Evaluating the applicability of current models of workload to peer-based human-robot teams

Conference Paper · January 2011

DOI: 10.1145/1957656.1957670 · Source: DBLP

3 authors, including:

Caroline E Harriott
Piaggio Fast Forward
24 PUBLICATIONS **96** CITATIONS

SEE PROFILE

Tao Zhang
Environmental Systems Research Institute (ESRI)
46 PUBLICATIONS **195** CITATIONS

SEE PROFILE

Evaluating the Applicability of Current Models of Workload to Peer-based Human-robot Teams

Caroline E. Harriott Tao Zhang Julie A. Adams

Department of Electrical Engineering and Computer Science
Vanderbilt University
Nashville, TN, USA
1-615-322-8481

caroline.e.harriott@vanderbilt.edu tao.zhang@vanderbilt.edu julie.a.adams@vanderbilt.edu

ABSTRACT

Human-Robot peer-based teams are evolving from a far-off possibility into a reality. Human Performance Moderator Functions (HPMFs) can be used to predict human behavior by incorporating the effects of internal and external influences such as fatigue and workload. The applicability of HPMFs to human-robot teams is not proven. The presented research focuses on determining the applicability of workload HPMFs in team tasks for first response mass casualty triage incidents between a Human-Human and a Human-Robot team. A model representing workload for each team was developed using IMPRINT Pro. The results from an empirical evaluation were compared to the model results. While significant differences between the two conditions were not found in all data, there was a general trend that workload in the human-robot condition was slightly lower than the workload experienced in the human-human condition. This trend was predicted by the IMPRINT Pro models. These results are the first to indicate that existing HPMFs can be applied to human-robot peer-based teams.

Categories and Subject Descriptors

H.1.2. [**Information Systems**]: User/Machine Systems – *Human Factors.*

I.2 [**Artificial Intelligence**]: Robotics.

General Terms

Performance, Experimentation, Human Factors

Keywords

human-robot peer-based teams, human-performance modeling

1. INTRODUCTION

Robotic technology continues to develop and humans are beginning to be partnered with robots for peer-based tasks [1, 2]. It is known that individual human performance can impact human team performance [3]. Similarly, human performance will impact the task performance of human-robot teams (HRTs). As human-robot team capabilities improve, it is necessary for the robotic team members to understand how the human's performance capabilities affect the task at hand. Future robots should potentially adapt their behavior, as humans do in human teams, to mitigate and accommodate performance changes in their human partners. Thus, it is necessary to understand if and how existing human performance moderator functions (HPMFs) apply to HRTs. Developing such understanding necessitates modeling HPMFs for HRTs, conducting evaluations to gather empirical results, and understanding how the empirical results relate to the modeled HPMFs.

HPMFs are equations derived from empirical results that predict human performance due to specific performance factors such as fatigue, mental workload or temperature. Approximately 500 HPMFs [4] are known to exist, and it is well known that a number of interactions exist across HPMFs. Our current research is focused on workload. HPMFs have been evaluated for various domains such as aviation [5] and nuclear power plants [6]. These domains are more regulated and controlled than the potential domains in which HRTs will be deployed, for example, first response to CBRNE disasters [7]. Thus, it is necessary to understand the applicability of HPMFs to HRTs.

A significant amount of research in the human-robot interaction domain has focused on measuring human performance [8, 9], but little research has focused on modeling the HPMFs for interaction between peer-based HRTs. Work by Howard [10, 11] is closely related to this research. It modeled and predicted human and system performance for repetitive collaborative tasks with a teleoperated robot. While the goal of Howard's research is similar to that of this research, Howard's research did not examine peer-based human-robot interaction.

Existing results indicate that heart rate variability (HRV), heart rate, and respiration rate can be employed to assess workload [12, 13, 14], with HRV being cited as a reliable measure of workload [15]. While no physiological measure perfectly reflects changes in workload, correlations between HRV and subjective workload measures have proven significant. Secondary tasks and subjective workload measures have also proven useful for assessing workload [16].

The reported research focused on teaming a human with a partner, human or robot to complete a task. Both human-human and human-robot teams were modeled and evaluated in order to validate that the workload HPMF for the humans are accurate

HRI'11, March 6–9, 2011, Lausanne, Switzerland.

before analyzing the HRT. It should be noted that in the reported research, the partner (human or robot) is instructing the primary human (e.g., the participant) on the task steps to complete. There are no shared decision making tasks.

2. BACKGROUND

The research scenario is a victim triage task for a mass casualty CBRNE incident. CBRNE incident response procedures dictate that first responders are not to enter the contaminated incident scene until a decontamination site is established, the potential hazards are identified, and personal protective equipment is donned [17]. Any response delay can result in significant civilian injury or death. Robots have the potential to enter the scene immediately in order to provide immediate feedback regarding victim locations, triage victims, etc. Such information can assist human responders in determining the appropriate response, including locating, treating, and transporting victims [7]. While robotic technology is not yet capable of all these tasks, robots may be able to enter a scene, identify an uninjured ambulatory victim, and recruit that victim to assist with victim triage. Such human volunteers can assist with locating and triaging victims. The robot can then relay information from the scene to first responders located outside the contaminated area.

