring Symposia, Its All In the Timing: Representing and Reasoning About Time in Interactive Behavior*, number SS-10-06:1153, pages 7–12.

Christmansson, M. (1994). The HAMA-method: A new method for analysis of upper limb movements and risk for work-related musculoskeletal disorders. In *Proceedings of the 12th Triennial Congress of the International Ergonomics Association/Human Factors Association of Canada*, pages 173–175.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates, Hillsdale, NJ, 2nd edition.

Colin, T., Smets, N., Mioch, T., and Neerincx, M. (2014). Real time modeling of the cognitive load of an urban search and rescue robot operator. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 874–879.

Consenzo, K., Parasuraman, R., Novak, A., and Barnes, M. (2006). Implementation of automation for control of robotic systems. Technical Report ARL-TR-3808, United States Army Research Laboratory.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research*, 169:323–338.

Crandall, J. W. and Cummings, M. L. (2007). Identifying predictive metrics for supervisory control of multiple robots. *IEEE Transactions on Robotics*, 23(5):942–951.

Crandall, J. W., Cummings, M. L., and Nehme, C. E. (2009). A predictive model for human-unmanned vehicle systems. *AIAA Journal of Aerospace Computing Information and Communication*, 6(11):585–603.

Crandall, J. W., Cummings, M. L., Penna, M. D., and De Jong, P. M. A. (2011). Computing the effects of operator attention allocation in human control of multiple robots. *IEEE Transactions on Systems, Man, and Cybernetics - Part A*, 41(3):385–397.

Crandall, J. W., Nielsen, C. W., and Goodrich, M. A. (2003). Towards predicting robot team performance. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pages 906–911.

Cutlip, R., Hsiao, H., Garcia, R., Becker, E., and Mayeux, B. (2000). A comparison of different postures for scaffold end-frame disassembly. *Applied Ergonomics*, 31:507–513.

Czerwinski, M., Tan, D., and Robertson, G. (2002). Women take a wider view. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 195–202.

Dahn, D. and Belyavin, A. (1997). The integrated performance modeling environment: A tool for simulating human-system performance. In *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*, pages 1037–1041.

De Vos, M., Gandras, K., and Debener, S. (2014). Towards a truly mobile auditory braincomputer interface: Exploring the p300 to take away. *International Journal of Psychophysiology*, 91:46–53.

Deary, I. and Der, G. (2005). Reaction time, age, and cognitive ability: Longitudinal findings from age 16 to 63 years in representative population samples. *Aging, Neuropsychology and Cognition: A Journal on Normal and Dysfunctional Development*, 12(2):187–215.

Deutsch, S. (2006). UAV operator human performance models. Technical Report AFRL-HE-WP-TR-2006-0158, AFRL Human Effectiveness Directorate, Warfighter Interface Division.

Deutsch, S., Cramer, N., Keith, G., and Freeman, B. (1999). The distributed operator model architecture. Technical Report AFRL-HE-WP-TR-1999-0023, AFRL Human Effectiveness Directorate, Warfighter Interface Division, Wright-Patterson Base, OH.

Dewar, R., Ellis, J., and Mundy, G. (1976). Reaction time as an index of traffic sign reception. *Human Factors*, 18:381–392.

Dixon, S. and Wickens, C. (2003). Control of multiple UAVs: A workload analysis. In *Proceedings of the 12th International Symposium on Aviation Psychology*.

Domestic Preparedness (2007). Scott health & safety introduces the BioPak 240 revolution closed-circuit breathing apparatus. http://www.domesticpreparedness.com/Updates/Industry_Update/Scott_Health_%26_Safety_Introduces_the_BioPak_240_Revolution_Closed-Circuit_Breathing_Apparatus/.

Donders, F. (1868). On the speed of mental processes. *Acta Psychologica*, 30:412–431.

Donmez, B., Cummings, M., and Graham, H. (2009). Auditory decision aiding in supervisory control of multiple unmanned aerial vehicles. *Human Factors*, 51(5):718 – 729.

Draper, J. and Blair, L. (1996). Workload, flow, and telepresence during teleoperation. In *IEEE International Conference on Robotics and Automation*, pages 1030–1035.

Duffy, S. A. and Pisoni, D. B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*, 35(4):351–389.

Duncan, M. and Mummery, W. (2007). GIS or GPS? A comparison of two methods for assessing route taken during active transport. *American Journal of Preventive Medicine*, 33(1):51–53.

Dustman, R. and Beck, E. (1965). Phase of alpha brain waves, reaction time and visually evoked potentials. *Electroencephalography and Clinical Neurophysiology*, 18(5):433–440.

