An Adaptive Supervisory-Based Human-Robot Teaming Architecture

By

Jamison R. Heard

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Electrical Engineering

August 31st, 2019

Nashville, Tennessee

Approved:

Dr. Julie A. Adams

Dr. D. Mitchell Wilkes

Dr. Terrence Fong

Dr. Matthew B. Weinger

Dr. Richard A. Peters

ACKNOWLEDGMENTS

Firstly, I want to express my appreciation and gratitude to my advisor Dr. Julie A. Adams for her guidance and emotional support throughout my doctoral studies. Her research expertise and high expectations pushed me to conduct high quality research and developed the fundamental skills to conduct such research. My doctoral committee members were also imperative in the formation of this doctoral thesis. I want to thank each member and express my appreciation for the constructive criticism and guidance.

I also want to thank the members of the Human-Machine Teaming Lab at Vanderbilt University and Oregon State University. I will always fondly remember their fellowship and support.

Lastly, I want to thank my parents, who curated my intellectual curiosity from an early age and believed in my abilities.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

List of Acronyms

Acronym	Meaning
EEG	Electroencphalogram
BOTH	The workload assessment algorithm trained on data from the Supervisory-based and Peer-based evaluations.
COMM	The NASA Multi-Attribute Task Battery's Communication Task.
CURRENT	The performance prediction model that estimates the current overall task performance.
fNIRS	Functional Near-Infrared Spectroscopy
H-H	The peer-based evaluation's human-human teaming condition.
H-R	The peer-based evaluation's human-robot teaming condition.
HR	Heart-Rate
HRV	Heart-Rate Variability
IMPRINT	Improved Performance Research Integration Tool
MATB	Multi-Attribute Task Battery
NL	Normal Load
NLVL	Noise-Level
OL	Overload
PEER	The workload assessment algorithm trained on data from the Peer-based evaluation.
PM	Posture Magnitude
POST-HOC	The workload assessment algorithm trained on data from the Real-Time Evaluation and was evaluated in a post-hoc fashion.

PREDICTED	The performance prediction model that predicts the overall task performance for 1-minute into the future.
RES	The NASA Multi-Attribute Task Battery's Resource Management Task.
RMSE	Root-Mean Squared Error
RR	Respiration Rate
RT	The workload assessment algorithm trained on data from the Real-Time Evaluation and was used to estimate workload in real-time.
SR	Speech-Rate
SUP	The workload assessment algorithm trained on data from the Supervisory Evaluation.
SYS	The NASA Multi-Attribute Task Battery's System Monitoring Task.
TRCK	The NASA Multi-Attribute Task Battery's Tracking Task.
UL	Underload
VI	Voice Intensity
VP	Voice Pitch

Chapter 1

Introduction

Supervisory human-machine teams incorporate humans commanding and monitoring robots' actions in order to complete a task and have been used in various environments, including extreme environments, such as space exploration and search and rescue. High task performance is imperative under such extreme conditions, as a mistake may cause significant monetary loss, mission failure, or loss of life. The importance of high task performance and costly mistakes place considerable stress and workload on the human supervisors, which can reduce task performance. A research goal of this dissertation is to demonstrate an adaptive human-robot teaming system capable of adapting, based on a complete estimate of the human's workload state in order to ensure high task performance.

Adaptive human-robot systems need to increase task performance by adapting to the human's workload level, since task performance decreases when workload is too high (i.e., overload) or too low (i.e., underload). Only a few adaptive workload systems exist, where the systems only adapt to high workload conditions, even though low workload conditions also can be as detrimental to task performance. Another limitation of current systems is that the system adaptation is based on one dimension of workload (i.e., cognitive or physical), even though tasks usually encompass multiple workload dimensions (e.g., a space exploration task may contain cognitive, auditory, speech, and visual).

The basis of the adaptive workload system is a real-time workload assessment algorithm that relies on objective metrics, such as physiological signals, to predict human workload levels. Typical workload assessment algorithms use machine learning to classify workload states (i.e., high or normal workload). Although, a discrete representation of the individual's workload state is important, adaptive workload systems need a continuous representation of the workload state to determine how much adaptation is needed in order to positively impact human performance. This dissertation uses an estimation based workload assessment algorithm to provide a multi-factored workload value, which can be used to determine if and how much adaptation is required in order to ensure team task performance.

The focus of this dissertation is to improve supervisory human-machine teams by developing an adaptive human-robot teaming system. The developed system relies on a real-time workload assessment algorithm and a performance prediction model. The results show that the workload assessment algorithm is capable of distinguishing between different workload levels across multiple workload components for two user evaluations. The developed algorithm was used in real-time for a proof-of-concept adaptive system that intelligently targeted interactions and improved task performance. This dissertation addresses the problems of workload estimation, individual differences, performance prediction, and intelligent interaction adaptation in human-robot teams.

Chapter II provides background information on workload theories, objective workload metrics, subjective workload metrics, and surveys the existing workload assessment algorithms and adaptive system architectures. Chapter III presents a diagnostic workload assessment algorithm and the associated post-hoc results from two user evaluations. Chapter IV examines the algorithm's real-time capabilities in a non-stationary supervisory-based environment, while Chapter V presents a proof-of-concept validation of the adaptive human-robot teaming architecture. Chapter VI presents conclusions and future directions.

Chapter 2

Related Work

Workload can be defined as the ratio of resources required to achieve tasks to the resources the human has available to dedicate to the task [126]. A high value of workload, or overload, occurs when a large amount of resources are required to complete assigned tasks, but only a small amount of resources are available [126]. The overload condition occurs when task demand increases, but the human has insufficient resources available to dedicate to the tasks, which causes task performance to decrease [133]. A typical task management strategy during the overload condition is to focus on tasks with higher importance or priority, while simultaneously shedding those tasks with lower importance [126].

A low workload value, or underload, occurs when a small amount of resources are required to achieve the assigned tasks, but a large amount of resources are available to dedicate to the tasks [126]. Underload can be very difficult to detect [41, 42], since as task demand increases, task performance can remain the same [133, 134]. Underload is frequently not a research focus, but can be as determinantal to task performance as overload [136]. The underload condition leads to reduced alertness, and vigilance as well as lowered attention [126].

McCraken and Aldrich [80] decomposed overall workload into four components or resource channels using the Visual, Auditory, Cognitive, and Psychomotor method. Mitchell [82] expanded upon this method to incorporate a speech workload component and split psychomotor into gross motor, fine motor, and tactile; thus, overall workload can be decomposed into seven components: cognitive, gross motor, fine motor, tactile, visual, speech, and auditory. This dissertation recombines gross motor, fine motor, and tactile into a physical workload component, as tactile and fine motor components are difficult to measure using objective metrics. Cognitive workload represents the difference of the total amount of mental processing resources available, relative to the amount the task requires [47]. Physical workload is defined as the amount of physical demands placed on a human when performing a task [47], while visual workload represents the demand when using the eyes to identify or separate objects [4]. Speech workload arises when a person uses their voice, while auditory workload represents demands to recognize words, tones, mood, and emotion through sound [15]. Each component varies its contribution to overall workload depending on the task requirements.

Understanding the overall workload state requires analyzing the workload components in order to target the system adaptations to the workload components that are contributing the most to the current workload state. For example, if the system is dependent on the overall workload value alone, it may reallocate a task that is not a primary contributor to the human's current workload state. If the human operator has a high physical task demand and a low cognitive demand, then adapting the system to reduce cognitive workload will have little impact on the physical demands and resolving the overall overload state. However, an adaptation that adjusts the physical workload demands can lower the overall overload state.

Chapters 2.1 and 2.2 reviews common workload metrics and the state-of-the-art workload assessment algorithms [50]. Chapter 2.3 presents the current theory surrounding adaptive system architectures and the state-of-the art systems.