The START [18] triage system is commonly employed to triage victims during emergency incidents [19, 20]. The START steps require about 60 seconds when completed by a trained responder and focus on assessing the immediacy of care required for a particular victim. START involves a number of steps intended to classify a victim into one of four triage levels: Minor, Delayed, Immediate and Expectant. Minor implies that the victim is ambulatory and coherently responsive. Delayed victims can survive while waiting up to a few hours for care. Immediate indicates that the victim needs to be treated as soon as possible. Expectant specifies that the victim has passed away or will soon expire. The first step determines if the victim is breathing or not (Respirations). If the victim is breathing, the number of breaths per minute is measured. If the victim is not breathing, an attempt should be made to open the airway. A non-breathing victim is classified as Expectant and a breathing victim with over 30 breaths per minute is classified as Immediate. A blanch test or the victim's pulse is measured for all other victims. A blanch test requires the responder to press the victim's fingernail until the color fades, let go and measure the time until normal color returns. If the fingernail takes longer than two seconds to refill, then the victim needs immediate care. Alternatively, the victim's pulse can be measured. If the pulse is not present or irregular, the victim is classified as Immediate. If the victim still is unclassified, the first responder assesses the victim's mental responsiveness by asking a question or asking the victim to open and shut the eyes. If the victim is unresponsive, he or she is classified as Immediate. If the victim is responsive, the classification is Delayed.

The presented research focused on analyzing workload for a situation in which a CBRNE incident has occurred and multiple non-ambulatory victims require triage. This research assumed that the volunteer has little first aid training, no prior experience with robots, and forms an ad-hoc team with either a human first responder located outside of the contaminated area or the robot, depending upon condition.

3. HUMAN PERFORMANCE MODELING

Human performance modeling simulates human behavior under a variety of conditions and tasks. HPMFs can be incorporated into Human Performance Models (HPMs) in order to improve model fidelity [21]. Several domains have employed HPMs in order to understand how system design, task assignments and environmental changes impact human behavior and performance. For example, the NASA Human Performance Modeling Project [5] incorporated multiple modeling techniques to investigate aspects of human performance in aviation tasks.

IMPRINT Pro is an event network modeling tool intended to assess human and system performance [22, 23]. It has been used to model personnel on a United States Navy destroyer bridge, the U.S. Army's Crusader System [24], and pilot performance for simulated unmanned air vehicles missions [25]. IMPRINT Pro simulates human behavior for a variety of conditions through the representation of task and event networks. IMPRINT Pro includes a number of pre-defined HPMFs (e.g., workload) and permits the incorporation of undefined HPMFs via the User Stressors module. IMPRINT Pro has been employed in the reported research to model both the Human-Human (H-H) and Human-Robot (H-R) conditions.

The developed models represent a team-based scenario involving first responders (both robot and human) instructing an ambulatory, uninjured victim/volunteer located in the contaminated incident area to perform triage on nearby non-ambulatory victims. The models represent the task activities and the uninjured volunteer's workload. The workload HPMF is decomposed into seven distinct channels: Cognitive, Auditory, Visual, Fine Motor, Gross Motor, Speech, and Tactile.

The triage scenario requires the volunteer to perform the START triage steps on six victims with differing levels of required triage and then repeat the triage steps on all victims who were not classified as Expectant during the initial triage. During second triage round, the order of attending to victims is based on triage level, with those having the most severe triage classification being visited first.

The modeled H-H scenario assumes that the uninjured victim has contacted 9-1-1 to report the incident and has volunteered to assist a remote (e.g., located outside of the contaminated incident area) first responder with the triage task. The scenario further assumes that the uninjured volunteer communicates with the first responder via cell phone and that the first responder provides step-by-step instructions that lead the volunteer through the triage steps. The volunteer provides responses that are recorded by the first responder to assist with incident response planning. The modeled H-R scenario assumes that the robot has been deployed into the contaminated incident area and has discovered the uninjured human victim who has volunteered to assist the robot with the triage task. The volunteer executes the instructions provided by the robot and reports results to the robot. The robot reports this information, as well as the location of the injured victim to remote first responders. The robot communicates with the volunteer using voice interaction. Both scenarios use the same task, victim order, triage instructions, and victim information. The only difference is that the H-R model accounts for the robot's slower speech pace and an extra step of placing a triage card on each victim.

The IMPRINT Pro models iterate through each triage scenario task. Tasks include each START triage step for the individual victim's needs, broken into discrete, atomic tasks based on each individual action the volunteer takes (e.g., feel for a pulse, count breaths for one minute, verbally report findings). Each task the volunteer is assigned has an associated workload value for each of the seven channels: Cognitive, Auditory, Visual, Fine Motor, Gross Motor, Tactile and Speech. A numeric value is assigned to each channel, which results in an overall workload associated with a particular atomic task. Each channel has an independent value scale and predefined guidelines for choosing an associated value. For example, the cognitive channel scale is 0 (minimum) to 7 (maximum) cognitive workload.