Eagle, D., Baunez, C., Hutcheson, D., Lehmann, O., Shah, A., and Robbins, T. (2008). Stop-signal reaction-time task performance: Role of prefrontal cortex and subthalamic nucleus. *Cerebral Cortex*, 18(1):178–188.

Eggemeier, F. T. and Wilson, G. F. (1991). Performance-based and subjective assessment of workload in multi-task environments. In Damons, D. L., editor, *Multiple-Task Performance*, pages 217–278. Taylor & Francis, Bristol, PA.

Elara, M. R., Calderon, C. A. A., Zhou, C., and Wijesoma, W. S. (2010). On the redefinition of fan out metric for human robot interactions with humanoid soccer robots. *International Journal of Humanoid Robots*, 7(4):565–586.

Finomore, V., Shaw, T., Warm, J., Matthews, G., Weldon, D., and Boles, D. (2009). On the workload of vigilance: comparison of the NASA-TLX and the MRQ. In *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting*, volume 53, pages 1057–1061.

Finomore, V., Warm, J., Matthes, G., Riley, M., Dember, W., Shaw, T., Ungar, N., and Scerbo, M. (2006). Measuring the workload of sustained attention. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*, volume 50, pages 1614–1618.

Fitbit, I. (2013). Fitbit wireless activity tracker. http://www.fitbit.com/.

Folds, D. and Gerth, J. (1994). Auditory monitoring of up to eight simultaneous sources. In *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, pages 505–509.

Ford, S., Birmingham, E., King, A., Lim, J., and Ansermino, J. (2010). At-a-glance monitoring: Covert observations of anesthesiologists in the operating room. *Anesthesia & Analgesia*, 111(3):653–658.

Foyle, D. and Hooey, B. (2008). *Human Performance Modeling in Aviation*. CRC / Taylor & Francis, Boca Raton, Florida, USA, 1st edition.

Friedman, B., Kahn Jr., P., Hagman, J., Severson, R., and Gill, B. (2006). The watcher and the watched: social judgments about privacy in a public place. *Human-Computer Interaction*, 21:235–272.

Gais, S., Lucas, B., and Born, J. (2006). Sleep after learning aids memory recall. *Learning and Memory*, 13:259–262.

Garet, M., Boudet, G., Montaurier, C., Vermorel, M., Coudert, J., and Chamoux, A. (2004). Estimating relative physical workload using heart rate monitoring: a validation by whole-body indirect calorimetry. *European Journal of Applied Physiology*, 94(1-2):46–53.

Gawron, V. (2008). *Human performance, workload and situational awareness*. CRC Press, London.

Gluck, K. A., Ball, J. T., Gunzelmann, G., Krusmark, M. A., Lyon, D. R., and Cooke, N. J. (2005). A prospective look at a synthetic teammate for UAV applications. In *Proceedings of the American Institute of Aeronautics and Astronautics InfotechAerospace Conference*, pages 1–13.

Goodrich, M. and Boer, E. (2003). Model-based human-centered task automation: A case study in ACC design. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 33(3):325–336.

Goodrich, M., McLain, T., Anderson, J., Sun, J., and Crandall, J. (2007). Managing autonomy in robot teams: observations from four experiments. In *Proceedings of ACM/IEEE International Conference on Human-Robot Interaction*, pages 25–32.

Goodrich, M. and Olsen, D. (2003). Seven principles of efficient human robot interaction. In *Proceedings of IEEE Conference on Systems, Man and Cybernetics*, pages 3943–3948.

Goodrich, M. and Schultz, A. (2007). Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275.

Gramann, K., Gwin, J., Ferris, D., Oie, K., Jung, T., Lin, C., Liao, L., and Makeig, S. (2011). Cognition in action: Imaging brain/body dynamic in mobile humans. *Reviews in the Neurosciences*, 22(6):593–608.

Groeneveld, R. and Meeden, G. (1984). Measuring skewness and kurtosis. *The Statistician*, 33:391–399.

Haddadin, S., Albu-Schäffer, A., and Hirzinger, G. (2007a). Safety evaluation of physical human-robot interaction via crash-testing. In *International Symposium on Robotics Research*, pages 439–450.

Haddadin, S., Albu-Schäffer, A., and Hirzinger, G. (2007b). Safety evaluation of physical human-robot interaction via crash-testing. In *Proceedings of the Robotics Science and Systems Conference*, pages 217–224.

Hansson, G. Å., Balogh, I., Ohlsson, K., Granqvist, L., Nordander, C., Arvidsson, I., Akesson, I., Unge, J., Rittner, R., Stromberg, U., and Skerfving, S. (2009). Physical workload in various types of work: Part I: Wrist and forearm. *International Journal of Industrial Ergonomics*, 39(1):221–233.

