Mental Workload and Task Performance in Peer-Based Human-Robot Teams

Caroline E. Harriott
Charles Stark Draper Laboratory
Glenna L. Buford and Julie A. Adams
Vanderbilt University
Tao Zhang
Purdue University

Successful human-robot (H-R) teams working with direct interaction and close coupling will have relationships that vary, depending on whether or not collaboration is prioritized and how collaboration changes the human mental workload and the team’s task performance. Modeling of representative functions can provide predictions of changes in mental workload and the potential impact on team performance. The presented research focuses on modeling and quantifying mental workload for H-R teams in which team members have some individually assigned responsibilities but must also make joint decisions with a teammate. IMPRINT Pro, a discrete event simulation modeling tool created by the U.S. Army Research Laboratory, was used to model human-human (H-H) and H-R teams completing a reconnaissance task in a building. This research evaluated H-H and H-R teams completing the same reconnaissance tasks. Predictions of mental-workload levels from the model and the evaluation results showed that mental workload was lower for the H-R teams. The results for the closely coupled teams were compared to results from a prior evaluation with a master-slave relationship; similar results were found for both evaluations. Mental workload was lower in H-R teams than in H-H teams, but task performance did not differ between the two.

Keywords: Mental workload, task performance, peer-based human-robot teams, human performance modeling, physiological metrics, subjective metrics

1. Introduction

Robots are becoming more integrated into everyday life (e.g., industry, civil servant tasks, homes). Time-critical domains, such as emergency response (Humphrey & Adams, 2009; Murphy, 2014) and military deployments (Freedy, DeVisser, Weltman, & Coeyman, 2007; Hutchins, Cosenzo, Barnes, Feng, & Pillalamarri, 2010), will incorporate human-robot (HR) teams into missions, and robots will act as peer team members (Scholtz, 2003). As this transition occurs, robot team members will interact and collaborate with humans (Hoffman & Breazeal, 2004; Kaupp, Makenenko, & Durrant-Whyte, 2010). They will share joint tasks and decisions and understand each other’s progress toward shared goals. They may need to redistribute tasks and goals as human...
team members’ mental workload fluctuates, as happens in human-only teams. Modeling and verification of such situations are a step toward real-time prediction of, and adaptation to, fluctuating human workload or other human performance attributes in collaborative H-R teams.

The development of accurate human performance predictions is a key to developing future robots that adapt their behavior and interaction with their human team members. Human performance moderator functions (Silverman, Johns, Cornwell, & O’Brien, 2006) can predict human performance, both prior to team subtask allocation and during task execution. Human-performance functions are derived from experimental data and provide predictions of human performance under a specific set of conditions (e.g., a military unit of three people flying an unmanned aerial vehicle in extremely hot weather conditions). Existing human performance functions have been developed for domains, such as driving (Salvucci, 2001), power plant operation (Mumaw, Roth, Vicente, & Burns, 2000), military applications (Weaver, Silverman, Shin, & Dubois, 2001), etc.

This research focuses on modeling mental workload, but other aspects of human performance can be modeled for use in creating adaptable robot behaviors. Our research has shown that a mental workload human-performance moderator function for non-robotic domains was applicable to H-R teams with a master-slave relationship. It is also important to understand the applicability of this mental workload human performance function to a more complex, collaborative peer-based team relationship involving shared goals and dynamic decision-making.

Considering rationale in choosing when to deploy a human or robot team member and how to form H-R teams is critical. Human-robot collaboration offers the potential of leveraging human-ability advantages (e.g., object manipulation, visual assessment) and robot capabilities (e.g., sensors, logical deduction). The presented research assesses task performance and mental workload levels of human-human (H-H) and H-R collaborative teams in order to address potential differences between them. The results are compared to a previous scenario featuring a master-slave relationship between a human and robot (i.e., the robot was the master) in order to compare mental-workload and task-performance results with differing team relationships. There are three primary questions addressed by this paper: 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 mental 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?

This paper presents results focused on mental workload and team performance from modeling (Section 4.1) and validating, via user evaluations (Section 4.4), a collaborative team task and the Collaborative model and evaluation, respectively. The results are compared to the modeled (Section 5.1) and evaluation results (Sections 5.2-5.10) for the master-slave H-R team relationship, the Guided model and evaluation, respectively. The primary focus is the Collaborative scenario, which demonstrated a flexible relationship requiring joint decision-making between participants and a human or robot partner. The Guided scenario featured a more rigid, instruction-based relationship between the participant and the human or robot partner (Harriott, Zhang & Adams, 2011a, 2011b, 2013). The respective models represent mental workload and team performance using the Improved Performance Research Integration Tool Pro (IMPRINT Pro) tool (Allender et al., 1995). This research applies a variety of mental-workload measurement techniques in a non-laboratory setting.

Section 2 provides relevant background related to mental-workload metrics and collaborative H-R teams. Section 3 provides the scenario descriptions and describes IMPринt Pro modeling. Section 4 provides results from the Collaborative model and the corresponding evaluation. Section 5 compares the Guided and Collaborative scenario results, while Section 6 provides a discussion of the implications on H-R interaction.
2. Background

2.1 Collaborative Human-Robot Relationships

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 H-R 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), working in a shared space toward the same goals, and using a shared pool of knowledge. There may be separate tasks for the individual teammates, but no one teammate can complete the job without the other(s).