The presented research focuses on the results for the six injured victims. Once the model completes execution, the list of tasks completed by the volunteer is provided. The results include the time required to complete the task, the associated workload value for each workload channel and an overall workload value. Figure 1 provides the total workload graph output for the H-R scenario from the initial start until the end of the first victim assessment. Over this time period, the workload changes correspond to the demands of individual tasks.

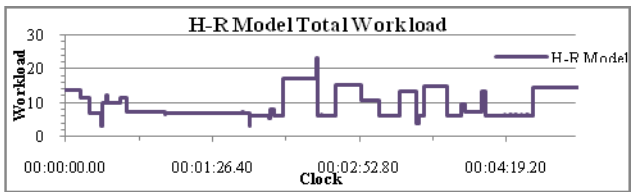

Figure 1. Workload output for the H-R task-first victim.

4. METHOD

It is necessary to conduct an empirical validation of the modeled workload results. The model results need to be compared to workload metrics from human user evaluations for both conditions, H-H and H-R. The model development and evaluation *hypothesis* was that a measurable difference in workload across the validation conditions exists. The purpose of the evaluation was intended to understand the workload differences between the two conditions and to understand how well the workload HPMF IMPRINT Pro model predicted actual human workload. The evaluation required an uninjured victim to form an ad-hoc team with either a remote human first responder, in the H-H condition or with local mobile robot in the H-R condition. The uninjured human received triage instructions from and provided necessary responses to the teammate while completing the triage steps.

4.1 Experimental Design

The experimental design was a mixed design with the participants as a random element. The experimental condition (H-H vs. H-R) differed between-subjects and the within-subject element included the series of triage victim assessment tasks. The independent variables were the experimental condition, the victim triage levels, the triage round (i.e., either the initial triage or the follow-up triage) and participant age, gender, experience with robots and first aid training. The dependent variables include both subjective and objective measures (see Section 4.3). The evaluation conditions corresponded to the two models (see Section 3). The H-H condition was completed prior to the H-R condition. During the H-H condition, an evaluator played the role of a first responder located outside of the contaminated area. The evaluator provided instructions to the uninjured victim – the participant. The H-R condition paired the participant with a robot. Both the participant and the robot were located in the contaminated incident area. A human evaluator supervised both the participant and the robot remotely.

4.2 Participants

Twenty-eight participants completed the evaluation, fourteen in each condition. All participants had at least some college education and were recruited by flyers around the Vanderbilt University area. Participant compensation was $15 for the approximately 90 minute evaluation. The participants were nearly evenly split by gender across the two conditions, with six males and eight females in the H-H condition and eight male and six females completing the H-R condition. The average age of all participants was 25.2 and age ranged between 18 and 57 years. The H-H condition mean age was 24.2 years and the H-R condition mean was 26.2. The participants rated their level of first aid experience on a Likert scale, with 1 representing no experience and 9 representing an expert level of experience. The average level of first aid experience was 3.6, with the H-H condition mean = 3.3 and the H-R condition mean = 3.9. All participants rated their level of robotics experience on the same scale. The average experience level was 2.7, with the H-H condition mean = 2.9 and the H-R condition mean = 2.6.

4.3 Evaluation Metrics

The objective metrics included: physiological data from a portable BioHarness ECG monitor [26] (HRV, breathing rate, beat-to-beat interval, heart rate, respiration rate, skin temperature, posture, vector magnitude data and acceleration data), time spent assessing each victim, correctness of responses to the secondary task questions, and accuracy of triage assessments. Subjective metrics included workload ratings collected after triaging each victim, post-experimental questionnaire responses and post-experimental NASA-TLX [27] responses. Due to space limitations, only on the HRV, heart rate, respiration rate, secondary task questions, in-task workload rating questions and NASA-TLX responses are reported.

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

After completing the triage steps for a particular victim, the participants ranked six workload channels on a scale from 1 (little to no demand) to 5 (extreme demand). The six workload channels were Cognitive, Auditory, Visual, Tactile, Motor and Speech. Each channel was defined during the first set of questions. The questions were adapted from the Multiple Resources Questionnaire [28] and the channels were chosen to facilitate comparison to IMPRINT Pro's seven workload channels. In order to prevent confusion, Imprint Pro's fine and gross motor channels were combined into a Motor channel

rating. When comparing the results to the predicted values, the fine and gross motor channels were added together. Each IMPRINT Pro's workload channel has a specified scale, minimum of 0 and maximum from 4 to 7. The provided workload responses were normalized to the corresponding IMPRINT Pro scale to facilitate comparison. The total workload for each victim assessment for the models was calculated using a time-weighted average of all workload values experienced while assessing the particular victim. These totals were compared directly to the re-scaled in-task subjective total workload results.

The NASA-TLX questionnaire was completed at the end of the entire evaluation [29]. This paper presents the results for the weighted overall NASA TLX score.