Hansson, G. Å., Balogh, I., Ohlsson, K., Granqvist, L., Nordander, C., Arvidsson, I., Akesson, I., Unge, J., Rittner, R., Stromberg, U., and Skerfving, S. (2010). Physical workload in various types of work: Part II: Neck, shoulder, and upper arm. *International Journal of Industrial Ergonomics*, 40(3):267–281.

Harriott, C. and Adams, J. A. (2013). Modeling human performance for human-robot systems. *Reviews of Human Factors and Ergonomics*, 9(1):94–130.

Harriott, C., Buford, G. L., Zhang, T., and Adams, J. (2012a). Assessing workload in human-robot peer-based teams. In *Proceedings of 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 141–142, Boston, Massachusetts.

Harriott, C., Buford, G. L., Zhang, T., and Adams, J. (2012b). Human-human vs. human-robot teamed investigation. In *Proceedings of 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, page 405, Boston, Massachusetts. Video.

Harriott, C., Zhang, T., and Adams, J. (2011a). Applying workload human performance moderator functions to peer-based human-robot teams. In *Proceedings of 6th ACM/IEEE International Conference on Human-Robot Interaction*, pages 45–52, Lausanne, Switzerland.

Harriott, C., Zhang, T., and Adams, J. (2011b). Predicting and validating workload in human-robot teams. In *Proceedings of 20th Conference on Behavioral Representation in Modeling Simulation*, pages 162–169, Sundance, Utah.

Harriott, C., Zhang, T., and Adams, J. A. (2013). Assessing physical workload for human-robot peer-based teams. *International Journal of Human-Computer Studies*, 71:821–837.

Harriott, C., Zhuang, R., Adams, J., and DeLoach, S. A. (2012c). Towards using human performance moderator functions in human-robot teams. In *Proceedings of International Workshop on Human-Agent Interaction Design and Models (HAIDM) of International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Valencia, Spain.

Hart, S. and Staveland, L. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Hancock, P. and Meshkati, N., editors, *Human Mental Workload*, pages 139–183. North Holland Press, Amsterdam.

Helander, M., Karwan, M., and Etherton, J. (1987). A model of human reaction to dangerous robot arm movements. In *Proceedings of the Human Factors and Ergonomics Society 31st Annual Meeting*, pages 191–195.

Hervey, A., Epstein, J., Curry, J., Tonev, S., Arnold, L., Hinshaw, S., Swanson, J., and Hechtman, L. (2006). Reaction time distribution analysis of neuropsychological performance in an adhd sample. *Child Neuropsychology*, 12(2):125–140.

Hick, W. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4:11–26.

Hinds, P., Roberts, T., and Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human-Computer Interaction*, 19:151–181.

Hoffman, G. and Breazeal, C. (2004). Collaboration in human-robot teams. In *Proceedings of the AIAA 1st Intelligent Systems Technical Conference*, pages 1–18.

Hohle, R. (1967). Component process latencies in reaction times of children and adults. In Lipsett, L. and Spiker, C., editors, *Advances in Child Development and Behavior*, volume 3, pages 225–261. Academic Press, New York.

Howard, A. and Paul, W. (2005). A 3d virtual environment for exploratory learning in mobile robot control. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 306–310.

Howard, A. M. (2007). A systematic approach to predict performance of human-automation systems. *IEEE Transactions on Systems, Man, and Cybernetics - Part C*, 37(4):594–601.

Humphrey, C. and Adams, J. (2009). Robotic tasks for CBRNE incident response. advanced robotics. *Advanced Robotics*, 23(9):1217–1232.

Humphrey, C. and Adams, J. (2011). Analysis of complex team-based systems:Augmentations to goal-directed task analysis and cognitive work analysis. *Theoretical Issues in Ergonomics Science*, 12(2):149–175.

Hunn, B. P. and Heuckeroth, O. H. (2006). A shadow unmanned aerial vehicle UAV improved performance research integration tool (IMPRINT) model to support future combat systems. Technical Report ARL-TR-3731, Army Research Laboratory.

Hyman, R. (1953). Stimulus information as a detriment of reaction time. *Journal of Experimental Psychology*, 45:423–432.

iSpeech, I. (2013). Free human quality text to speech and speech recognition. http://www.ispeech.org/.

James, L. and Brett, J. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, 69(2):307–321.

Jameson, A., Kiefer, J., Müller, C., Großmann-Hutter, B., Wittig, F., and Rummer, R. (2010). Assessment of a user's time pressure and cognitive load on the basis of features of speech. In *Resource-Adaptive Cognitive Processes*, pages 171–204. springer, Berlin Heidelberg.

Jastrow, J. (1890). *The time-relations of mental phenomena. Fact and theory papers No. VI.* N.D.C. Hodges, New York.