Collaboration between humans and robots has been investigated in a variety of contexts, including varying levels of autonomy (Brookshire, Singh, & Simmons, 2004; Stubbs, Hinds, & Wettergreen, 2007) and modeling robot behavior after human collaborative behaviors (Briggs & Scheutz, 2011; Chernova & Breazeal, 2010; St. Clair & Matarić, 2011). Research has not yet focused on understanding the implications of a collaborative relationship between a human and a 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 H-R 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 H-R team’s performance, whereas this paper also examines the collaborative relationship of human and robot’s effect on mental workload of the human.

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

Hinds, Roberts, and Jones (2004) assessed human feelings of personal responsibility, blame, and credit in H-R 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 H-H and H-R teams, given the different relationship types, were not investigated; however, this research question is addressed in the presented investigation.

The benefit to team performance of an autonomous robot has been shown, but the benefit or detriment of collaborating with a robot instead of a human partner has not been investigated in regard to mental workload and team performance. The presented results focus on comparing task performance and mental workload between H-H and H-R teams in order to understand how collaboration with a robot can impact human performance in relation to mental workload by evaluating the Collaborative scenario on its own and comparing the Collaborative scenario results...
to the results from the Guided scenario. The Collaborative scenario explores the comparison between H-H and H-R collaborative teams. The comparison between the Collaborative and Guided scenarios analyzes mental workload and task performance trends seen in collaborative teams in contrast to the master-slave relationship.

2.2 Mental Workload

Humans have limited mental resources, and because of this fact, mental workload can be defined as the difference between the amount of available mental processing resources and the amount required by a task (Hart & Staveland, 1988). For instance, an easy, routine task may require only 10 percent of a person’s available resources, whereas a very difficult task might require 90 percent of the person’s mental resources. Mental workload overload occurs when there are too few resources available to allocate to the required tasks, increasing stress and inaccuracies, whereas mental workload underload occurs when tasks consume too few available resources, increasing boredom. Both conditions can hinder overall performance (Nachreiner, 1995).

Mental workload can be measured using subjective and objective measures. Subjective measures include self-report surveys, such as the Multiple Resource Questionnaire (Boles, Bursk, Phillips, & Perdelwitz, 2007) and the NASA Task Load Index (Hart & Staveland, 1988). Subjective measures of mental workload require participants to rate the experienced demand along multiple channels. Subjective mental workload measurements can provide insight into how participants perceive demand, but typically must take place after completing a task or subtask, which can limit the granularity of the knowledge about how mental workload changes over time. Additionally, many subjective measures of mental workload are not absolute measures and cannot be directly compared between studies.

Objective metrics include both 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 et al., 2003), while heart rate variability (Aasman, G. Mulder, & L. J. Mulder, 1987; Castor et al., 2003; Roscoe, 1992) and respiration rate decrease as mental workload increases (Roscoe, 1992; Keller, Bless, Blomann, & Kleinbohl, 2001). Other physiological measures of mental 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, Reimer, Coughlin, & Dusek, 2009; Veltman & Gaillard, 1998; Wilson, Purvis, Skelly, Fullenkamp, & Davis, 1987). For example, an increase of mental workload results in a longer duration between blinks and an increase in galvanic skin response (Veltman & Gaillard, 1998).

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, Casson, & Rodriguez-Villegas, 2009; Brodbeck et al., 2011). Physiological sensors are subject to noise. Discriminating between signal and noise often requires the use of filters for noise reduction, but ultimately it 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) presented common H-R interaction metrics including subjective and objective mental-workload metrics. For example, mental workload can be estimated from the number of interventions (i.e., unplanned robot interactions) per unit of time. The number of interventions specifically refers to the number of interactions necessary in the supervisory relationship, with both the human and the robot intervening on the other. Alternatively, mental workload can be estimated as the ratio of operator-engaged task time to robot execution time or fan out (Goodrich & Olsen, 2003). Fan out refers to the number of robots that a single human can interact with successfully. These measures of mental workload are applicable to teleoperation and supervisory interaction but are not typically relevant in peer-based or collaborative interaction.

Other applications within the H-R interaction field define and measure mental workload in the same ways as in H-H interaction. Subjective mental workload surveys, such as the NASA Task Load Index survey (Hart & Staveland, 1988), have been applied to H-R interaction evaluations (e.g., Draper & Blair, 1996; Goodrich, McLain, Anderson, Sun, & Crandall, 2007; Kaber, Onal, & Endsley, 2000; Ruff, Narayanan, & Draper, 2002). Although physiological measures of mental workload are not commonly used in H-R interaction, they have been used to represent some affective states that may be related to mental workload (e.g., Liu, Rani, & Sarkar, 2006; Rani, Sims, Brackin, & Sarkar, 2002).