2.1 Workload Metrics

2.1.1 Common Objective Metrics

The objective workload metrics focus on identifying workload levels and are viable options when developing workload assessment algorithms. The objective metrics are categorized as a physiological response or task demand, and both categories correlate to overall workload. Three established criteria are used to evaluate the metrics: sensitivity, diagnosticity, and selectivity [89]. Sensitivity refers to the metric's ability to reliably detect workload levels. Diagnosticity, in the context of the visual, auditory, cognitive, and psychomotor workload theory, refers to the metric's ability to "discriminate between different types of workload (e.g., visual versus cognitive workload)" [82]. Lastly, selectivity refers to the metric's ability to reject unrelated demands (i.e., emotional stress). Each metric is categorized as being conforming, non-conforming, or requiring additional evidence with the stated criteria, where conforming is defined as complying to the criterion. A metric conforms with a criterion if at least three citations indicate that the metric fits the criterion (i.e., three citations indicate that heart-rate is sensitive to workload). Likewise, a metric is non-conforming if at least three citations indicate that the metric does not fit the criterion (i.e., respiration rate has low sensitivity to workload). It is important to note that if a metric is classified as non-conforming for a criterion, the metric may still supply useful information to a workload assessment algorithm. If a metric does not have at least three citations for a particular criterion, then additional evidence is required. Table 2.1 lists the objective workload metrics and their associated response under high workload conditions for overall workload and the workload components along with the sensitivity, diagnosticity, and selectivity classifications. The category column determines if a metric is a response to a change in workload (**response**) (i.e., heart-rate variability), or a direct measurement of the task demand (**demand**) (i.e., posture sway). Metrics categorized as demand require evidence based on the response metrics in order to determine how well the demand metric meets the criteria.

Directly comparing workload metrics in Table 2.1 is challenging, as the metrics are evaluated in different task environments that may contain different workload levels. Ideally, the metrics are directly compared in multiple task environments with various workload compositions in the presence of unrelated demands; however, such an analysis is extremely challenging and infeasible. Thus, the classifications are provisional, but the classifications for highly studied metrics, such as heart-rate variability, are unlikely to change.

The **Electroencephalogram** (EEG) collects neurophysiological signals from different

Table 2.1: Workload Metrics Overview

Metric	Category	Correlation	Workload Component(s)	Sensitivity	Diagnosticity	Selectivity
EEG: Power Spectral Density	Response	Both	Cognitive	Conf	Conf	Conf
EEG: Event Related Potentials	Response	Increases	Cognitive	Conf	Conf	Conf
fNIRS	Response	Increases	Cognitive	Conf	Conf	Conf
Heart Rate Variability	Response	Decreases	Cognitive	Conf	Conf	Conf
Heart Rate	Response	Increases	Cognitive, Physical	Conf	Conf	Conf
Respiration Rate	Response	Decreases	Speech, Physical	Non-Conf	Non-Conf	Non-Conf
Galvanic Skin Response	Response	Decreases	Cognitive, Physical	Conf	Non-Conf	Non-Conf
Skin Temperature	Response	Decreases	Cognitive, Physical	Conf	Non-Conf	Non-Conf
Blink Frequency	Response	Both	Cognitive, Visual	Conf	Non-Conf	Non-Conf
Pupil Dilation	Response	Increases	Cognitive	Conf	Conf	Conf
Fixation Duration	Both	Increases	Cognitive, Visual	Conf	Conf	Conf
Blink Duration	Response	Decreases	Cognitive, Visual	Non-Conf	Non-Conf	Non-Conf
Blink Latency	Response	Increases	Cognitive, Visual	Conf	Non-Conf	Conf
Noise Level	Demand	Increases	Cognitive, Auditory	Conf	Conf	Conf
Speech Response Time	Response	Increases	Cognitive, Auditory, Speech	Conf	Conf	Conf
Speech Rate	Response	Increases	Cognitive, Speech	Conf	Conf	Conf
Number of Fragments	Response	Increases	Cognitive, Speech	Req Evid	Req Evid	Req Evid
Number of False Starts	Response	Increases	Cognitive, Speech	Req Evid	Req Evid	Req Evid
Number of Syntax Errors	Response	Increases	Cognitive, Speech	Req Evid	Req Evid	Req Evid
Filler Utterances	Response	Increases	Cognitive, Speech	Req Evid	Req Evid	Req Evid
Utterance Repetitions	Response	Increases	Cognitive, Speech	Req Evid	Req Evid	Req Evid
Utterance Length	Response	Decreases	Cognitive, Speech	Req Evid	Req Evid	Req Evid
Variance in Posture	Demand	Increases	Physical	Req Evid	Req Evid	Req Evid
Postural Load	Demand	Increases	Physical	Req Evid	Req Evid	Req Evid
Vector Magnitude	Demand	Increases	Physical	Req Evid	Req Evid	Req Evid
Task Density	Demand	Increases	Task Dependent	Req Evid	Req Evid	Req Evid
Task Switches and Interruptions	Demand	Increases	Task Dependent	Req Evid	Req Evid	Req Evid
Secondary Task Failure Rate	Demand	Increases	Task Dependent	Conf	Conf	Conf

Note: Conf = Conforming, Non-Conf = Non-Conforming and Req Evid = Requires Additional Evidence

brain regions. The power spectral density, specifically the alpha (8 - 13 Hz) and theta (4 - 8 Hz) frequency bands, and the event-related potentials are sensitive to a range of cognitive workload levels, when at least thirty seconds of data are processed (e.g., [18, 92, 113]). Event related potentials suffer from low signal-to-noise ratios and require a known stimulus. EEG signals are selective to cognitive workload, since they are affected by fatigue, anxiety, and emotional stress. EEG signals are also sensitive to muscle artifacts and may not accurately reflect cognitive workload when a participant is physically active. Recent research showed improved EEG signal accuracy during physical activity, but did not measure workload during movement [40]. Incorrect sensor placement can create inaccuracies. The EEG conforms with all criteria: sensitivity, diagnosticity, and selectivity.

Functional Near-Infrared Spectroscopy (fNIRS) measures blood oxygen levels in the brain and increases with cognitive workload (e.g., [53, 54, 105]). fNIRS is resistant to movement artifacts, unlike the EEG [105]. The metric is sensitive, diagnostic, and selective, (depending on the specific sensor equipment used and the human's activity [130]); thus, fNIRS conforms with all of the specified criteria.

Heart rate variability (HRV) measures the variation in the heart rate's beat-to-beat interval and is sensitive to large cognitive workload variations, but requires thirty seconds to two minutes of data (e.g., [1, 65, 121]). Sensitivity decreases when using less than thirty seconds or more than two minutes of data [22]. HRV is sensitive, diagnostic and selective, if an individual's skills, training, fatigue, and distractions remain constant.

Heart rate (HR) refers to the number of heart beats per minute and increases with physical and cognitive workload, but is only sensitive to large changes (e.g., [22, 43, 66]). Heart rate is diagnostic if a task contains only cognitive or physical workload and requires thirty seconds to two minutes of data. Heart rate is selective if an individual's skills, training, and fatigue remain constant [22]; however, sensor movement can impact the metric's accuracy. HR conforms with all criteria.

Respiration rate (RR) represents the number of breaths taken per minute. Respiration Rate decreases as speech and physical workload increases, but is not a good predictor of workload when used independently (e.g., [68, 99]). Respiration rate may be sensitive to the number of tasks performed [85], but is not diagnostic nor selective, as physical movement can decrease the metric's accuracy [22]. Respiration rate does not conform with the criteria.

Galvanic Skin Response is the conductivity of the skin [88, 120], which decreases as cognitive and physical workload increases (e.g., [88, 110, 120]). Galvanic skin response is not diagnostic or selective, since the metric is impacted by sweating and changes in the sympathetic nervous system (i.e., emotional and physical behavior). Galvanic skin response conforms with sensitivity.

Skin Temperature decreases as physical or cognitive workload increases (e.g., [81, 83, 84]). Skin temperature is sensitive to workload variations, but is not diagnostic or selective, as fatigue and changes in the sympathetic nervous system impact skin temperature [91]. Skin temperature conforms with sensitivity, but not with diagnosticity and selectivity.

Blink frequency or blink rate, captures the number of blinks per minute. An eye-tracker or electrooculogram can measure blink frequency, which has been shown to increase with higher cognitive workload and decrease with higher visual workload (e.g., [20, 78]). Blink frequency is sensitive to medium to large changes in workload, but is not diagnostic. There are considerable individual differences in blink frequency, which is not selective, as the metric is highly correlated with fatigue. Blink frequency is conforming with sensitivity, but is non-conforming with diagnosticity and selectivity. Blink frequency may be unreliable for longer duration tasks [20, 22].