Jawbone (2013). Up. know yourself live better. https://jawbone.com/up.

Jevas, S. and Yan, J. (2001). The effect of aging on cognitive function: A preliminary quantitative review. *Research Quarterly for Exercise and Sport*, 72:A–49.

Ji, Q., Zhu, Z., and Lan, P. (2004). Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE Transactions on Vehicular Technology*, 53(4):1052–1068.

Johansson, G. and Rumar, K. (1971). Drivers' brake reaction times. *Human Factors*, 13(1):23–27.

John, B. E. and Newell, A. (1989). Cumulating the science of HCI: From s-R compatibility to transcription typing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 109–114.

Johnstone, J., Ford, P., Hughes, G., Watson, T., and Garrett, A. (2012). BioHarness multivariable monitoring device: Part. I: Validity. *Journal of Sports Science & Medicine*, 11(3):400–408.

Jung, Y., Choi, Y., Park, H., Shin, W., and Myaeng, S. (2007). Integrating robot task scripts with a cognitive architecture for cognitive human-robot interactions. In *IEEE International Conference on Information Reuse and Integration*, pages 152–157.

Kaber, D., Onal, E., and Endsley, M. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness and subjective workload. *Human Factors and Ergonomics in Manufacturing*, 10(4):409–430.

Kaber, D. B., Wang, X., and Kim, S.-H. (2006). Computational cognitive modeling of operator behavior in telerover navigation. In *Proceedings of the 2006 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3210–3215.

Karhu, O., Kansi, P., and Kuorinka, I. (1977). Correcting working postures in industry: A practical method for analysis. *Applied Ergonomics*, 8(4):199–201.

Katzenbach, J. and Smith, D. (1993). The discipline of teams. *Harvard Business Review*, 71:111–120.

Keller, J., Bless, H., Blomann, F., and Kleinbohl, D. (2001). Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary control. *Journal of Experimental Social Psychology*, 47(4):849–852.

Kessel, C. J. and Wickens, C. D. (1982). The transfer of failure-detection skills between monitoring and controlling dynamic systems. *Human Factors*, 24(1):49–60.

Klein, M., Lio, C., Grant, R., Carswell, C., and Strup, S. (2009). A mental workload study on the 2D and 3D viewing conditions of the Da Vinci surgical robot. In *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting*, volume 53, pages 1186–1190.

Kramer, A. (1990). Physiological metrics of mental workload: A review of recent progress. Technical Report AD-A223 701, Navy Personnel Research and Development Center.

Kunlun, S., Yan, L., and Ming, X. (2011). A safety approach to predict human error in critical flight tasks. In *The 2nd International Symposium on Aircraft Airworthiness*, volume 17, pages 52–62.

Land, M. and McLeod, P. (2000). From eye movements to actions: How batsmen hit the ball. *Nature Neuroscience*, 3(12):1340–1345.

Lasley, D., Hamer, R., Dister, R., and Cohn, T. (1991). Postural stability and stereo-ambiguity in man-designed visual environments. *IEEE Transactions on Biomedical Engineering*, 38(8):808–813.

Lee, J. (2008). Fifty years of driving safety research. *Human Factors*, 50(3):521–528.

Lee, J., Caven, B., Haake, S., and Brown, T. (2001). Speech-based interaction with in-vehicle computers: The effect of speech-based e-mail on drivers' attention to the roadway. *Journal of the Human Factors and Ergonomics Society*, 43(4):631–640.

Leedal, J. and Smith, A. (2005). Methodological approaches to anaesthetists' workload in the operating theatre. *British journal of anaesthesia*, 94(6):702–709.

Lei, S. and Roetting, M. (2011). Influence of task combination on eeg spectrum modulation for driver workload estimation. *Human Factors*, 53(2):168–179.

Li, G. and Buckle, P. (1999). Current techniqes for assessing physical exposture to work-related muscoskeletal risks, with emphasis on posture-based methods. *Ergonomics*, 42(5):674–695.

Liang, Y., Reyes, M., and Lee, J. (2007). Real-time detection of driver cognitive distraction using support vector machines. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):340–350.

Liu, C., Rani, P., and Sarkar, N. (2006). Human-robot interaction using affective cues. In *15th IEEE International Symposium on Robot and Human Interactive Communication*, pages 285–290.

Liu, C., Rani, P., Sarkar, N., and Chen, S. (2009). Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. *International Journal of Human-Computer Interaction*, 25(6):506–529.

Lo, C.-C. and Wang, X.-J. (2006). Corticobasal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature Neuroscience*, 9(7):956–963.

Luce, R. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press, New York.