2.3 Task Performance

Task performance represents how well a person accomplishes a task and is measured objectively by gauging error, efficiency (e.g., time), and/or accuracy when completing the task (Gawron, 2008). Examples of task performance measures include how many times a person 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, Rahimi, & Casali. (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. It 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 of the task. The dissociation theory of workload and performance (Yeh & 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 person 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 H-R 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 H-R interaction scenario (Sheridan, 1992). Task performance metrics can be used to reflect the nature of the relationship between the human and robot in this manner.
2.4 Team Performance

Teams are able to accomplish complex, stressful, and ill-defined tasks, but measuring team performance is a difficult problem (Salas, Cooke, & Rosen, 2008). Team performance progresses as each member of the team copes with individual-level and team-level task work and teamwork. Individual task work represents what the team member completes without interaction (Kozlowski & Klein, 2000). Shared cognition is also a critical, and measurable, aspect of team performance that involves information processing during team-level work (Salas & Fiore, 2004; Cooke, Salas, Cannon-Bowers, & Stout, 2000). Shared cognition encompasses shared mental models and team situational awareness.

The aim of this paper is not to assess an encompassing team performance measure but to examine a human’s task performance and mental workload while working in a team, given either a human or robot partner and a collaborative or master-slave relationship style. Shared cognition has not been measured and the individual task work of the human’s partner (either human or robot) is not of interest in this investigation; therefore, this study does not assess the complete team performance picture.

3. Scenario Descriptions

Human performance models were created and evaluated for two scenarios: Collaborative (Section 3.1) and Guided (Section 3.2). Each scenario was modeled for H-H and H-R teams. The model results predict mental workload and team performance and were validated with user evaluations.

Human performance modeling is a useful tool in human-systems interaction and can be used for understanding H-R interaction. Modeling tools, such as IMPRINT Pro (Allender et al., 1995), incorporate human-performance functions, such as mental workload, derived from user evaluations in other domains. An open question is whether or not these performance functions can model the mental workload of H-R teams and the associated team performance.

3.1 Collaborative Scenario

The Collaborative scenario involved the reconnaissance of an academic building following a bomb threat. 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 & 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.

3.1.1 Scenario 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 Fig. 1. 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 Fig. 1.

![Figure 1](image1.png)

**Figure 1.** The map of the scenario environment indicating each investigation area shaded and labeled. All item locations are shown.

Nineteen suspicious and non-suspicious items were placed in the environment by the experimenters; see Table 1 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 1, for example, bulletin boards and white boards, recycling bins and trash cans, and fire extinguishers. Fig. 1 labels each item in the table with the corresponding item number while also delineating locations of benign objects.

Fig. 2.a demonstrates an individual working with the robot to assess a suspicious item. Some items were not clearly visible and required individuals to expend more effort to locate and identify, for example the fake pipe bomb placed under the eye wash station, as shown in Figs. 2.a and 2.b. Other items were clearly visible, even from a distance, such as a suspicious backpack placed on a bench (Fig. 2.c).

![Figure 2](image2.png)

**Figure 2.** (a) Evaluation participant and robot teammate inspecting a fake bomb, (b) fake bomb detail, and (c) an easily spotted suspicious backpack.
Table 1. The items listed by investigation area with a description and indication as to whether or not the item was suspicious.

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

3.2 Guided Scenario Description

The Guided scenario tasked untrained individuals with performing medical triage steps using regimented instructions provided by a first responder partner (Harriott et al., 2011a, 2011b). The triage steps included assessing the victims’ breathing rate, pulse, and heart rate, and checking for injuries. Each victim was categorized into one of three triage levels: delayed, immediate, or expectant. The first response partner directed the participant in conducting eleven victim assessments (1-11) on six simulated victims (1-6) (automated medical mannequins) in two triage rounds, during which some victims’ triage levels changed between rounds. The Guided scenario assumed that the first response partner was strictly acting in an instructional role, guiding the participant through the triage steps necessary to assess the victims. The scenario assumed that a human first responder was unable to access the environment due to unknown contaminants and provided instructions via a mobile electronic device (e.g., cell phone). The scenario also assumed that the robot first responder entered the environment, moved with the individual from victim to victim, and provided instructions directly. This scenario represents a master-slave relationship.

3.3 Modeling Description

Human performance models were developed using IMPRINT Pro for both scenarios, where each teaming condition was modeled (Allender et al., 1995; U.S. Army Research Laboratory [ARL], 2009). IMPRINT Pro was developed in order to assess alternate human-system configurations (Archer, Gosakan, Shorter, & Lockett, 2005). 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 inputs to the model. 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 scenario and for the two teaming conditions: H-H and H-R. The resulting models are sequential networks representing subtasks that individuals perform during the scenario. IMPRINT Pro’s mental workload representation provides predictions by integrating seven individual mental workload channels (i.e., Cognitive, Auditory, Visual, Fine Motor, Gross Motor, Speech, and Tactile) (ARL, 2009).

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) (ARL, 2009).

The H-H and the H-R 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.

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 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. For example, the backpack seen in Fig. 2c is visible, but either partner is equally likely to report the bag to the other. 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 Guided model only produced one value, due to a lack of uncertainty or alternate paths in the model. 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.