Pupil dilation, or pupillometry, is the change in pupil diameter, which increases as cognitive load increases (e.g., [3, 20, 71]). Pupil dilation is highly sensitive to small workload variations. Pupil dilation is also diagnostic and selective, if the amount of light in the environment is remains constant. Lighting changes can greatly impact the metric's sensitivity to workload. Pupil dilation conforms with all criteria. Cheaper eye-tracking devices may not be precise enough to measure pupil dilation for cognitive workload assessment, as pupil dilation requires precise measurements, on the order of tenths of a millimeter [20].

Fixation Duration represents the number of eye fixations during a defined period (e.g., [6, 78]). A fixation occurs when a human stares at an object longer than a predetermined time. A larger number of fixations correlates to higher cognitive workload, but this metric is limited to high density visual display environments containing multiple objects. Fixation duration is sensitive and diagnostic. Fixation duration is selective on an individual basis, since the metric is dependent on scanning strategies and the visual display environment. The metric conforms with all criteria.

Blink Duration is measured as the length of a blink and decreases as cognitive and visual workload increase (e.g., [20, 22, 78]). Blink duration is not sensitive or diagnostic. This metric is selective, if fatigue levels remain constant; thus, blink duration is not suitable for long duration tasks. Blink duration is non-conforming with all criteria.

Blink Latency represents the time between consecutive blinks, which increases as cognitive and visual workload increases (e.g, [20, 22, 78]). Blink latency is sensitive, but is not diagnostic. Fatigue does not impact blink latency; thus, the metric is selective. Blink frequency conforms with selectivity and selectivity, but not diagnosticity.

Noise Level correlates to an increase in auditory and cognitive workload (e.g., [17, 46, 86, 114]). The metric's sensitivity is dependent on the amplitude, variability, duration, and intermittency of the task environment's noise level [26]. Noise level is diagnostic and selective, since the metric is a measurable task demand. Noise level conforms to all criteria.

Speech Response Time is the amount of time required to respond to an auditory stimulus [8]. A longer response time represents a higher level of cognitive, auditory, and speech workload. Speech response time is sensitive to workload variations, diagnostic, and selective (e.g., [8, 24, 87]). Thus, speech response time is conforming with all criteria. Accurate measurement requires known times for each stimuli, which can be difficult to obtain if the stimulus is not computer generated.

Speech Rate captures the articulation and pause rate of verbal communication, which affects the listener's speech and cognitive workloads (e.g., [7, 8, 63]). Speech rate is sensitive, diagnostic, and selective; thus, speech rate conforms with all criteria. Individual differences are vast; thus, speech rate is difficult to standardize across a population.

Number of Fragments, False Starts and Syntax Errors each increase as cognitive and speech workload increases. Fragments represent a sentence that does not complete a thought, while false starts are incomplete sentences. The metrics are sensitive to workload variations, diagnostic, and selective. Although, the speech context can effect the metric's sensitivity (e.g., [8, 63]); however, additional evidence is required. Recognizing the number of fragments, false starts, and syntax errors is not trivial [8]. Speech recognition software, such as Dragon® Naturally Speaking, require low noise environments and/or training, but recent advances in microphones (e.g., bone microphones) [93] and speech recognition software can overcome this limitation [5].

Filler or Delay Utterances are words, such as “um” or “you know”, that are meaningless in the context of a sentence. Filler utterances increase as cognitive and speech workload increase and are sensitive to workload variations (e.g., [8]). Filler utterances are diagnostic and selective; however, additional evidence is needed. Automatic detection of filler utterances is non-trivial and requires tracking recently spoken words.

Utterance repetitions increase as cognitive and speech workloads increase and are sensitive, diagnostic, and selective (e.g., [8]). Additional evidence is required to substantiate claims. Speech recognition software can be used to detect repetitions [70].

Utterance Length is the length of time between long pauses and can be measured via speech recognition software or speech-envelope detection. Utterance length decreases as speech and cognitive workload increases and is sensitive, diagnostic, and selective (e.g., [8, 74]). However, additional evidence is required.

Posture sway, or the **Variance in Posture** captures the change in the human’s center of gravity and is calculated by determining the mean squared deviation from the mean postural position, which increases as physical workload increases [72]. Posture sway is sensitive to workload variations, diagnostic, and selective; but only if the task requires posture changes. Additional evidence is required to substantiate claims. A task with physical workload, such as walking, may result in low variance in posture.

Postural load represents the time during which a participant’s trunk is flexed more than 45° and requires a known time period for framing the postural load value [94]. Increasing postural load increases physical workload; thus, the metric’s sensitivity, diagnosticity, and selectivity depends on the task. Additional evidence is required to determine if the metric is sensitive to task’s that requires flexing of the trunk.

Vector Magnitude represents physical movement [112] and refers to the magnitude, or mathematical size, of the walking vector. High vector magnitude levels indicate high levels of walking, a factor in physical workload. Vector magnitude is only relevant to tasks incorporating walking; thus, the metric's sensitivity, diagnosticity, and selectivity are task dependent. Additional research is required to substantiate claims.

Task Density calculates the number of tasks initiated during a specific time period [124] and increases overall workload. This metric is sensitive, diagnostic, and selective, if the task's workload composition is known. Task density requires maintaining consistent and comparable measurements throughout the task, which can be difficult. Task density needs additional evidence to determine if the metric conforms with the criteria.

The **Number of Task Switches or Interruptions** in a specified time period increases overall workload (e.g., [77, 103]. The timing of the task switch or interruption impacts overall workload as well. The number of task switches and interruptions are sensitive and selective, but not diagnostic. This metric requires additional evidence to determine if the metric conforms with the established criteria.

Secondary Task Failure Rate refers to the incorrect completion of a secondary task and is used to measure spare workload capacity. Secondary tasks include constant monitoring tasks (e.g., verbal recital of a word) or discrete prompted tasks, such as card sorting, memory recall, or mental math problems [37]. Secondary task failure rate increases as overall workload increases and is sensitive, diagnostic, and selective, depending on the secondary task's workload composition (e.g., [20, 22, 37]). This metric conforms with each criterion. The metric's sensitivity also depends on the individual's workload management strategies and learning effects, if the secondary task is usually not performed [22].

2.1.2 Common Subjective Workload Metrics

Subjective workload metrics provide insight into the human's perceived workload [37] and are typically inexpensive and easily administered [21]. Subjective metrics are infeasible for real-time workload assessment, since they do not output a continuous value. Further, subjective metrics encounter difficulty distinguishing between task difficulty and workload and cannot assess the unconscious processing of information that humans cannot rate [89].

The **NASA Task Load Index** (NASA-TLX) defines workload as a weighted mean of subjective ratings along six demand channels: mental, physical, temporal, own performance, effort, and frustration (e.g., [20, 37, 47]). The overall value is a score between 0 - 100, with 100 representing exceptionally high workload. A limitation is the time to administer the survey; however, auditory adaptations can be investigated. Additional limitations include a lack of continuous measurement throughout the trial and subjectivity. Participants may be unable to recall workload experienced during a task when providing responses after task completion. NASA-TLX conforms with sensitivity, selectivity, and diagnosticity.

The Cooper-Harper Scale evaluates air-craft handling using a decision tree based on the task, aircraft characteristics and workload demand [28] and has been shown to correlate with task performance and workload [37]. The primary limitation is that the tool is specific to the aircraft-handling domain, which was overcome by the **Modified Cooper-Harper Scale** that allows for the assessment of cognitive workload [127]. The modified tool rates mental workload instead of controllability and emphasizes the task difficulty [37]. The Modified Cooper-Harper scale is sensitive and selective, but not diagnostic (e.g., [22, 37, 127]). The modified tool conforms with sensitivity and selectivity, but not diagnosticity.

The less variable **Subjective Workload Assessment Technique** [97] measures cognitive workload using three different scales: time, cognitive effort, and psychological stress (e.g., [97]). The time component measures the amount of spare time available to dedicate to a task, while the cognitive effort component measures how much conscious mental effort is needed to complete the task. The psychological stress component measures the amount of risk, confusion, frustration, and anxiety associated with the task [37]. SWAT conforms with sensitivity, diagnosticity, and selectivity (e.g., [22, 37, 89]).

The **Multiple Resource Questionnaire** rates seventeen workload dimensions that encompass auditory, facial, memory, manual, spatial, tactile, visual, and vocal process on a rating scale ranging from 0 - 4 (e.g., [14, 20, 37]). The MRQ conforms with sensitivity, diagnosticity, and selectivity.