4.4 Experimental Environment

The evaluation occurred in the Center for Experiential Learning and Assessment within Vanderbilt University's School of Medicine. During the evaluation, the lights were dimmed and a background noise track created a more realistic environment. The background track incorporated explosion noises, street noise, people coughing and screaming, construction noise, and sirens. The volume was low enough that participants were able to clearly hear the teammate.

Six medical mannequins with various mechanical capabilities were distributed in the evaluation room. Table 1 details each mannequin's capabilities. Figure 2 shows the four of the six mannequins each identified by its victim number.

Table 1. Capabilities of each mannequin.

Mannequin	Age	Gender	Capabilities
Victim 1	Newborn	Female	• Crying
Victim 2	Adult	Female	• Pulse • Respiration Rate • Voice
Victim 3	Child	Male	• None
Victim 4	Adult	Male	• Pulse • Respiration Rate • Voice
Victim 5	Toddler	Female	• Pulse • Respiration Rate
Victim 6	Adult	Male	• Pulse • Respiration Rate • Blinking Eyes

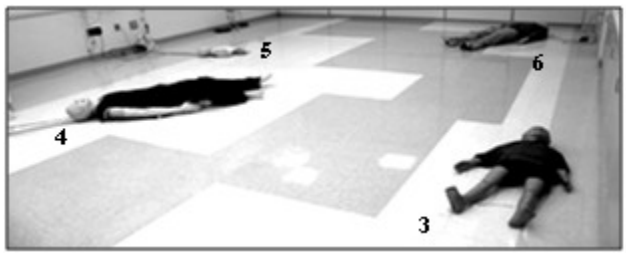

Figure 2. 4 of the 6 medical mannequins used in evaluation.

The Pioneer 3-DX robot teammate was equipped with a laser range finder and navigated the room autonomously on a pre-planned path [30, 31]. The robot's speech was scripted and controlled by the evaluator [30]. When the robot asked a question, the evaluator logged the response in the robot's script and moved the speech process to the next instruction.

4.5 Method

After completing initial forms and questionnaires, participants donned a BioHarness ECG monitor [26]. A baseline heart rate was measured, after which all heart rate channels were recorded continuously throughout the evaluation. Once the heart rate monitor was functioning properly, a script was read that introduced the disaster response scenario and informed the participant that he or she would be working with a teammate (either human or robot).

Each participant viewed a four minute video, setting the scene of a mass-casualty incident. The video showed scenes from David Vogler's live footage from the September 11th attacks in New York City [32]. After the video, the participant was instructed that his or her role was an uninjured, ambulatory and "contaminated" victim that is unable to leave the contaminated incident area until responders set up a decontamination area.

During the briefing, the participants assigned to the H-H condition were told that they had called 9-1-1, could not yet leave the contaminated area, that human responders were not permitted into the contaminated incident area, and asked if they would be willing to assist a human first responder to triage victims. Participants were told that they were transferred to a human first responder who would to lead them through the triage steps, record the participant's responses to questions, and the GPS location of the victims based on the participant's cell phone GPS signal. The participants identified which victim to treat next. The participants used a walkie-talkie with a headset and microphone (in place of a cell phone) to communicate with the remote human teammate - an evaluator acting as a first responder. The evaluator was in a remote location, from which he or she could not be seen or heard.

During the H-R condition briefing, participants were told they would be co-located with the robot because human responders were not permitted in the contaminated incident area. They were asked if they would be willing to work with the robot. The robot led the participants to the victims. The robot communicated with the participants using a digitally synthesized voice projected by a speaker mounted on the robot, while leading the participants through the tasks. The robot's speech was monitored and advanced by the remote evaluator. The participants wore a wireless microphone that transmitted responses to the voice interaction system and the evaluator. The participants were able to ask questions and in a Wizard of Oz manner, the remote evaluator either had the system repeat the robot's statement/question or provided a pre-programmed response.

The victims were positioned such that it almost forced the H-H condition participants to visit the victims in the same order as the H-R condition during the initial triage, see Figure 2. It was possible for participants to visit victims in a different order than planned during the H-H condition. If this occurred, usually a switch of Victims 3 and 4, the evaluator adjusted the script to assess the alternate order during the first round. During the follow-up triage, the first responder provided instructions to the H-H condition participants that guided them to the proper victim based upon the initial triage results and the GPS location collected from the participant's "cell phone."

The triage instructions provided and questions asked were identical across conditions. The teammate guided the participant through the steps to identify a victim's triage level. The participants in both conditions started at the same position in the room and moved from victim to victim during the initial triage (Round 1). After completing the initial triage of all six victims, the participant was led back to the five surviving victims for a second triage check (Round 2). During the second triage for the H-H condition, the next victim was specified by referring to the victim by the order in which they were first visited, for example, "please go to the first victim you triaged." The robot led the participant to the appropriate victim during the H-R condition. Upon reaching a victim, the teammate provided a summary and asked the participant to perform the triage assessment again. The H-R condition required the participant to place a color-coded triage card on the victim upon completing the triage. The cards were located on the robot platform and the robot instructed the participant which color card to choose. The H-H condition participants were simply told the victim's triage level. Table 2 details the mannequin settings for each victim, their age, expected triage level, the order each victim was visited, and the type of symptoms for each victim by triage round. Respiration rate is represented as breaths per minute (bpm). Note that the victim triage order during the second triage round was ordered by the most severe triage level after the initial triage.