4. Collaborative Evaluation

The Collaborative evaluation implemented the Collaborative scenario presented in Section 3.1. The evaluation hypotheses were as follows: 1) the model will accurately predict the human’s mental workload in both conditions, 2) the mental workload will be lower in the H-R team than in the H-H team, and 3) task performance will not be significantly different between H-H and H-R teams. Although this scenario incorporates uncertainty and joint decision-making, the hypotheses were partially based on our previous Guided scenario results. In that scenario, we found that the model accurately predicted mental workload for both teaming conditions and that mental workload
was lower in the H-R team than in the H-H team (Harriott et al., 2011a, 2011b). The Collaborative evaluation featured team collaboration, arguably a more complex interaction between teammates, but the H-R team mental workload is hypothesized to be lower due to the longer time that the robot partner required to complete tasks.

4.1 Model 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 then 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 (SD = 0.58) and 12.96 (SD = 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 2. A Kruskal-Wallis test indicated a significant main effect of investigation index for 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 2. 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.

Table 2. Mean modeled mental workload by investigation area index and condition. Standard deviations are listed in parentheses.

<table>
<thead>
<tr>
<th>Investigation Index</th>
<th>Across Both Models</th>
<th>H-H Model</th>
<th>H-R Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>12.78 (0.71)</td>
<td>12.86 (0.80)</td>
<td>12.70 (0.61)</td>
</tr>
<tr>
<td>Medium</td>
<td>13.24 (0.33)</td>
<td>13.41 (0.31)</td>
<td>13.08 (0.27)</td>
</tr>
<tr>
<td>High</td>
<td>13.19 (0.36)</td>
<td>13.28 (0.35)</td>
<td>13.10 (0.35)</td>
</tr>
</tbody>
</table>
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 Eq. 1:

\[
Task\ Density = \frac{(Number\ of\ Items) \times (Size\ of\ Area)}{Area\ Time}.
\]  

Area Time refers to the amount of time the model predicted that participants would require to investigate each area; the Number of Items per area is provided in Table 1; and the Area Size (meters squared) is presented in Table 3. 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 (SD = 0.16) for the H-H condition, while the H-R condition was 0.34 (SD = 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.

### Table 3. Investigation area size (m²).

<table>
<thead>
<tr>
<th>Investigation Area</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area 1</td>
<td>52.43</td>
</tr>
<tr>
<td>Area 2</td>
<td>53.28</td>
</tr>
<tr>
<td>Area 3</td>
<td>84.74</td>
</tr>
<tr>
<td>Area 4</td>
<td>57.32</td>
</tr>
<tr>
<td>Area 5</td>
<td>43.73</td>
</tr>
<tr>
<td>Area 6</td>
<td>124.04</td>
</tr>
</tbody>
</table>

The mean task density for low investigation index areas was 0.45 (SD = 0.13), 0.38 (SD = 0.13) for medium investigation index areas, and 0.41 (SD = 0.20) for high investigation index areas in the H-H model. The mean task density was 0.37 (SD = 0.09) for low investigation index, 0.30 (SD = 0.09) for medium investigation index areas, and 0.35 (SD = 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, Lee, Liu, & Gordon-Becker, 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 (Eq. 2):

\[
Work\ Completed = Mental\ Workload \times AreaTime.
\]  

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. Eq. 2 factors out timing by multiplying workload by area time and results in only the work completed, as shown in Eq. 3:

\[
Work = \frac{Work}{Time} \times Time.
\]
Eq. 3 is a simple estimation, but offers insight into whether or not the time the teams’ requirement 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.

4.2 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 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.

4.3 Collaborative Evaluation Apparatus

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.

The evaluation environment was divided into six investigation areas, each with a corresponding investigation index representative of an expected mental workload level: low, medium, or high (as detailed in Section 3.1). The independent variables were the experimental condition, the area investigation indices, and participant age, gender, experience with robots, and first response training. The dependent variables include subjective and objective mental workload metrics (please see Section 4.3.2) and objective task performance metrics.

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 Fig. 1. 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.
4.3.1 Participants

Thirty-six participants completed the evaluation, with eighteen in each condition. The 19 male and 17 female participants were not experts in first response or robotics and ranged in age between 18 and 56 years old. The mean age of H-H participants was 27.4 ($SD = 8.4$ years), while the mean age of H-R participants was 24.1 ($SD = 4.6$ years). Participants rated their search and rescue response expertise on a Likert scale from 1 (little or no experience) to 5 (very experienced), with a median response of 1 for both conditions. The minimum response was 1 and maximum response was 3 in both conditions. Participants provided a median response of 1 in both conditions when rating their experience with robots on the same scale, with a minimum response of 1 for both conditions, a maximum response of 4 in the H-H condition, and a maximum response of 5 in the H-R condition. Each participant had normal or corrected-to-normal visual acuity.

4.3.2 Evaluation Metrics

The objective metrics captured physiological metrics via a portable BioHarness electrocardiography monitor (Biopac.com, 2013) and included: heart rate variability, beat-to-beat interval, heart rate, respiration rate, skin temperature, posture, vector magnitude data, and vector acceleration data. Additional objective metrics included walking data from a Garmin footpod pedometer (Garmin.com, 2011) and time spent searching each Investigation Area. The objective metrics also included the correctness of responses to the secondary task questions (STQs), STQ response time, and the number of times an STQ was repeated. Other metrics included: Item Response Time, Item Find Type, the number of items found, the number of photographs taken, and the number of items recalled post-experiment\(^1\). The subjective metrics included in situ mental workload ratings collected after searching each investigation area, post-experimental questionnaire responses\(^2\), and NASA-TLX mental workload (Hart & Staveland, 1988) responses.