Macflynn, G., Montgomery, E., Fenton, G., and Rutherford, W. (1984). Measurement of reaction time following minor head injury. *Journal of Neurology, Neurosurgery and Psychiatry*, 47:1326–1331.

Mahoney, P. (1994). Businesses and bombs: Preplanning and response. *Facilities*, 12(10):14–21.

Maltz, M. and Shinar, D. (1999). Eye movements of older and younger drivers. *Human Factors*, 41(1).

McAtamney, L. and Corlett, E. (1993). RULA: A survey method for the investigation of work-related upper limb disorders. *Applied Ergonomics*, 24(2):91–99.

McCracken, J. and Aldrich, T. (1984). Analyses of selected lhx mission functions: Implications for operator workload and system automation goals. Technical Report ASI479-024-84, Anacapa Sciences, Inc., Fort Rucker, AL.

Medvec, B. (1994). Productivity and workload measurement in ambulatory oncology. *Seminars in Oncology Nursing*, 10(4):288–295.

Mehler, B., Reimer, B., Coughlin, J., and Dusek, J. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record*, 2139:6–12.

Mio, G. (2014). Mio. train with heart. http://www.mioglobal.com/.

Moray, N. (1982). Subjective mental workload. *Human Factors*, 24(1):25–40.

Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., and Wittig, F. (2001). Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In *User Modeling 2001*, pages 24–33. Springer, Berlin Heidelberg.

Murphy, R. (2004). Human-robot interaction in rescue robotics. *IEEE Transactions on Systems, Man, and Cybernetics - Part C*, 34(2):138–153.

Murphy, R. and Schreckenghost, D. (2013). Survey of metrics for human-robot interaction. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot interaction*, pages 197–198.

Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., and Hagita, N. (2009). Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction*, pages 61–68.

Nachreiner, F. (1995). Standards for ergonomics principles relating to the design of work systems and to mental workload. *Applied Ergonomics*, 26(4):259–263.

Nagler, C. and Nagler, W. (1973). Reaction time measurements. *Journal of Experimental Psychology*, 2:261–274.

Neto, O., Pacheco, M., Bolander, R., and Bir, C. (2009). Force, reaction time, and precision of Kung Fu strikes. *Perceptual and Motor Skills*, 109(1):295–303.

Nike, I. (2015). Nike+fuelband. http://www.nike.com/fuelband.

Niogi, S., Mukherjee, P., Ghajar, J., Johnson, C., Kolster, R., Sarkar, R., Lee, H., Meeker, M., Zimmerman, R., Manley, G., and McCandliss, B. (2008). Extent of microstructural white matter injury in postconcussive syndrome correlates with impaired cognitive reaction time: A 3T diffusion tensor imaging study of mild traumatic brain injury. *American Journal of Neuroradiology*, 34(4):967–973.

Noble, C., Baker, B., and Jones, T. (1964). Age and sex parameters in psychomotor learning. *Perceptual and Motor Skills*, 19:935–945.

Norman, D. and Bobrow, D. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, 7:44–64.

Novak, D., Beyeler, B., Omlin, X., and Riener, R. (2014). Workload estimation in physical human-robot interaction using physiological measurements. *Interacting with Computers*.

Nuance (2014). Dragon speech recognition software. http://www.nuance.com/dragon/index.htm.

Nygaard, L. C. and Pisoni, D. B. (1995). Speech perception: new directions in research and theory. In Miller, J. L. and Eimas, P. D., editors, *Speech, language, and communication*, pages 63–96. Academic Press, San Diego, CA, 2nd edition.

Olsen Jr., D. R. and Wood, D. B. (2004). Fan-out: measuring human control of multiple robots. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 231–238.

Paul, P., Kuijer, M., Visser, B., and Kemper, H. (1999). Job rotation as a factor in reducing physical workload at a refuse collecting department. *Ergonomics*, 43(9):1167–1178.

Petkosek, M., Warfield, L., and Caretta, T. R. (2005). Development of a human performance model of a UAV sensor operator: lessons learned. Technical Report AFRL-HE-WP-TR-2005-0118, Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter interface division, Wright-Patterson, OH.

Pew, R. W. and Mavor, A. S. (2007). *Human-system integration in the system development process: a new look*. National Academic Press, Washington, DC.

Pew, R. W. and Mavor, A. S. (2008). *Modeling human and organizational behavior: applications to military simulations*. National Academic Press, Washington, DC.

Poh, M., Kim, K., Goessling, A., Swenson, N., and Picard, R. (2011a). Cardiovascular monitoring using earphones and a mobile device. *IEEE Pervasive Computing*, 99:1–13.

Poh, M., McDuff, D., and Picard, R. (2011b). Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on biomedical engineering*, 58(1):7–11.