Verbal In-Situ Ratings assess six percieved workload components: auditory, visual, cognitive, speech, tactile and motor [45]. Each component ranges from 1 (little or no demand) to 5 (extreme demand). The limitations of verbal in-situ ratings include the lack of a continuous measure and the dependency on choice in administration time. Verbal in-situ ratings are sensitive, diagnostic, and selective; however, additional evidence is required.

2.2 Workload Aggregation Algorithms Approaches

Thirty-one workload aggregation algorithms across eleven task environments were identified and reviewed. An overview of the algorithms by task environment is provided in Table 2.2. The algorithms are classified using the following criteria: sensitivity, diagnosticity, suitability, and generalizability. Some of the algorithms in Table 2.2 are not described in this dissertation, but are described in Heard et al. [50]. Selectivity was not considered in the evaluation criteria, as none of the reviewed manuscripts analyzed the effect of unrelated demands (i.e., emotional stress or fatigue). The suitability and generalizability criteria were added in order to further assess an algorithm's viability for inclusion in an adaptive workload system. Each criterion is classified as conforming or non-conforming. An algorithm is classified as conforming if an algorithm meets all of the requirements specified for each criterion below, while a non-conforming classification is given otherwise. The requirements were chosen based on logical thought that is grounded in the literature [12, 82], as there exists no established criteria for comparing and critiquing workload assessment algorithms.

Sensitivity refers to an algorithm's ability to reliably detect workload levels, which depends on the algorithm's accuracy and levels of workload classified. An algorithm is classified as conforming if the algorithm detects at least three workload levels with $\geq$80% accuracy or $\leq$5 root mean squared error (RMSE). Detecting at least three workload levels was chosen as the threshold, as it is highly desirable that an algorithm detects both overload and underload conditions. Such a classification is not feasible with a binary representation; thus, an acceptable number of detected workload levels is three. Multiple factors deter-

Table 2.2: Summary of Workload Assessment Algorithms

Task Environment	**Paper**	**Sensitivity**	**Diagnosticity**	**Suitability**	**Generalizability**
Multi-Attribute Task Battery	Wilson and Russell[129] *	Conf	Conf	Conf	Non-Conf
	Christensen et al.[25] *	Req Evid	Non-Conf	Conf	Non-Conf
	Durkee et al. [31]	Non-Conf	Non-Conf	Conf	Non-Conf
Automated-Enhanced Cabin Air Management System	Wang et al. [123] *	Non-Conf	Non-Conf	Non-Conf	Non-Conf
	Ting et al. [118, 119]	Req Evid	Conf	Non-Conf	Non-Conf
	Zhang et al. [140] *	Conf	Non-Conf	Non-Conf	Non-Conf
	Zhang et al. [139]	Conf	Non-Conf	Conf	Non-Conf
	Yin and Zhang [135]	Conf	Non-Conf	Conf	Non-Conf
	Zhang and Wang [138]	Non-Conf	Conf	Non-Conf	Non-Conf
Flight or Driving Simulator	Hoogendoorn and van Arem [57]	Conf	Non-Conf	Non-Conf	Non-Conf
	Putze et al. [96]	Non-Conf	Non-Conf	Non-Conf	Non-Conf
	Oh et al. [90]	Req Evid	Non-Conf	Non-Conf	Non-Conf
	Besson et al. [9, 10] *	Conf	Non-Conf	Non-Conf	Non-Conf
	Wilson and Fisher [128] *	Conf	Non-Conf	Conf	Non-Conf
	Fan et al.[34] and Zhang et al. [141] *	Non-Conf	Non-Conf	Conf	Non-Conf
	Manawadu et al. [76]	Non-Conf	Non-Conf	Conf	Non-Conf
Remotely Piloted Vehicle	Rusnock et al. [101] *	Conf	Non-Conf	Conf	Non-Conf
	Borghetti et al. [16]	Conf	Non-Conf	Non-Conf	Non-Conf
	Durkee et al. [32]	Conf	Req Evid	Non-Conf	Non-Conf
	Durkee et al. [33]	Conf	Non-Conf	Conf	Non-Conf
Cognitive-Based	Zhang et al. [137]	Conf	Conf	Non-Conf	Non-Conf
	Hogervorst et al. [56] *	Non-Conf	Conf	Non-Conf	Non-Conf
	Massari et al. [79]	Non-Conf	Non-Conf	Non-Conf	Non-Conf
	Zhang et al. [142]	Non-Conf	Non-Conf	Conf	Non-Conf
Augmented Reality	Schultze-Kraft et al. [106] *	Non-Conf	Non-Conf	Non-Conf	Non-Conf
Normal Day	Ghosh et al.[38]	Non-Conf	Non-Conf	Conf	Non-Conf
Air-Traffic Control	Abbass et al. [2]	Req Evid	Conf	Non-Conf	Non-Conf
Misc. Tasks	Popovic et al. [95]	Non-Conf	Conf	Conf	Conf
Robot Surveillance	Teo et al. [115]*	Req Evid	Non-Conf	Conf	Non-Conf
Anomaly Detection	Zhao et al. [143]	Conf	Non-Conf	Conf	Non-Conf

Note: Conf = Conforming, Non-Conf = Non-Conforming and Req Evid = Requires Additional Evidence
* indicates a participant-specific algorithm

mine an algorithm's accuracy (i.e., validation method, human workload levels, difference between workload levels, overfitting), which make setting threshold values difficult. Thus, the classification accuracy thresholds were chosen based on the reviewed algorithms' accuracy distribution. The metrics incorporated in the algorithm and epoch size also impacts an algorithm's sensitivity; however, it is difficult to objectively analyze their impact on a reviewed algorithm's sensitivity without direct access to the algorithm. Further, an algorithm may use an insufficient epoch size for a metric or low sensitivity metrics, but the algorithm may still achieve a $\geq$80% accuracy. Thus, if an algorithm uses an insufficient epoch size or low sensitivity metrics, it will be noted in the algorithm analysis. An algorithm may require additional evidence for the sensitivity criterion, if there is insufficient information to classify the algorithm or the algorithm was not developed in a practical setting. For example, training an algorithm using randomly chosen folds or data segments may increase classification accuracy, due to the high correlations that exist between samples in time.

Diagnosticity refers to an algorithm's ability to "discriminate between different types of workload (e.g., visual vs. cognitive) [82]," but the majority of algorithms are designed to assess a single workload component. Another workload component's presence may confound the algorithm's workload assessment (i.e., presence of physical workload may confound the algorithm's cognitive workload assessment). It is difficult to objectively state how much the confounding component affects the workload assessment; however, diagnosticity is reliant on the workload metrics used. For example, if heart-rate is used for cognitive workload assessment, the presence of physical workload may greatly confound the cognitive workload assessment. An algorithm conforms with diagnosticity, if $\leq$20% of the metrics correlate to multiple workload components.

An algorithm's **Suitability** determines if the algorithm is capable of assessing the complete overall workload state imposed by the task environment in which it was deployed in. For example, a driving task imposes cognitive and visual workload; thus, the complete overall workload state is composed of said workload components. If an algorithm only assesses cognitive workload, then the algorithm only assesses a subset of the driving task's overall workload state. An algorithm can assess a workload component if the algorithm contains workload metrics sensitive to said workload component. An algorithm is conforming, if the algorithm assesses the complete overall workload state for the task and non-conforming otherwise. An algorithm is capable of assessing the complete overall workload state, if the algorithm contains metrics that correlate to the task's workload components.

Generalizability represents the algorithm's ability to generalize across tasks and populations. An algorithm generalizes across tasks, if the algorithm assesses the complete overall workload state (i.e., each workload component). An algorithm generalizes across populations if the algorithm is not participant-specific and achieves $\geq$80% accuracy, or $\leq$5 RMSE. An algorithm is classified as conforming if the algorithm generalizes across tasks and populations. There is a difficulty in classifying algorithms that are participant-specific, as there is no good measure to determine the population generalizability of such algorithms. However, the accuracy of participant-specific algorithms trained on a population will decrease, due to individual differences [144]. Thus, participant-specific algorithms can only receive a non-conforming rating, but will be denoted by an asterisk in Table 2.2.