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

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

All victims were triaged during Round 1. The third victim was not triaged during Round 2 because this victim was classified as Expectant during the initial triage. Four of the five remaining victims' triage levels changed prior to the second triage. The secondary task required the participants to memorize a list of names (see Section 4.3). Throughout all triage tasks, the participants were asked a question related to the list of names.

After triaging a victim, participants responded to the subjective workload questions. After completing the questions, the participant proceeded to the next victim. Upon evaluation completion, a post-experimental questionnaire and the NASA-TLX workload questionnaire were completed. A second baseline heart rate was collected and the heart rate monitor was removed after questionnaire completion.

5. RESULTS

5.1 Physiological Measures

The HRV was analyzed by victim, triage level and condition. The overall H-H mean was 13.01 (St. Dev. = 12.56) and the H-R mean was 12.70 (St. Dev. = 14.77). A t-test found no significant difference across conditions for overall HRV. The mean HRV by victim and triage round are plotted in Figure 3, error bards represent one standard deviation above and below the mean. The x-axis represents the victim assessed and is abbreviated by the victim's number and round number. Mean (M.) and standard deviation (St. Dev.) for HRV, heart rate and respiration rate by triage level and condition are provide in Table 3.

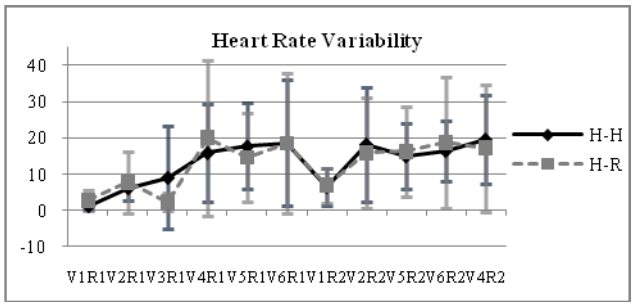

Figure 3. Mean HRV by victim and condition.

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

The heart rate descriptive statistics by condition and triage level are provided in Table 3. A t-test found that the H-H condition had significantly higher heart rate, $t(306) = 3.59$, $p < 0.01$. A two-way ANOVA assessed the main effects and interaction of both condition and triage level, with heart rate as the dependent variable and both condition and triage level as independent variables. Results showed that H-H heart rate was significantly higher than that of the H-R condition, with $F(302,1) = 12.78$, $p < 0.01$. There was no main effect of triage level and no interaction effect of triage level and condition.

The respiration rate descriptive statistics, by condition and triage level are also provided in Table 3. A t-test found that the H-H condition had a significantly higher mean respiration rate, $t(306) = 2.65$, $p = 0.01$. A two-way ANOVA assessed the main effects and interaction of both condition and triage level, with respiration rate as the dependent variable and both condition and triage level as independent variables. Results showed that the H-H respiration rate was higher than H-R, $F(302, 1) = 6.93$, $p < 0.01$. No main effect of triage level on respiration rate or interaction effect of triage level and condition were found.

Table 3. Descriptive Statistics for all physiological metrics.

Metric	Triage Level	Statistic	H-H	H-R
HRV	Delayed	M.	17.50	18.09
		St. Dev.	15.38	18.43
	Immediate	M.	12.57	12.71
		St. Dev.	10.82	13.81
	Expectant	M.	7.61	4.56
		St. Dev.	10.58	4.44
	Overall	M.	13.01	12.70
		St. Dev.	12.56	14.77
Heart Rate (beats per minute)	Delayed	M.	85.03	81.13
		St. Dev.	12.21	9.40
	Immediate	M.	85.20	80.42
		St. Dev.	12.00	10.27
	Expectant	M.	87.36	82.31
		St. Dev.	12.65	11.45
	Overall	M.	85.55	80.96
		St. Dev.	12.13	10.22
Respiration Rate (breaths per minute)	Delayed	M.	18.88	18.01
		St. Dev.	3.69	2.64
	Immediate	M.	19.13	17.71
		St. Dev.	3.91	3.25
	Expectant	M.	19.01	18.69
		St. Dev.	4.50	3.25
	Overall	M.	19.04	17.97
		St. Dev.	3.94	3.10

5.2 In-Task Subjective Workload Ratings

The individual channel subjective workload ratings gathered at the completion of each victim assessment were combined into a total workload value. The overall mean for H-H workload was 16.26 (St. Dev. = 6.31), while the H-R workload mean was 13.48 (St. Dev. = 4.80). A t-test indicated that the H-H condition rated workload higher, $t(306) = 4.35$, $p < 0.01$. The mean H-H condition workload for Delayed victims was 19.31 (St. Dev = 5.94) and 15.02 (St. Dev. = 4.70) for the H-R condition. The mean for the H-H Immediate victims, was 16.06 (St. Dev. = 5.89) and for H-R was 13.48 (St. Dev. = 4.60). The Expectant victims H-H mean was 15.21 (St. Dev. = 5.97) and the H-R mean was 11.14 (St. Dev. = 4.74). Figure 4 provides the total workload for each condition at each victim assessment point.