Porges, S. (1972). Heart rate variability and deceleration as indexes of reaction time. *Journal of Experimental Psychology*, 92(1):103–110.

Prewett, M., Saboe, K., Johnson, R., Coovert, M., and Elliott, L. (2009). Workload in human-robot interaction: A review of manipulations and outcomes. In *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting*, pages 1393–1397.

Raby, M. and Wickens, C. (1994). Strategic workload management and decision biases in aviation. *The International Journal of Aviation Psychology*, 4(3):211–240.

Rani, P., Sarkar, N., and Adams, J. A. (2007). Anxiety-based affective communication for implicit human-machine interaction. *Advanced Engineering Informatics*, 21(3):323–334.

Rani, P., Sarkar, N., Smith, C. A., and Kirby, L. D. (2004). Anxiety detecting robotic system - towards implicit human-robot collaboration. *Robotica*, 22(1):85–95.

Rani, P., Sims, J., Brackin, R., and Sarkar, N. (2002). Online stress detection using psychophysiological signal for implicit human-robot cooperation. *Robotica*, 20(6):673–686.

Reimer, B., Mehler, B., Coughlin, J., Godfrey, K. M., and Tan, C. (2009). An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 115–118.

Reiser, L. and Schlenk, E. (2009). Clinical use of physical activity measures. *Journal of the American Academy of Nurse Practitioners*, 21(2):87–94.

Reitter, D. and Lebiere, C. (2010). A cognitive model of spatial path-planning. *Computational and Mathematical Organization Science*, 16:220–245.

Riecke, B., Bodenheimer, B., McNamara, T., Williams, B., Peng, P., and Feuereissen, D. (2010). Do we need to walk for effective virtual reality navigation? Physical rotations alone may suffice. In Hölscher, C., Shipley, T., Olivetti Belardinelli, M., Bateman, J., and Newcombe, N., editors, *Spatial Cognition VII (Vol. 6222)*. Springer, Berlin / Heidelberg.

Riggs, A., Melloy, B., and Neyens, D. M. (2014). The effect of navigational tools and related experience on task performance in a virtual environment. In *Proceedings of the Human Factors and Ergonomics Society 58th Annual Meeting*, pages 2378–2382.

Ritter, F. E., Kukreja, U., and St. Amant, R. (2007). Including a model of visual processing with a cognitive architecture to model a simple teleoperation task. *Journal of Cognitive Engineering and Decision Making*, 1(2):121–147.

Ritter, F. E., Van Rooy, D., St. Amant, R., and Simpson, K. (2006). Providing user models direct access to interfaces: An exploratory study of a simple interface with implications for HRI and HCI. *IEEE Transactions on Systems, Man, and Cybernetics - Part A*, 36(3):592–601.

Roberts, S. and Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2):358–367.

Rooti Labs, L. (2014). Phyode w/me: Know your inner self. http://www.phyode.com/.

Roscoe, A. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological Psychology*, 34:259–287.

Ruff, H., Narayanan, S., and Draper, M. (2002). Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned air vehicles. *Presence*, 11(4):335–351.

Ryu, H., Park, E., Park, Y., Han, K., and Lim, J. (2003). A workload analysis of a visiting nursing service based on a health center in Soeul. *Journal of Korean Academic Nursing*, 33(7):1018–1027.

Salvucci, D. D. and Taatgen, N. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review*, 115(5):101–130.

San Agustin, J., Skovsgaard, H., Mollenbach, E., Barret, M., Tall, M., Hansen, D. W., and Hansen, J. P. (2010). Evaluation of a low-cost open-source gaze tracker. In *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications*, pages 77–80.

Scheutz, M., Schermerhorn, P., Kramer, J., and Anderson, D. (2007). First steps toward natural human-like hri. *Autonomous Robots*, 22(4):411–423.

Schneider, E., Villgrattner, T., Vockeroth, J., Bartl, K., Kohlbecher, S., Bardins, S., Ulbrich, H., and Brandt, T. (2009). EyeSeeCam: An eye movementdriven head camera for the examination of natural visual exploration. *Annals of the New York Academy of Sciences*, 1164(1):461–467.

Scholtz, J. (2003). Theory and evolution of human robot interactions. In *IEEE 36th International Conference on System Sciences*, pages 125–134.

Schreiber, B. T., Lyon, D. R., Martin, E. L., and Confer, H. A. (2002). Impact of prior flight experience on learning Predator UAV operator skills. Technical Report AFRL-HE-AZ-TR-2002-0026, Air force research laboratory, warfighter training research division, Mesa, AZ.

Scosche Industries (2014). Best Heart Rate Monitor — Rhythm Plus. http://www.scosche.com/rhythm-plus.