The reported results included physiological measures of normalized heart rate, normalized respiration rate, and heart rate variability (HRV). Heart rate and respiration rate were normalized by determining the mean difference between a participant’s raw heart/respiration rate during the investigation of each area and the mean of a baseline heart/respiration rate captured during training. The absolute value was not taken, because the directionality from baseline is an important indicator of change in respiration rate. Normalized heart rate is measured in beats per minute, while normalized respiration rate is measured in breaths per minute. HRV was calculated by taking the mean ratio of low-frequency HRV over the high frequency HRV.

The STQ responses represented the participants’ spare mental capacity for recognizing items from a list and recalling specific associated values. The participants were given one minute to memorize a list of six chemicals, each with an associated Danger Level:

- Chlorobenzene, 75
- Acetate, 25
- Naphtha, 100
- Ethylamine, 10
- Pyridine, 5
- Ammonia, 50

\(^1\)Please note that the following objective metrics are not reported in this manuscript: beat-to-beat interval, skin temperature, posture, vector magnitude data, acceleration data, walking data, STQ response time, Item Response Time, Item Find Type, number of photographs taken, and number of items recalled post-experiment.

\(^2\)The post-experimental questionnaire responses are not presented.
Twelve STQs related to the chemicals were asked, independent of condition. The standard question structure was as follows: “Was X on the list of chemicals you received?” If the participants responded “Yes,” regardless of the ground truth, the structure of the follow-up question was, “What was the danger level of X?” The chemicals (represented by X in the questions) in order of inquiry were: Ethylamine, Acetate, Acetone, Ammonia, Chlorine, Naphtha, Ethane, Propane, Chlorobenzene, Nicotine, Acrolein, and Pyridine. The dependent variable, On List Correct, represented if participants answered correctly when an item was on the list and the Danger Level Correct metric determined if participants correctly provided an item’s danger level.

Task performance represented the number of items found in the environment by each team. The numbers of items found by participants were determined in two ways: 1) the total number of items found and 2) the total number of additional items reported. A total of 26 items were located throughout the environment, as described in Section 3.1.

The in situ subjective mental workload metric required the participants to verbally rank six mental workload channels upon completing the search of each investigation area. The six mental workload channels matched those used in the performance modeling (see Section 3.3): Cognitive, Auditory, Visual, Tactile, Motor, and Speech. The rating range was a scale from 1 (little to no demand) to 5 (extreme demand), which was selected to match the normalized results from the modeling. Each channel was defined during training, but participants were able to ask for channel definitions during any of the mental workload questions. The questions were asked verbally and the verbal responses were recorded. The definitions were adapted from the Multiple Resources Questionnaire (Boles et al., 2007):

- Auditory: recognizing words, tones, mood and emotions through sound.
- Visual: recognizing faces and objects, sustaining visual attention, judging distances, spatial reasoning and reading.
- Speech: instances when you used your own voice.
- Motor: movement and control of face muscles, arms, hands, fingers, legs and feet.
- Tactile: recognition and judgment of shapes using the sense of touch.
- Cognitive: making judgments, estimation, learning, problem-solving, decision-making, reasoning, memorization and recalling from memory.

The NASA-TLX questionnaire assessed mental workload upon evaluation completion. Due to an experimenter error, data from eight participants in the H-H condition were not recorded.

Total investigation time began when the human or robot partner was introduced and ended upon completion of last set of in situ mental workload questions. Individual area investigation times were recorded, beginning when the previous area’s in situ mental workload questions were completed and ending when the next set of in situ mental workload questions began.

Two video coders determined the start and end times for each total investigation time, area investigation time (not presented), item response time, and secondary task question response time. STQ responses, in situ subjective mental workload ratings, items found, and item find type were also coded. Intercoder reliability was calculated to be 0.71 using exact timing matches and 0.86 when times were within one second between coders. The differences in coder timing responses were handled by taking a mean of the two if the difference was below a threshold (e.g., two seconds for response times), and the video was checked by a third coder for data entry error if the timing difference was over the threshold.

4.3.3 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, the in situ subjective workload questions were explained to the participant. Definitions were provided and the participants were provided an opportunity to ask any questions. The participants were then given one minute to memorize the list of STQ chemicals. 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.

The responder halted the immediate investigation area search when the area border was reached, as shown in Fig. 1. 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 mental workload ratings. Two STQs were posed during each of the six areas for a total of 12 STQs. After all six investigation areas were completed, participants responded to a post-trial questionnaire and the NASA-TLX mental workload questionnaire.

4.4 Results

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

4.4.1 Physiological Measures

Higher levels of low-frequency heart rate variability have been shown to correspond with lower mental workload (Aasman et al., 1987; Castor et al., 2003; Roscoe, 1992). The mean low-frequency HRV for the H-H condition was 402431 (SD = 1081004) ms² and 249469 (SD = 1025228) ms² 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 HRV for the H-R condition.