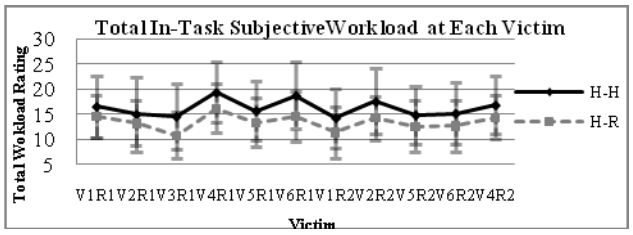

Figure 4. Total workload by victim and condition.

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

5.3 Secondary Task Question Correctness

The number of correct answers to secondary task questions was compared between conditions. Thirteen questions (Q.) were asked in total – one during the introduction to the task (Q. 1), one during the triage of each the victim during Round 1 (Q. 2–7), one between the two rounds (Q. 8), and one during the triage of each victim during the second round (Q. 9-13). Overall, the mean number of correct responses was 12.71 (St. Dev. = 0.61) during the H-H condition and 12.43 (St. Dev. = 0.65) for the H-R condition. T-tests across conditions found no significant difference. Analysis was conducted based upon triage level without any significant results. The division of correct answers by condition is provided in Table 4.

Table 4. Average number of correct responses by triage level

Triage Level	Statistic	H-H	H-R
Delayed Q. 5, 7, 10	M.	2.93	2.86
	St. Dev.	0.27	0.36
Immediate Q. 2, 3, 6, 11, 12, 13	M.	5.86	5.76
	St. Dev.	0.36	0.43
Expectant Q. 4, 9	M.	1.93	1.93
	St. Dev.	0.27	0.27
Overall	M.	12.71	12.43
	St. Dev.	0.61	0.65

5.4 NASA-TLX

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

5.5 Correlations Analysis

A partial Pearson's correlation was performed to analyze the correlation between HRV, heart rate, respiration rate and the total in-task subjective workload rating from each victim while adjusting for the independent variables of victim being assessed and victim triage level. Across both conditions, in-task subjective workload ratings were significantly negatively correlated to respiration rate, $r(290) = -0.15$, $p = 0.01$ and had a significant positive correlation to heart rate, $r(290) = 0.16$, $p < 0.01$. The correlation between HRV and subjective workload ratings was nearly significant with $r(290) = 0.10$, $p = 0.10$. The literature [12-15] implies that these three physiological measures may be able to represent workload. Since the physiological metrics were correlated to the in-task subjective workload ratings, the trends shown by these three physiological measures can be considered when assessing the difference in workload between conditions. The literature also reports a positive correlation between both HRV and heart rate and workload, and a negative correlation between respiration rate and workload.

5.6 Model Analysis

The participants' in-task subjective workload ratings and the average workload for each victim predicted by the IMPRINT Pro model were compared. Model total workload was calculated

by adding together each workload channel. The empirical total workload was calculated by rescaling each of the channel results to the IMPRINT Pro's workload channel scales and then totaling the results at each time point. Figure 5 compares the H-H condition results compared to the H-H model workload results. Figure 6 provides the results for the H-R condition. As can be seen in the figures, the majority of the empirical workload data points by victim and round were higher than the predicted model values. The calculated difference between the modeled workload values and the in-task subjective workload values demonstrate how effective the models were at predicting human behavior. The average difference between the H-H subjective values and the model results at each time point was 3.23 (St. Dev. = 2.03). The mean delta between the H-R condition and the model was 2.64 (St. Dev = 2.01). These data imply that the H-R model was slightly closer to empirical results than the H-H model.

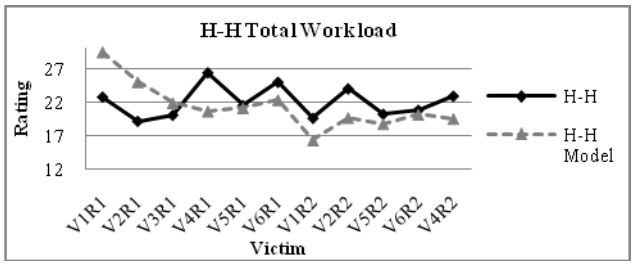

Figure 5. Total workload: H-H model and H-H condition.

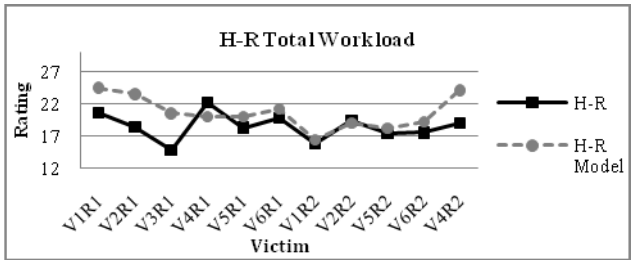

Figure 6. Total workload: H-R model and H-R condition.