Sharek, D. (2009). NASA-TLX online tool. http://www.nasatlx.com/.

Sheridan, T. (1992). *Telerobotics, Automation, and Human Supervisory Control*. MIT Press, Cambridge, MA.

Sidner, C., Kidd, C. D., Lee, C., and Lesh, N. (2004). Where to look: A study of human-robot engagement. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, pages 78–84.

Sidner, C., Lee, C., Morency, L. P., and Forlines, C. (2006). The effect of head-nod recognition in human-robot conversation. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, pages 290–296.

Silverman, B., Johns, M., Cornwell, J., and O'Brien, K. (2006). Human behavior models for agents in simulators and games: Part I: Enabling science with PMFServ. *Presence: Teleoperators and Virtual Environments*, 15(2):139–162.

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138.

Singh, P. (2005). *EM-ONE: An architecture for reflective commonsense thinking*. PhD thesis, Massachusetts Institute of Technology.

Sjorgren, P. and Banning, A. (1989). Pain, sedation and reaction time during long-term treatment of cancer patients with oral and epidural opiods. *Pain*, pages 5–11.

Skantze, G. and Schlangen, D. (2009). Incremental dialogue processing in a micro-domain. In *Proceedings of 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 745–753.

Spinks, J. A., Zhang, J. X., Fox, P. T., Gao, J. H., and Hai Tan, L. (2004). More workload on the central executive of working memory, less attention capture by novel visual distractors: evidence from an fMRI study. *Neuroimage*, 23(2):517–524.

Squire, P., Trafton, G., and Parasuraman, R. (2006). Human control of multiple unmanned vehicles: Effects of interface type on execution and task switching times. In *Proceedings of the 1st Annual ACM SIGCHI/SIGART Conference on Human-robot Interaction*, pages 26–32.

St. Amant, R., Horton, T. E., and Ritter, F. E. (2004). Model-based evaluation of expert cell phone menu interaction. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, pages 343–350.

St. Clair, A. and Matarić, M. (2011). Task coordination and assistive opportunity detection via social interaction in collaborative human-robot tasks. In *Proceedings of the 2011 International Conference on Collaboration Technologies and Systems Workshop on Collaborative Robots and Human Robot Interaction*, pages 168–172.

Steele, B., Holt, L., Belza, B., Ferris, S., Lakshminaryan, S., and Buchner, D. M. (2000). Quantitating physical activity in COPD using a triaxial accelerometer. *CHEST Journal*, 117(5):1359–1367.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., and Goodrich, M. (2006). Common metrics for human-robot interaction. In *Proceedings of the 1st Annual ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, pages 33–40.

Sternberg, S. (1969). Mental processes revealed by reaction time experiments. *American Scientist*, 57(4):421–457.

Strurm, J. A. and Seery, C. H. (2007). Speech and articulatory rates of school-age children in conversation and narrative contexts. *Language, Speech, and Hearing Services in Schools*, 38(1):47–59.

Stubbs, K., Hinds, P., and Wettergreen, D. (2007). Autonomy and common ground in human-robot interaction: A field study. *IEEE Intelligent Systems*, 22(2):42–50.

Taylor, J. L., Yesavage, J. A., Morrow, D. G., and Dolhert, N. (1994). The effects of information load and speech rate on younger and older aircraft pilots' ability to execute simulated air-traffic controller instructions. *Journals of Gerontology*, 49(5):P191–P200.

Thrun, S. (2004). Toward a framework for human-robot interaction. *Human-Computer Interaction*, 29(1-2):9–24.

Tiberio, L., Cesta, A., and Olivetti Belardinelli, M. (2013). Phychophysiological methods to evaluate user's response in human robot interaction: A review and feasibility study. *Robotics*, 2:92–121.

Tobii Technology (2013). Tobii Glasses 2 - Mobile eye tracking. http://www.tobii.com.

Tomaka, J., Blascovich, J., Kibler, J., and Ernst, J. M. (1997). Cognitive and physiological antecedents of threat and challenge appraisal. *Journal of Personality and Social Psychology*, 73(1):63.

Topal, C., Dogan, A., and Gerek, O. (2008). A wearable head-mounted sensor-based apparatus for eye tracking applications. In *IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems*, pages 136–139.

Trafton, J., Altmann, E., Brock, D., and Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies*, 58(5):583–603.

Trafton, J. and Monk, C. (2007). Task interruptions. *Reviews of Human Factors and Ergonomics*, 3(1):111–126.

Trafton, J. G., Bugajska, M. D., Fransen, B. R., and Ratwani, R. M. (2008). Integrating vision and audition within a cognitive architecture to track conversations. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 201–208.