The mean normalized heart rate for the H-H condition was -3.15 (SD = 26.95) beats per minute and 3.83 (SD = 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 (SD = 3.62) breaths per minute and 2.54 (SD = 4.03) breaths per minute for the H-R condition. Fig. 3 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 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 4. 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).

<table>
<thead>
<tr>
<th>Investigation Index</th>
<th>Across Both Conditions</th>
<th>H-H</th>
<th>H-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>1.40 (4.08)</td>
<td>0.20 (3.62)</td>
<td>2.60 (4.18)</td>
</tr>
<tr>
<td>Medium</td>
<td>0.99 (4.13)</td>
<td>-0.27 (3.77)</td>
<td>2.25 (4.14)</td>
</tr>
<tr>
<td>High</td>
<td>1.31 (3.96)</td>
<td>-0.14 (3.56)</td>
<td>2.76 (3.86)</td>
</tr>
</tbody>
</table>

Table 4. Mean normalized respiration rate by condition and investigation index.

Figure 3. Mean normalized respiration rate by condition and investigation area.

4.4.2 Timing

The mean total time taken to complete the entire investigation for H-H condition participants was 1964.18 (SD = 373.79) seconds (s) and was 2473.90 (SD = 213.16) s for the H-R condition participants. A Kruskal-Wallis test revealed that the H-R condition took significantly longer, which can be attributed to the longer time it took the robot to speak and move throughout the environment, \( \chi^2(1) = 14.58, p < 0.01 \). This result suggests that the longer H-R condition times may impact the lower H-R condition mental workload levels, which is analyzed via the strength of the correlation between the investigation times and the in situ subjective mental workload ratings (as provided in Sections 4.4.5 and 4.4.8, respectively).
4.4.3 Secondary Task Questions

A total of 172 correct and 37 incorrect responses to the STQs 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 5. 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 the 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.

Table 5. Total number of correct (C) and incorrect (I) responses to On List questions, by investigation index and condition.

<table>
<thead>
<tr>
<th>Investigation Index</th>
<th>Across Both Conditions</th>
<th>H-H</th>
<th>H-R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>I</td>
<td>C</td>
</tr>
<tr>
<td>Low</td>
<td>119</td>
<td>21</td>
<td>55</td>
</tr>
<tr>
<td>Medium</td>
<td>111</td>
<td>28</td>
<td>61</td>
</tr>
<tr>
<td>High</td>
<td>98</td>
<td>42</td>
<td>56</td>
</tr>
</tbody>
</table>

The On List Correct results demonstrate that mental workload reserves were not significantly different 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 STQs 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 6. 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.

Table 6. Total number of correct (C) and incorrect (I) responses to Danger Level questions, by investigation index and condition.

<table>
<thead>
<tr>
<th>Investigation Index</th>
<th>Across Both Conditions</th>
<th>H-H</th>
<th>H-R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>I</td>
<td>C</td>
</tr>
<tr>
<td>Low</td>
<td>66</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>Medium</td>
<td>43</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>High</td>
<td>19</td>
<td>13</td>
<td>11</td>
</tr>
</tbody>
</table>

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 STQs correctly with similar frequency, showing that the two conditions resulted in similar levels of spare mental capacity.

4.4.4 Task Performance

Task performance was based on the number of items assessed by participants, including both the items in Table 1 and any extra items. The mean number of items from Table 1 assessed by the H-H condition participants was 24.11 ($SD = 0.81$), while the H-R condition participants assessed an average of 24.33 items ($SD = 0.82$). A Kruskal-Wallis test found no significant effect of condition for the number of 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 ($SD = 1.73$), while the H-R participants found a mean of 1.44 extra items ($SD = 1.21$). 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.

4.4.5 In Situ Subjective Mental 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 7. 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 mental workload ratings. Fig. 4 provides the median in situ mental workload rating values for each investigation area.

Table 7. In situ subjective mental workload ratings by condition and investigation index. Med = median, Min = minimum, and Max = maximum.

<table>
<thead>
<tr>
<th>Investigation Index</th>
<th>Across Both Conditions</th>
<th>H-H</th>
<th>H-R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Med</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Low</td>
<td>11.5</td>
<td>6.0</td>
<td>29.0</td>
</tr>
<tr>
<td>Medium</td>
<td>10.5</td>
<td>6.0</td>
<td>27.0</td>
</tr>
<tr>
<td>High</td>
<td>12.5</td>
<td>6.0</td>
<td>27.0</td>
</tr>
</tbody>
</table>
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.

### 4.4.6 Comparison of Modeled Mental Workload and In Situ Subjective Mental 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. Fig. 5 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 ($SD = 5.24$) and the H-R condition’s result was 10.97 ($SD = 3.98$). The mean total mental workload from the H-H Model was 13.18 ($SD = 0.58$), while the H-R model resulted in a mean of 12.96 ($SD = 0.47$), as presented in Section 4.1. The mean total modeled mental workload was 13.07 ($SD = 0.53$) across both models, while the mean total subjective mental workload rating was 12.48 ($SD = 4.88$) 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 ($p < 0.01$) (see Section 4.1). 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 & Pashler, 2000). Table 8 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).
Table 8. Confidence interval (C.I.) results of the model and evaluation results.