It is worth noting that the criteria were developed to evaluate an algorithm's viability for use in an adaptive workload system, which may not have been the original authors' intended use of an algorithm. The algorithms were also evaluated based on their ability to assess the complete overall workload state, although the original authors only focus on a single workload component assessment (i.e., only cognitive). It is also difficult to compare directly the algorithms' performance, as the algorithms are not evaluated in standard environments. The ratings are still valuable, as the algorithms are compared in task environments with similar workload component compositions.

2.2.1 Overview of Machine Learning Classifiers

Workload assessment algorithms typically use machine learning to classify workload. Common machine learning classifiers include artificial neural networks, linear regression, linear discriminant analysis, fuzzy logic, support vector machines, model based, and ensemble.

Artificial neural networks mimic computation within the human brain and consist of at least three layers: input, processing, and output [12]. The input layer contains a node for each feature to be classified (i.e., heart-rate, respiration rate), while the output layer contains a node for each class (i.e., overload, medium workload). The processing layer contains a set of adaptive weights that are tuned by a learning algorithm and are used to determine the contribution of each metric for a class, based on the input and output data.

A linear regression classifier seeks to find a line, plane, or hyperplane that divides the input features into corresponding classes [12]. A simple two class, one feature equation takes the linear form $Y = B_0 + B_1 * X$, where B_0 and B_1 are weights and X is the input feature. The class of X is 0 if Y is below the line, or 1 if Y is above the line.

Linear discriminant analysis attempts to project the input features onto a smaller feature space that minimizes the in-class distance and maximizes the between-class distance [132].

Fuzzy Logic is a set of rules that formalizes human-reasoning [131]. Each feature is described as a membership function (i.e., temperature is high/low) in the range of [0, 1], where the number represents the degree of membership. The fuzzy rules use membership functions to determine the data's class. For example, determining if a person has a fever may use the rule: *IF (temperature is high), THEN (the person has a fever).*

Support vector machines map features to a feature space and identify the boundary between them. This boundary can be linear or non-linear, depending on the function used. The radial-basis function is a common non-linear function that creates circular boundaries.

A model-based algorithm creates a model tailored to the application domain that may or may not use machine learning algorithms [13].

An ensemble of classifiers aggregates machine-learning algorithms in order to reduce an individual classifier's variance and bias [29]. The classifiers can be the same or different.

2.2.2 Multi-Attribute Task Battery

The Multi-Attribute Task Battery (MATB) is an simulated aircraft monitoring system that contains optional subtasks (i.e., responding to communication requests) [27]. The supervisory-based task environment incorporates the cognitive, visual, auditory, and physical workload components.

Wilson and Russell [129] sought to assess workload in real-time using an artificial neural network. Seven participants completed tasks with two levels of difficulty. which was manipulated by varying the number of tasks to be completed within five minutes. The baseline condition required the participants to stare at a screen. The experiment included a single low workload condition task and a single high workload condition task. Six EEG electrodes, two electrooculography electrodes, and electrocardiogram physiological data were collected for each task. The power spectral density was calculated for five frequency bands pertaining to the EEG and electrooculography data, while heart rate variability, respiration rate, and blink frequency were calculated using the electrocardiogram and electrooculography data. A total of 43 features were used for classification. 75% of the physiological data was randomly segmented into ten second epochs, with 50% overlap and fed into an artificial neural network. The remaining physiological data served as test data for classifying the human's cognitive workload level as baseline, low workload, or high workload. The algorithm only assessed cognitive workload, although the task has a visual workload component. A participant-specific classifier was trained and evaluated in order to remove individual differences. The mean artificial neural network's accuracy was 84.3% across all participants; however, the algorithm's accuracy for individual participants ranged from 69% - 97%. Accuracy may increase if a larger epoch size is used, as heart-rate and heart-rate variability require at least thirty seconds of data to be sensitive to workload. An adapted aiding scenario removed subtasks when the neural network classified a high workload condition, which resulted in a 44% reduction in cognitive workload and a 33% reduction in task error. The reduction in workload and task error is expected, given that the adaptive aiding decreased the task density. The algorithm's **sensitivity**, **diagnosticity**, and **suitability** are classified as conforming, as the algorithm assesses three workload levels with 84.3% accuracy, <25% of the metrics correlate to multiple workload components, and the algorithm assesses each task workload component (i.e., cognitive and visual, respectively). The algorithm is classified as non-conforming for **generalizability**, as the algorithm is unable to assess each individual workload component.

Christensen et al. [25] expanded on Wilson and Russell's [129] results by focusing on the day-to-day variability of cognitive workload. The MATB was used with three task difficulty levels with eight participants working for five days, distributed over a month. Each day consisted of three five-minute tasks with random difficulty, resulting in a total of twenty-five tasks per participant. The physiological data recorded each day included 19 EEG channels, electroculography, heart rate, blink rate, and blink duration. The metrics were segmented into 40 second epochs, with a 35 second overlap. The artificial neural network was trained using individual participants' physiological data from the low and high workload tasks; thus, creating a participant-specific classifier. A neural network, linear discriminant analysis classifier, and support vector machine were evaluated using the same-day and inter-day tasks. The neural network had an accuracy of 99% when classifying low and high workload conditions for the same-day tasks and 83% for the inter-day tasks, while the linear discriminant analysis classifier and support vector machine achieved a day-to-day accuracy of 65% and 68%, respectively. When the classifiers were trained on a sub-set of days ranging from two to five days, the additional days increased the classifiers' accuracy and day-to-day generalizability. The classification accuracy across days also improved when the classifiers were trained on the first day's data and 2.5 minutes of data at the start of each day, although the increase was not significant. Even though the algorithm achieved high same-day accuracy with the neural network, the decrease in inter-day accuracy and the large number of features (i.e., 198), suggest that overfitting of the data occurred. Thus, the algorithm's **sensitivity** is classified as "needs additional evidence", as the algorithm may not perform well in a practical setting. **Diagnosticity** and **generalizability** are classified as non-conforming, since visual workload impacts 60% of the metrics. The algorithm conforms to **suitability**, as the algorithm assesses the task's workload components.

The Airforce Research Laboratory and Aptima, Inc designed a model-based classifier for cognitive workload with a range of 0 to 100, synonymous with the NASA-TLX score [31]. Seven participants completed fifteen five-minute MATB tasks each containing three task difficulty levels spanning from low to high workload, for a total of fifteen task difficulty levels. EEG, electrocardiogram, pupil diameter, blink rate, and fixation duration workload metrics were collected with an epoch of 5 seconds. A continuous workload model used for training and validation was developed by injecting noise into the NASA-TLX scores, where the amount of noise was dependent on the task context (i.e., task difficulty and number of tasks). The authors justify the noise, because the NASA-TLX is a static measure, but operator states are non-static. Thus, the need for injecting noise. The continuous workload model changed over time by dynamically updating the model weights using physiological, situational, behavioral, and human input. An unspecified machine-learning algorithm was trained on the workload model and classified cognitive workload in the range of 0 - 100, which produced a mean absolute difference of 35%. The high error is attributed to individual differences. The authors also examined the algorithm's ability to classify low and high cognitive workload, which produced an accuracy of 76%. The following indexes are based on the binary classifier, rather than the continuous classifier. The algorithm's **sensitivity** and **diagnositicity** are non-conforming, due to the <80% classification and the metrics used correlate to cognitive and visual workload. The algorithm's **suitability** is conforming. **Generalizability** is classified as non-conforming, since the algorithm does not assess each workload component.

Christensen et al.'s [25] algorithm is the most suitable for assessing workload in the MATB task environment and achieves the highest classification accuracy. However, the algorithm may not be practical in real-world situations, due to the presence of overfitting.

2.2.3 Automated-Enhanced Cabin Air Management System

The Automated-Enhanced Cabin Air Management System is a supervisory based task incorporating cognitive and visual workload [75]. The task environment requires participants to monitor oxygen level, while keeping different subsystems within range.