Trafton, J. G., Cassimatis, N. L., Bugajska, M. D., Brock, D. P., Mintz, F. E., and Schultz, A. C. (2005). Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man, and Cybernetics - Part A*, 35(4):460–470.

Tyler, S., Neukom, C., Logan, M., and Shively, J. (1998). The MIDAS human performance model. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, pages 320–325.

United States Army Research Laboratory (2009). *Improved performance research integration tool Pro (IMPRINT Pro) User's Guide*. Human Research and Engineering Directorate, 3.0 edition.

Van der Schatte Olivier, R., vant Hullenaar, C., Ruurda, J., and Broeders, I. (2009). Ergonomics, user comfort, and performance in standard and robot-assisted laparoscopic surgery. *Surgical Endoscopy*, 23(6):1365–1371.

Veltman, J. and Gaillard, A. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, 41(5):656–669.

Venkatagiri, H. S. (1999). Clinical measurement of rate of reading and discourse in young adults. *Journal of Fluency Disorders*, 24(3):209–226.

Vicente, K., Thornton, D., and Moray, N. (1987). Spectral analysis of sinus arrythmia: A measure of mental effort. *Human Factors*, 29(2):171–182.

Vicente, K. J. (1992). Memory recall in a process control system: a measure of expertise and display effectiveness. *Memory and Cognition*, 20(4):356–373.

Vicente, K. J. and Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review*, 105(1):33–57.

Vogler, D. (2001). Raw video footage / wtc 9.11.01. http://davidvogler.com/911.

Weinger, M., Herndon, O., Zornow, M., Paulus, M., Gaba, D., and Dallen, L. (1994). An objective methodology for task analysis and workload assessment in anesthesia providers. *Anesthesiology*, 80(1):77–92.

Weinger, M., Reddy, S., and Slagle, J. (2004). Multiple measures of anesthesia workload during teaching and nonteaching cases. *Anesthesia & Analgesia*, 98(5):1419–1425.

Welford, A. (1977). Motor performance. In Birren, J. and Shaie, K., editors, *Handbook of the Psychology of Aging*, pages 450–496. Van Norstrand Reinhold, New York.

Welford, A. (1980). Choice reaction time: Basic concepts. In *Reaction Times*, pages 73–128. Academic Press, New York.

Wickens, C. (2002). Multiple resources and performance prediction. *Theories in Ergonomic Science*, 3(2):159–177.

Wickens, C., Lee, J., Liu, Y., and Gordon-Becker, S. (2003). *An Introduction to Human Factors Engineering*. Prentice Hall, Upper Saddle River, NJ, 2nd edition.

Wickens, C. D. and Andre, A. D. (1990). Proximity compatibility and information display: Effects of color, space, and objectness on information integration. *Human Factors*, 32(1):61–77.

Wickens, C. D. and Liu, Y. (1988). Codes and modalities in multiple resources: A success and a qualification. *Human Factors*, 30(5):599–616.

Wierwille, W., Rahimi, M., and Casali, J. (1985). Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity. *Human Factors*, 27(5):489–502.

Wierwille, W. W., Gutmann, J. C., Hicks, T. G., and Muto, W. H. (1977). Secondary task measurement of workload as a function of simulated vehicle dynamics and driving conditions. *Human Factors*, 19(6):557–565.

Williges, R. C. and Wierwille, W. W. (1979). Behavioral measures of aircrew workload. *Human Factors*, 21(5):549–574.

Wilson, G. (1993). Air-to-ground training missions: A psychopsysiological worklaod analysis. *Ergonomics*, 36(9):1071–1087.

Wilson, G. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *International Journal of Aviation Psychology*, 12(1).

Wilson, G., Purvis, B., Skelly, J., Fullenkamp, P., and Davis, I. (1987). Physiological data used to measure pilot workload in actual flight and simulator conditions. In *Proceedings of the Human Factors and Ergonomics Society 31st Annual Meeting*, volume 31, pages 779–783.

Witmer, B., Bailey, J., Knerr, B., and Parsons, K. (1996). Virtual spaces and real world places: Transfer of route knowledge. *International Journal of Human-Computer Studies*, 45(4):413–428.

Wu, H.-C. and Wang, M. (2002). Relationship between maximum acceptable work time and physical workload. *Ergonomics*, 45(4):280–289.

Yeh, Y. and Wickens, C. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, 30(1):111–120.

Yin, B. and Chen, F. (2007). Towards automatic cognitive load measurement from speech analysis. In *Human-Computer Interaction, Interaction Design and Usability*, pages 1011–1020. Springer, Berlin Heidelberg.

YouEye (2013). Agile user testing, webcam eye tracking, and emotion recognition. http://www.youeye.com.