<table>
<thead>
<tr>
<th>Area</th>
<th>Area 1</th>
<th>Area 2</th>
<th>Area 3</th>
<th>Area 4</th>
<th>Area 5</th>
<th>Area 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Mean</td>
<td>12.6</td>
<td>12.1</td>
<td>14.1</td>
<td>13.3</td>
<td>13.8</td>
<td>13.3</td>
</tr>
<tr>
<td>Model SD</td>
<td>1.24</td>
<td>1.2</td>
<td>1.34</td>
<td>1.31</td>
<td>1.34</td>
<td>1.34</td>
</tr>
<tr>
<td>C.I. Low Cutoff</td>
<td>11.7</td>
<td>11.3</td>
<td>13.2</td>
<td>12.4</td>
<td>12.8</td>
<td>12.8</td>
</tr>
<tr>
<td>C.I. High Cutoff</td>
<td>13.5</td>
<td>12.9</td>
<td>15.1</td>
<td>14.3</td>
<td>14.7</td>
<td>14.7</td>
</tr>
<tr>
<td>Evaluation Mean</td>
<td>12.7</td>
<td>10.8</td>
<td>13.3</td>
<td>10.2</td>
<td>14.3</td>
<td>11.5</td>
</tr>
<tr>
<td>Within C.I.?</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Within One SD of Model Mean?</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

4.4.7 NASA-TLX Responses

The mean total NASA-TLX score for the H-H condition was 41.32 (SD = 17.80) and 30.89 (SD = 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.

4.4.8 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 (SD = 0.09), while the H-R task density was 0.20 (SD = 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 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 (SD = 0.07), 0.23 (SD = 0.08) in medium investigation areas, and 0.23 (SD = 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.
4.4.9 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 (see Section 4.1). The amount of work completed by H-H participants was 4344.23 (SD = 2512.33) mental workload seconds and 4226.55 (SD = 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 Fig. 6. The values between the conditions are very similar for the individual investigation areas; thus, both conditions resulted in similar levels of work.

![Figure 6. Amount of work completed by investigation area and condition.](image)

4.4.10 Correlations Analysis

Pearson’s product-moment correlations were performed to correlate physiological data with the in situ mental workload ratings. HRV 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.

4.5 Discussion of Collaborative Evaluation Results

The hypothesis that the models are a good predictor of mental workload in both evaluation conditions was supported for the H-H condition only. 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 second hypothesis stated that the H-R condition will lead to lower mental workload and was supported. The model data and the in situ mental workload ratings were both lower in the H-R condition. Respiration rate was significantly higher in the H-R condition and resulted in a significant negative correlation to in situ mental workload values, supporting the claim from the literature (Keller et al., 2001; Roscoe, 1992) that respiration rate decreases as mental workload increases.
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, Altmann, Brock, & Mintz, 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.

The third hypothesis was shown to be true; 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 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 model and evaluation directly compare H-H and H-R collaborative teams. The research seeks to answer whether or not robots and humans can collaborate to the same degree as humans and has demonstrated successful collaborations. As the development of robot technology moves toward incorporating H-R collaboration, it is important to determine how H-H and H-R collaboration are similar and different.

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. Future work will analyze the aspect of how verbal communication slows the collaboration and how other communication options (e.g., gesture) may circumvent the language bottleneck experienced in this evaluation.

This research demonstrates 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

82
saw no significant detriments to task performance nor significantly lower levels of work completed.

5. Comparison Across Teaming Scenarios

Prior work focused on the Guided scenario (Section 3.2), in which the relationship with the participant was more of a master-slave relationship. The participant’s teammate, an experimenter or a robot, provided instructions to perform victim triage. This section compares the Guided and Collaborative scenario results.

The Guided evaluation assessed mental workload via heart rate variability, subjective in situ and post-trial ratings, and secondary task question responses (Harriott et al., 2011a, 2011b). The human partner communicated with participants over a walkie-talkie from a remote location. The robotic partner, a Pioneer 3-DX, was physically present and moved autonomously between victims with participants. Participants rated in situ mental workload after completing the triage of each victim, and the NASA-TLX ratings were elicited from participants following the evaluation.

The Guided scenario models, as described in Section 3.3, were validated in a similar manner to the Collaborative scenario (Harriott et al., 2011a, 2011b). Twenty-eight participants completed the triage evaluation, with fourteen in each condition (H-H and H-R). The mean age of the H-H participants was 24.2 (SD = 10.3) years, while the mean was 26.2 (SD = 7.8) years for the H-R participants.

The results indicated that subjective mental workload was lower for the H-R condition. Additionally, physiological measures demonstrated the same trends; H-R condition participants experienced lower mental workload. Participants also took a significantly longer time to complete H-R condition tasks (Harriott et al., 2011a).

This section seeks to determine if the same mental workload trends occurred across the Guided and Collaborative interaction relationships. New analysis is provided for some metrics, while summaries of previously presented physiological measures, STQs, NASA-TLX ratings, and correlations results (Harriott et al., 2011a, 2011b; Harriott, Buford, Zhang, & Adams, 2012) 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.