6. DISCUSSION

The research hypothesis was that a difference between the H-H and H-R condition workload values exists. This hypothesis was partly supported based upon the empirical validation and comparison to the IMPRINT Pro model of HRT workload values. The H-R condition resulted in a slightly lower workload level, apparent in the model predictions and the subjective and physiological measures of workload.

Both HRV and in-task subjective workload measures indicated a main effect of triage level on workload, showing that the evaluation's controlled manipulation of workload did affect the participants' physiological responses and subjective ratings.

There was a main effect of condition on heart rate, respiration rate and in-task subjective workload ratings. All of these results indicated that H-H workload was higher than H-R. This difference in workload was corroborated by the NASA-TLX results, showing that while not significant, the H-H workload ratings tended to be higher than the H-R ratings. The models' predictions align with subjective empirical data, showing a lower level of workload in the H-R condition.

The IMPRINT Pro models predicted the trend in workload very similarly to the actual in-task subjective workload rating results. The models created using current models of workload and adjusted for the slight differences in the H-H and H-R scenarios. Overall, the models provided a valuable tool for prediction of workload. The data imply that the H-R condition participants experienced a lower level of workload than participants in the H-H condition, but the models more accurately predicted the H-R subjective workload. Current workload HPMFs may be applicable to human-robot peer based teams with the understanding that H-R teams may experience slightly less workload than H-H teams; however, additional analysis of more complex relationships and tasks are planned and required.

There are currently two working hypotheses as to why the H-R condition resulted in lower workload. One working hypothesis is that the embodiment of the robot may directly result in lowering the human's workload during the H-R condition. The second working hypothesis is that the robot's slower movement speed from victim to victim and slower interaction with the participant during the triage tasks may result in a lower workload for the H-R condition participants. These hypotheses will be explored in a planned evaluation that requires the participant to complete joint, peer-based decision making and task activities with either a human or a robotic partner.

7. CONCLUSION

The research studied two teams, one human-human and one human-robot, performing the same series of medical triage tasks. The workload HPMF for both conditions was modeled using IMPRINT Pro. An empirical evaluation assessed workload for the corresponding conditions. Both the models and the empirical evaluation found that the human-robot condition generally results in lower workload than the human-human condition. As well, the empirical results generally mirror the model results for each condition, with the human-robot peer-based team model slightly overestimating the measured workload. This research provides initial support for the applicability of a current workload HPMF (for human-only teams) to human-robot peer-based teams. Further research is required incorporating more complex, shared decision making tasks before further generalizing the applicability of existing workload HPMFs to human-robot peer-based teams.

8. ACKNOWLEDGMENTS

The authors thank Matthais Scheutz, Paul Schermerhorn and David Bender for providing the DIARC architecture. We also thank Matt Weinger, Brad Immekus, Ray Booker, Andrew Cross, Thad Hoffman and Sean Hayes. This research is supported by AFOSR award FA9550-09-1-0108, National Science Foundation Grant IIS-0643100, and an Office of Naval Research MURI Program award N000140710749.

9. REFERENCES

[1] Scholtz, J. 2003. "Theory and Evolution of Human Robot Interactions," in *Proc. of IEEE 36th Int. Conf. on System Sciences*, 5 (Hawaii, 2003), 125 - 134.

[2] Goodrich, M.A., and Schultz, A.C. 2007. "Human-Robot Interaction: A Survey," in *Foundations and Trends in Human-Computer Interaction*. 1,3 (2007), 203-275.

[3] Katzenbach, J.R., and Smith, D. K. "Best of HBR 1993-The discipline of teams." *Harvard Business Review*. (July-Aug. 2005), 1-10.

[4] Silverman, B.G., Johns, M., Cornwell, J., and O'Brien, K. 2006. "Human behavior models for agents in simulators and games: Part I: Enabling science with PMFServ." *Presence: Teleoperators and Virtual Environments.* 15, 2 (2006). 139-162.

[5] Foyle, D. C., and Hooey, B.L. 2008. *Human Performance Modeling in Aviation.* Boca Raton: CRC/Taylor & Francis.

[6] Mumaw, R.J., Roth, E.M., Vicente, K.J., and Burns, C.M. "There is more to monitoring a nuclear power plant than meets the eye." *Human Factors*. 42, 1 (Spr. 2000), 36-55.

[7] Humphrey, C.M. and Adams, J.A. "Robotic Tasks for CBRNE Incident Response." *Advanced Robotics*, 23, (2009), 1217-1232.

[8] Chen, J.Y.C., Haas, E.C., Barnes, M.J. "Human performance issues and user interface design for teleoperated robots." *IEEE Transactions on Systems, Man and Cybernetics - Part C*. 37, 6 (Nov. 2007), 1231 – 1245.

[9] Murphy, R.R. "Human-robot interaction in rescue robotics," *IEEE Transactions on Systems, Man and Cybernetics - Part C*. 34, 2 (May. 2004), 138-153.