Wang et al. [123] utilized a neural network with ant colony foraging to classify cognitive workload. Eleven participants completed tasks. The number of subsystems monitored (1 - 5) was used to manipulate task difficulty and task performance was determined by the time in range value, or the percentage of time the system was in range for two seconds. Participants subjectively assessed workload (ranging from 0 to 100) after each task via an onscreen questionnaire. The subjective workload results were used for classifier training, while validation combined the time in range and subjective values using the formula

$$OFS = 0.7xTIR + 0.3x(100 - effort),$$

where OFS represents the workload level in the range of 0 - 100, TIR is the time in range as a percentage, and effort is the subjective assessment value. The formula weights were determined by maximizing the linear correlation between the OFS value and the EEG index of workload. Thirty-two EEG channels, heart rate, and heart rate variability data were collected, using an epoch of two seconds for each metric. The task load and engagement indexes were calculated from the EEG data and used as additional algorithm metrics. The workload metrics were used as features for a feed forward neural network trained by using the ant colony-based adaptive differential evolution algorithm, which uses foraging techniques to find the optimal weights. The trained neural network estimated the operator functional state and was cross validated using the leave-one-out method. The average RMSE of the algorithm's estimated vs. the modeled operator functional state ranged from 9.1 to 16.4 for all participants. The algorithm's **sensitivity** and **diagnosticity** are non-conforming, given the >5 RMSE and that the presence of physical workload may confound the workload assessment. The algorithm's **suitability** and **generalizability** are also non-conforming, as the participant-specific algorithm is unable to assess all of the task's workload components. The algorithm may achieve better results if a longer epoch time is used, as a two second epoch is insufficient for heart rate and heart-rate variability.

Cognitive workload was classified using fuzzy logic [118, 119]. Nine participants controlled up to five sub-systems during a four-hour session that included four tasks. The first task was an Automated-Enhanced Cabin Air Management System baseline adjustment period, which collected baseline physiological data used to normalize each participant's values. The second and fourth tasks incorporated fuzzy logic adaptive control, while the third task triggered adaptation based on system errors (i.e., a simulated oxygen level outside the defined operating range). The fuzzy modeling logic depended on the EEG task load index and heart rate variability metrics, with an epoch of one-hundred and twenty-eight seconds. The model's output was a time in range percentage predictor for the oxygen level ranging from low, a small time, to very high, a large time. A low output equated to higher workload, which caused the system to assume more autonomy. The fuzzy modeling controller improved operator performance and reduced the reported anxiety, fatigue, and effort levels. The algorithm's accuracy was not analyzed, only the performance outcome on the time in range value; thus, the algorithm's **sensitivity** requires additional evidence. The **diagnosticity** is conforming, since EEG and heart-rate variability only correlate to cognitive workload. The algorithm's **suitability** and **generalizability** are classified as non-conforming, as the algorithm is unable to assess visual workload.

Zhang et al. [140] developed a fuzzy clustering algorithm. Eleven participants completed two task sessions, consisting of nine task load conditions, each lasting 15 minutes. The task load was determined by the number of subsystems to be controlled. Two secondary tasks incorporated tank leveling and alarm response. Thirty-two EEG channels, heart rate, and heart rate variability metrics were recorded with an epoch of 1 minute. All metrics were normalized to the range of [0, 1] and the EEG task load index was computed. The metrics were fed into a fuzzy c-means clustering classifier, with a preset c of three, corresponding to good (low workload), average (medium workload), and poor/risky (high workload) operator performance. The fuzzy c-means classifier was participant-specific and had a mean accuracy of 85.7%. The algorithm conforms to **sensitivity**, due to the high accuracy and detecting three workload levels. The **diagnositicity** is non-conforming, as the presence of physical workload may confound the algorithm's cognitive workload assessment. The algorithm's **suitability** and **generalizability** are also non-conforming, as the algorithm is unable to assess visual workload and is unable to generalize across tasks.

This prior work [140] was extended by using a bias [135] and a bounded support vector machine [139]. The bias support vector machine used the thirty-two EEG channels, heart rate, heart rate variability, and electrooculography metrics, with non-overlapping epochs of 30 seconds. An adaptive exponential smoothing function and a local preservation technique was applied to improve the support vector machine's accuracy and reduce the feature space's dimensionality. Adaptive exponential smoothing uses the standard exponential smoothing formula, but adaptively chooses the value of λ based on if there is a large or small shift in the data. After adaptive exponential smoothing and feature reduction, the new feature matrix, $a_0, a_1, ..., a_{i-1}$, was used to train and verify the bias support vector machine's accuracy, which classified cognitive workload as baseline, normal, and high. The bias support vector machine achieved a 95% accuracy with adaptive exponential smoothing and feature reduction for individual participants, as compared to an accuracy of 88% without adaptive exponential smoothing and feature reduction. The bias support vector machine was generalized across all participants and achieved a 92% accuracy, with ten-fold cross validation. The adaptive exponential smoothing may explain the high accuracy of the generalized algorithm and is worth investigating for real-time domains. The algorithm's **sensitivity** is conforming, given the high accuracy and detecting three workload levels. The **diagnosticity** is non-conforming, since the presence of physical workload may confound the cognitive workload assessment. The algorithm's **suitability** is conforming, as the algorithm is capable of assessing all of the task's workload components. The algorithm's **generalizability** is non-conforming, given that the algorithm is capable of assessing two workload components: cognitive and visual.

Zhang et al.'s [137] bias support vector machine is the most suitable for workload classification in the Automated-Enhanced Cabin Air Management System task environment and has the highest classification accuracy. The algorithm has acceptable population generalizability, which may be attributed to the feature smoothing. Further research is needed to determine if adaptive exponential smoothing can improve generalizability for other machine-learning based algorithms.

2.2.4 Flight or Driving Simulator

Flight or driving simulators are operator-based environments that incorporate the cognitive, physical, and visual workload components. If the task requires the participant to listen and respond to a stimulus, speech and auditory workload components may be incorporated.

A support vector machine was used to classify cognitive workload in a driving simulator [96]. Eighteen participants completed a primary task and two variable secondary tasks, where the primary task was lane changing on a three lane highway without any traffic. The secondary tasks vary, with difficulty corresponding to low, medium, and high workload and are strongly correlated with the NASA-TLX scores, although the NASA-TLX scores are unreported. The first secondary task was a visual search task, which required identifying distractors with a different line thickness from the others. The difficulty of the visual search task varied based on how evident the distractor was different from the others. The other secondary task, a mathematical cognitive task, presented a list of numbers and required determining if the last number was divisible by a divisor. Task difficulty varied by changing the divisor from 2 (low difficulty) to 7 (high difficulty). EEG, skin conductance, heart rate, and respiration rate metrics were collected during the first minute of each task, with an unknown epoch size. The physiological data was pre-processed and used as features for a radial basis function support vector machine, which was cross-validated for all participants for each secondary task and was subject-independent. The algorithm classified high, medium, and low workload with an accuracy of 70% for the visual task and 43% for the mathematical task. When only the low and high workload classes were considered, the classifier accuracy improved to 95% for the visual task and 73% for the mathematical task. The low classifier accuracy for the mathematical task may be explained by the low task difficulty variability [96]. The results highlight the importance of verifying the workload level. The algorithm's **sensitivity** is classified as non-conforming, given the classification accuracy and detecting two workload levels. The algorithm is non-conforming to **diagnositicity**, given that a change in physical workload impacts the algorithm's cognitive workload assessment. The **suitability** and **generalizability** are non-conforming, as the algorithm is unable to assess visual workload.

Oh et al. [90] investigated an ensemble classifier for assessing cognitive workload. Thirty-nine participants used a flight training simulator to complete three tasks, that corresponded to low, medium, and high workload. Audible beeps occurred at random times throughout each task. The low workload task required maintaining straight and level flight, while the medium and high workload tasks required descending/ascending and vice-versa during turning. Four EEG channels, electrocardiogram, and operator performance data were collected throughout each task. The event related potential was calculated from the EEG data, with a one second epoch time-locked to each audible beep and the average heart rate was calculated from the electrocardiogram data every five minutes. Operator performance metrics were calculated once a minute and consisted of deviations from the target altitude, speed, descent/ascent angles, and heading. The NASA-TLX was completed after each task and served as a classifier feature. Four ensemble learning approaches were examined: bagging, boosting, stacking, and voting. Each ensemble classifier was optimized by taking the minimum balance error to reduce large biases and was evaluated using a confusion matrix and the receiver-operator curve. The bagging, boosting, stacking, and voting ensemble algorithms achieved a generalized accuracy of 88%, 96%, 94%, and 97%, respectively. The algorithm's **sensitivity** to workload is classified as needs additional evidence, even though the algorithm had a >80% accuracy and detected three workload levels. The inclusion of the NASA-TLX scores attributes to the high classification accuracy, but are impractical for real-world domains. The **diagnosticity** is non-conforming, as the presence of physical workload may confound the algorithm's cognitive workload assessment. The algorithm's **suitability** and **generalizability** are also non-conforming, as the algorithm is unable to assess visual workload.