5.1 Guided Scenario Model Results

The Guided scenario model did not include uncertainty, thus the results are derived from a single simulation trial (see Section 3.3). The Guided scenario total modeled mental workload values were normalized to the 6–30 range (see Section 4.4.5) as a new presentation of these results. The mean mental workload for the H-H model across the 11 triage assessments was 15.73 (SD = 2.36) and 15.17 (SD = 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.

5.2 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 HRV found a mean H-H condition low-frequency HRV of 122681.1 ms² (SD = 424825.9), while the mean low-frequency HRV for the H-R condition was
264364.5 ($SD = 875409.6$) 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 HRV 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 HRV. The Guided evaluation HRV 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).

5.3 Timing

Timing was calculated by measuring the amount of time it took for participants to perform each triage assessment. The time began when the participant arrived at a victim and stopped when the victim triage level was determined. The mean assessment time was 136.42 ($SD = 60.92$) s for the H-H condition and 151.53 ($SD = 70.73$) s for the H-R condition. A Kruskal-Wallis test determined that the H-R participants took significantly longer to triage the victims, $\chi^2(1) = 5.55$, $p = 0.02$. The Guided evaluation timing results echo those from the Collaborative evaluation, in which the H-R participants took longer to complete tasks.

5.4 Secondary Task Questions

The STQs 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 STQ responses between the two conditions in either scenario.

5.5 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 9. 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 10. 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.
Table 9. Mean reported age by condition and victim number, in years.

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

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.

Table 10. Mean reported respiration rate by condition and assessment number in breaths per minute.

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

5.6 In Situ Subjective Workload Ratings

The Guided evaluation total in situ subjective mental workload ratings ranged from 6 to 30 and were gathered after each victim assessment in the same manner as in the Collaborative evaluation (see Section 4.4.5).

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 11. 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
mental workload ratings from both evaluations indicate the same trend: The H-R condition participants reported lower reported subjective mental workload values.

Table 11. Median total subjective mental workload ratings by condition and triage level for the Guided evaluation.

<table>
<thead>
<tr>
<th>Triage Level</th>
<th>Across Both Conditions</th>
<th>H-H</th>
<th>H-R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Med</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Delayed</td>
<td>15.0</td>
<td>6.0</td>
<td>30.0</td>
</tr>
<tr>
<td>Immediate</td>
<td>14.0</td>
<td>6.0</td>
<td>30.0</td>
</tr>
<tr>
<td>Expectant</td>
<td>13.0</td>
<td>6.0</td>
<td>30.0</td>
</tr>
</tbody>
</table>

5.7 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.

5.8 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 mental workload across both conditions (H-H and H-R models) was 15.45 (SD = 2.04) and 14.87 (SD = 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 (Section 4.4.6).

5.9 Work Completed

The amount of work completed during each victim assessment was computed as defined in Section 4.1. The H-H participants had a mean work completion value of 2235.93 \( (SD = 1524.69) \) workload seconds and the H-R participants had a mean value of 2101.70 \( (SD = 1331.11) \) 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.

5.10 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, \( r(290) = -0.15, p < 0.01 \).
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.

5.11 Discussion of Comparison Across Teaming Scenarios

The Guided evaluation focused on a relationship between a human and a robot, which was predicated on 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 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 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). 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 evaluation 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 mental 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 mental 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 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.
6. General Discussion

The Guided and Collaborative scenarios were each designed to develop human performance models and validate their predictions via evaluation data. 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 validation 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 paper discusses three physiological measures (i.e., heart rate, respiration rate, and heart rate variability), two workload equations, one secondary task measure, and two subjective measures of workload. All of the workload metrics, and their advantages and disadvantages, are presented in Table 12. 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, Lee, Liu, & Gordon-Becker, 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.
Table 12. Summary of evaluated workload metrics, with data collection method and associated advantages and disadvantages.

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

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.

The presented research 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.

6.1 Limitations

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

6.1.1 Robot 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. As discussed in Section 4.5, 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 of the robot cutting its human partner off mid-sentence or beginning to talk at the same time as when the human partner was performing a task involving concentration. 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.

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 companions 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 & 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 Lee, Morency, & Forlines, 2006). Employing mutual gaze and gazing at objects that are the subjects of conversation can improve the conversation (Mutlu, Shiwa, Kanda, Ishiguro, & Hagita, 2009; Sidner, Kidd, Lee, & Lesh, 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 & 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.

6.1.2 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 & 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 executor 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.

Acknowledgements

The authors thank Sara Kiesler and three anonymous reviewers for their detailed review of this manuscript. The authors also thank Sean Hayes, Eli Hooten, Travis Service, and Corey Wollaeger. This research was performed at Vanderbilt University, supported by AFOSR Award FA9550-09-1-0108 and partially supported by NSF Grant IIS-0643100 and the ONR Award N00014-12-1-0987.
References


Authors’ names and contact information: C. E. Harriott, Charles Stark Draper Laboratory, Cambridge, MA, USA. Email: charriott@draper.com; G. L. Buford, Vanderbilt University, Nashville, TN, USA. Email: glenna.l,buford@gmail.com; T. Zhang, Purdue University, West Lafayette, IN, USA. Email: zhangtao2000@gmail.com; J. A. Adams, Vanderbilt University, Nashville, TN, USA. Email: julie.a.adams@vanderbilt.edu.