[10] Howard, A.M. "Role allocation in H-R interaction schemes for mission scenario execution." In *Proc. of IEEE Int. Conf. on Robotics and Automation*. (Orlando, FL, May 2006).

[11] Howard, A. M. "A systematic approach to predict performance of human- automation systems." *IEEE Transactions on Systems, Man and Cybernetics – Part C*, 37, 4 (2007), 594-601.

[12] Reimer, B., Mehler, B., Coughlin, J.F., Godfrey, K.M., and Tan, C. "An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers." In *Proc. of the First International Conference on Automotive User Interfaces and Interactive Vehicular Applications.* (Essen, Germany. Sep. 2009).

[13] Vincente, K.J., Thornton, D.C., and Moray, N. "Spectral analysis of sinus arrythmia: a measure of mental effort." *Human Factors*. 29, 2 (1987), 171-182.

[14] Aasman, J., Mulder, G., and Mulder, L.J.M. "Operator effort and the measurement of heart-rate variability." *Human Factors.* 29, 2 (1987). 161-170.

[15] Roscoe, A.H. "Assessing pilot workload. Why measure heart rate, HRV and respiration?" *Biological Psychology*. 34 (1992), 259-287.

[16] Gawron, V. J. *Human Performance, Workload, and Situational Awareness Measures Handbook. 2nd ed*. Boca Raton: CRC, 2008.

[17] NATO. Civil Emergency Planning Civil Protection Committee. *Guidelines for First Response to a CBRN Incident*. <http://www.nato.int/cps/en/natolive/topics_49158.htm?selectedLocale=en>.

[18] Benson, M., Koenig, K.L., and Schultz, C.H., "Disaster triage: START then SAVE – a new method of dynamic triage for victims of a catastrophic earthquake." *Prehospital and Disaster Medicine*. 11, 2 (Apr.-Jun. 1996), 117-24.

[19] "Simple Triage and Rapid Treatment (START)." *Community Emergency Response Team Los Angeles.* Web. 15 Sept. 2010. <http://www.cert-la.com/triage/start.htm>.

[20] "Disaster Preparedness & Response Network -Resources." Web. 15 Sep. 2010. <http://www.scahec.net/prepares/resources/media.html>.

[21] van Lent, M., Silverman, B.G., McAlinden, R., O'brien, K., Probst, P., and Cornwell, J. "Enhancing the Behavioral Fidelity of Synthetic Entities with Human Behavior Models," in *Proc. of 13th Conference on Behavior Representation in Modeling and Simulation,* (Arlington, VA, May 2004).

[22] Allender, L., Kelley, T. D., Salvi, L., Lockett, J., Headley, D. B., Promisel, D., Mitchell, D., Richer, C., and Feng, T. "Verification, validation, and accreditation of a soldier-system modeling tool." *Proc. of the Human Factors and Ergonomics Society 39th Annual Meeting*, (San Diego, CA., 1995), 1219-1223.

[23] Archer, S., Gosakan, M., Shorter, P., and Lockett III, J. F., "New capabilities of the army's maintenance manpower modeling tool," *Journal of the International Test and Evaluation Association*, 26, 1 (2005), 19-26.

[24] Allender, L. "Modeling human performance: impacting system design, performance, and cost." In *Proc. of Military, Government and Aerospace Simulation Symposium, 2000 Advanced Simulation Technologies Conference*. (Washington, D.C., 2000) 139-144.

[25] Wickens, C.D., Dixon, S., and Chang, D. "Using interface models to predict performance in a multiple-task UAV environment – 2 UAVs" *Technical Report for the Aviation Human Factors Division Institute of Aviation prepared for Micro Analysis and Design*. Apr. 2003.

[26] "BioHarness Data Logger and Telemetry System." Data Acquisition - *BIOPAC.* Web. 31 Aug. 2010. <http://www.biopac.com/bioharness-data-logger-telemetry-system-acqknowledge>.

[27] "NASA TLX Homepage." Human Systems Integration Division at NASA Ames. Web. 15 Sept. 2010. <http://humansystems.arc.nasa.gov/groups/TLX/>.

[28] Boles, D.B., Bursk, J.H., Phillips, J.B., and Perdelwitz, J.R., "Predicting dual-task performance with the Multiple Resources Questionnaire (MRQ)." *Human Factors*. 49, 1 (Feb. 2007), 32–45.

[29] NASA TLX - Online NASA-TLX Workload Measurement Tool. Web. 15 Sept. 2010. <http://tlx.playgraph.com/>.

[30] Scheutz, M., Schermerhorn, P., Kramer, J., and Anderson, D. "First steps toward natural human-like HRI." *Autonomous Robots*. 22, 4 (May. 2007), 411-423.

[31] Montemerlo, M., Roy, N., and Thrun, S. "Perspectives on standardization in mobile robot programming: The Carnegie Mellon navigation (CARMEN) toolkit." In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems,"* 2003, 2410-2414.

[32] Vogler, David. "Raw Video Footage / WTC 9.11.01." 2001. Web. 31 Aug. 2010. <http://davidvogler.com/911>