Various algorithms were analyzed by Fan et al.[34] and Zhang et al. [141] in a driving simulator tailored to children with autism spectrum disorder. Twenty participants completed eighteen driving scenarios with various task difficulty levels across six different days. Participants restarted a scenario if a certain number of errors were made. Experts continuously rated each scenario from 1 to 9 across several dimensions: engagement, enjoyment, boredom, frustration, and difficulty. Scenarios with a difficulty rating $\geq$5 were determined to be high cognitive load, while those below the threshold were determined to be low cognitive load. EEG, galvanic skin response, electromyography, respiration-rate, skin-temperature, and performance data were fed into five algorithms (i.e., support-vector machine, k-nearest neighbor, decision tree, discriminant analysis, and a neural network) that classified a scenario as low or high cognitive load; thus, data from the entire scenario (approximately five minutes) was used. The multi-modal data were fused at the feature level, decision level, and hybrid level to determine the optimal fusion level performance. Fusing the data at the feature level and using a k-nearest neighbor classifier achieved the highest accuracy (84.43%) when classifying the two cognitive load levels. The algorithm's **sensitivity** and **diagnosticity** are classified as non-conforming, due to only classifying two cognitive load levels. The algorithm is **suitable** (conforming) for classifying overall workload in a driving simulator based on the incorporated workload metrics, but does not conform with **generalizability** due to being participant-specific and not assessing the complete workload state.

Deep learning techniques were used to classify driver perceived workload [76]. Fourteen participants completed nine traffic trials, which manipulated traffic and pedestrian density in order to elicit varying workload levels. After each trial, participants rated their workload level on a Likert scale from 1 to 5. These ratings served as the “true” labels for the workload classification algorithm, which relied on driver behavioral measures (i.e., steering control smoothness), subjective metrics (i.e., driving style questionnaire, and workload sensitivity questionnaire), traffic information, EEG, electrocardiogram, skin conductance, and eye-tracking data as features with an unknown epoch size. The deep-learning algorithm architecture consisted of an input layer, a drop-out layer, two long short-term memory layers with 100 neurons, a fully-connected layer with 100 neurons, and an output layer. Five-fold cross-validation demonstrated that the algorithm achieved 74.5% classification accuracy; thus, the algorithm is non-conforming with the **sensitivity** criterion. It is assumed that the algorithm only classifies overall workload given that the workload ratings are the class labels, which means that the algorithm does not conform with the **diagnosticity** criterion. The algorithm is **suitable** (conforming) to classify workload in a driving simulator, but is not **generalizable** across tasks or populations (non-conforming).

Oh et al.’s [90] algorithm achieves the highest classification indexes of the driving simulator-based algorithms. The algorithm also achieves the highest accuracy among the algorithms validated using flight simulators.

2.2.5 Remotely Piloted Vehicle Simulators

Remotely piloted vehicle simulators are operator-based task environments that contain cognitive and visual workload components. These environments may contain speech and auditory workload, if the participant must verbally respond to audible stimuli.

A multi-layer perceptron, linear regression classifier, and a model tree classifier assessed overall workload in the range of 0 - 56 [101]. Twelve participants completed surveillance and tracking tasks. The surveillance task required searching a desert market place for a high valued target and tracking the target until the target left the marketplace. Task difficulty was determined by the number of non-target people (12 or 48), and the video quality (low or high). The tracking task required tracking a person on a motorcycle and task difficulty was determined by the number of people to track (1 or 2), and terrain (rural or urban). Seven EEG channels, pupil dilation, blink rate, fixation duration, heart rate, heart rate variability, and respiration rate metric were recorded. Each workload metric was pre-processed and re-sampled to 1 Hz before being used as a feature. The classifiers had an unknown epoch size and were trained on and validated against an Improved Performance Research Integration Tool (IMPRINT) model of workload for each task [4]. The workload models ranged from 0 (low workload) to 56 (high workload). Each classifier was trained using 75% of the data and validated with the remaining 25%, with performance determined by the RMSE value between the classifier output and the workload model. The model tree classifier had a RMSE of 2, while the perceptron and linear regression classifiers had a RMSEs of 3 and 5, respectively. The algorithm's **sensitivity** is classified as conforming, due to the $<$5 RMSE value and the algorithm's ability to estimate multiple workload levels. The algorithm's **diagnosticity** is non-conforming, since the algorithm only assessed overall workload (i.e., the algorithm is incapable of determining if an increase in overall workload is due to cognitive, visual, or physical workload). The **suitability** is conforming, given that the algorithm can assess all of the task's workload components, due to the metrics used. The algorithm's **generalizability** is non-conforming, as the algorithm is participant-specific. The approach needs to be extended for tasks that incorporate the auditory and speech components by including additional workload metrics.

This prior work [101] was extended by using a random forest algorithm to estimate overall workload [16] using the same experimental design. Six participant's EEG data from seven electrodes were used to evaluate the algorithm's ability to generalize across a population. The EEG data was sampled at 480 Hz and filtered using a 0.2 Hz low pass and a 40 Hz high pass third-order Butterworth filters. This filtered data was subjected to a 10-second Hanning window with a 1-second stride for workload estimation. Leave-one-participant-out cross-validation evaluated the algorithm's population generalizability when trained on an overall IMPRINT Pro workload model. The random forest algorithm achieved an RMSE value of 0.158 on a range from 11 - 27.8. However, the algorithm did not correlate significantly with the IMPRINT Pro workload model. The algorithm's sensitivity is classified as conforming due to the low RMSE value and detecting multiple workload levels, but the algorithm's diagnosticity is non-conforming due to only detecting overall workload. The algorithm is unable to assess all of the task's workload components, due to solely relying on EEG data. This approach is classified as non-conforming for generalizability due to not generalizing across tasks, but the approach was evaluated across a population.

Durkee et al.'s [31] work was extended by investigating how to improve cross-subject accuracy in a model based classifier [32], called the Functional State Estimation Engine, which classified cognitive workload as a value from 0 - 100 for use in a vigilant spirit control system paired with multi-modal communication. The primary task tracked a high value target by keeping the target centered in the unmanned aerial vehicle's crosshairs. Two secondary tasks consisted of monitoring screen gauges and verbally responding to communication requests. Task difficulty was dependent on the target's speed and motion path, and the number of gauge events and communications requests. Twenty-five participants completed fifteen five minute trials, with task difficulty varying from easy to hard. Three functional state estimation engine models were evaluated: standard, reduced, and expanded. The standard model had six EEG channels and electrocardiogram data, while the reduced model had three EEG channels and electrocardiogram data. The expanded model used the standard model metrics, along with pupil dilation. Each algorithm model was evaluated for second-by-second accuracy and aggregated workload classifier accuracy (accuracy averaged over five minutes). The second-by-second classifier accuracy was determined by the relative classification state of each task (desired output) and the model, which was generated by the NASA-TLX scores and additive noise. The workload classes were low, medium-low, medium-high, and high, with boundaries determined by the NASA-TLX scores. The mean classification accuracy was highest with the standard model at 81%.

The aggregated workload classifier accuracy was evaluated using absolute mean difference between the model output and NASA-TLX scores. The lowest mean difference was the standard model at 10.6. Performance may increase if a larger epoch size is used, due to the incorporation of the electrocardiogram-based metrics. Overall, the algorithm's **sensitivity** is conforming, due to the high classifier accuracy and number of workload levels detected. The **diagnosticity** is classified as needs additional evidence, since the exact electrocardiogram metrics are unknown. If the metrics only include heart-rate variability, then the diagnosticity is conforming. Likewise, if the electrocardiogram metrics only include heart-rate, then the diagnosticity is non-conforming. The algorithm's **suitability** is non-conforming, as the algorithm is unable to assess visual or speech workload. The algorithm's **generalizability** is non-conforming, since the algorithm may only be capable of assessing one workload component.

A multi-model approach was evaluated to improve cross-participant accuracy for the same vigilant spirit control system [33]. The multi-model classifier trained an individual classifier for each NASA-TLX sub-scale and combined the six classifiers' output into a composite score. The multi-model classifier outputted a value, 0 - 100 every second and used EEG, electrocardiogram, and pupil diameter metrics as physiological inputs and behavior and situational data as contextual inputs. The multi-model classifier was trained using nineteen participants' data and validated using the other six. The evaluation metrics were the Pearson's correlation coefficient and absolute difference between the classifier's output and the NASA-TLX score. Noise injection was used for the continuous workload model, but the injected noise was different for each NASA-TLX sub-scale and dependent on different physiological data (i.e., mental demand noise was based on EEG data and physical demand noise was based on electrocardiogram data). The multi-model classifier outperformed the single-model classifier with a Pearson's coefficient of 0.94 and a mean difference of 5.0. The **sensitivity** is conforming, given the low mean difference and number of workload levels detected. The algorithm's **diagnosticity** is non-conforming, since each individual classifier was trained on all the physiological data. Thus, the classifier for physical demand is impacted by an increase in cognitive workload. The **suitability** is conforming, given that the metrics correlate to either cognitive, visual, or physical workload. The algorithm's **generalizability** is non-conforming, due to the algorithm's inability to assess speech or auditory workload.

The multi-model algorithm outperforms the other algorithms, which is attributed to aggregating classifiers trained on each NASA-TLX subscale. However, the algorithm has low diagnosticity. A better approach may be aggregating the workload component classifiers.

2.2.6 Cognitive-Based Tasks

A cognitive-based task environment typically includes tasks that are used as secondary tasks (i.e., n-back test). This task does not fit any human-machine teaming paradigm and may contain a visual workload component (i.e., a reading text).

Zhang et al. [137] addressed inter-subject variability when assessing cognitive workload by using common spatial patterns to filter EEG signals and a large margin unbiased regression machine to classify cognitive workload. Sixteen participants completed six cognitive tasks from cognitivefun.net: go/no-go visual reaction test, visual forward digit span test, stroop test, fast counting test, speed run test, and the n-back test. The experiment had three stages, where the participants rested during the first and third stages, while the second stage contained the six tasks. The large unbiased regression machine classifier was trained on fifteen EEG channels, heart rate, and galvanic skin response data with an unknown epoch size. The EEG data was filtered using a joint spatial-spectral filter, which produces eight features to generate one column for the feature vector **g**. The bias coefficient was removed from the linear regression classifier by solving the zero-bias learning problem, which generates the end result of the classifier $z = \mathbf{u}^T d(g)$. z represents the data's class as the change in workload, **u** is the zero-bias matrix, and $d(\mathbf{g})$ is the recursive feature algorithm. The algorithm outputs the change in workload (i.e., low to high workload or high to low workload). Three unbiased regression machines, each dependent on a different workload metric, were validated using a leave-one-subject out approach. The common spatial patterns based unbiased regression machine classifier achieved the highest accuracy, 87.5%, while the galavanic skin response and heart rate variability classifiers achieved an accuracy of 75% and 62.5%, respectively. The approach of joint spatial-spectral filtering of the EEG data and removing the bias of the regression machine attributed to the algorithm's generalizabilty. The classification indexes are based on the EEG-based regression machine classifier. The algorithm's **sensitivity** to cognitive workload is classified as conforming, due to the high accuracy and detecting three workload levels. The **diagnosticity** is conforming as well, since the EEG metrics only correlate to cognitive workload. The algorithm's **suitability** is non-conforming, as the visual reaction task contains an unassessed visual workload component. The **generalizability** was classified as non-conforming, given that the algorithm can only assess cognitive workload. Additional metrics need to be added in order to classify other workload components.

A linear discriminant analysis and a support vector machine were used to classify three tasks corresponding to different cognitive workload levels [79]. Five participants completed tasks with the "eXperience Induction Machine", a mixed-reality task environment. Each participant completed three tasks corresponding to either high or low cognitive workload: spatial navigation, reading, and calculation. The spatial navigation task required exploring a square spiral maze until the center was reached. A red sphere at each corner of the maze told participants to take part in a reading or calculation task. EEG data was collected during the entire experiment and visually inspected to exclude data containing muscle or eye artifacts. The power spectral density of each frequency band was computed with a five second window and a four percent overlap. The participant-independent classifiers were trained on the EEG data and validated using the leave-one-out cross validation method and the Matthews correlation coefficient. Two classification schemes were investigated: each of the three tasks and high/low cognitive workload. The linear discriminant analysis classifier achieved an 83% accuracy, with a correlation coefficient of 0.72 for the task classification scheme and an 88.5% accuracy with a correlation coefficient of 0.74 for the workload classification scheme. The support vector machine achieved an accuracy of 66% with a 0.45 correlation coefficient for the task classification scheme and an 87% accuracy with a 0.63 correlation coefficient for the workload classification scheme. The algorithm's **sensitivity** is non-conforming, due to only detecting two levels of workload. The **diagnosticity** is conforming, as only EEG metrics were used. The algorithm's **suitability** and **generalizability** are non-conforming, due to being incapable of assessing visual workload

A recurrent three-dimensional convolutional neural network was used to classify two cognitive workload levels across cognitive-based tasks [142]. Twenty participants completed the n-back task (with n = 1 and 3) and an arithmetic task, which showed two numbers (1 - 9) for 0.5 seconds and had two task difficulty levels. The first level required determining if the two numbers summed to 10 or not, while the second level displayed two numbers and then another two numbers 2.5 seconds later for which participants determined if the sum of all four numbers was equal to 20 or not. 16-channel EEG data were recorded during each task and were filtered using a band-pass filter and artifact removal. This filtered EEG data was transformed into three-dimensions using Morlet wavelet transformation to calculate the power spectral density of each 1 Hz band from 1 - 40 Hz, creating a 16 (number of EEG channels) by 40 (power spectral density) matrix. Each row is used to create an EEG topographic map resulting in 40 maps, which are interpolated together using cubic spline interpolation to generate a 20x20x40 matrix. A sequence of 20 cubes (representing 1 second of data) was fed into the workload classification algorithm, which consisted of a 3-D convolutional layer, two stacked bidirectional long short-term memory layers, and a fully-connected layer. The algorithm was validated by being trained on data from one task and tested on the other task, which achieved a 89% classification accuracy. The algorithm's **sensitivity** is non-conforming, due to the binary workload classification. The algorithm's **diagnosticity** is also non-conforming, but the algorithm is **suitable** (conforming) for classifying the task's workload components. The algorithm was shown to **generalize** across two cognitive-based tasks, but is unable to classify workload in other task domains with different workload component demands; thus, the algorithm's generalizability is non-conforming.

The recurrent convolutional neural network workload classification algorithm outperforms the other cognitive-based algorithms, given the high classification accuracy. The algorithm was shown to generalize across two tasks; however, the algorithm has limited viability for an adaptive workload teaming system, due to relying solely on EEG data, which is insufficient to classify physical workload. It is unclear how the algorithm performs on real-world tasks that have varying workload compositions, given the limited real-world nature of the n-back task.

2.2.7 Uncategorized Algorithms

The following algorithms are uncategorized; thus, no common theme is presented.

Schultze-Kraft et al. [106] investigated the use of spatial filters with a linear regression algorithm and linear discriminant analysis to classify cognitive workload. Ten participants captured triplets of colored screws in a bucket in a simulated environment, which contains cognitive and visual workload components. An error occurred if a screw hit the bottom of the screen or was caught in the wrong colored bucket. Task difficulty was manipulated by changing the occurrence rate of screws, until an error rate of 10% and 25% was achieved for the low and high workload conditions, respectively. Each session lasted 24 minutes with 4 blocks of alternating high and low workload conditions. The workload metrics included 64 EEG channels, heart rate, respiration rate, and galvanic skin response, each with an epoch of ninety seconds. The number of EEG-based features was reduced using the spatio-spectral decomposition algorithm and a source power co-modulation spatial filter was applied to the EEG features in order to improve the classifier's accuracy. The participant-specific classifiers predicted low and high cognitive workload. The linear regression classifier achieved a 91.9% mean accuracy, while the linear discriminant analysis achieved a 94% mean accuracy. The algorithm's **sensitivity** is non-conforming, due to detecting only two workload levels. The **diagnosticity** is non-conforming, as an increase in physical workload may confound the cognitive workload assessment. The algorithm's **suitability** is non-conforming, since the algorithm does not assess visual workload. The **generalizability** is non-conforming, since the algorithm is participant-specific.

An aircraft traffic control system adapted to a human's workload level, as determined from EEG data and task complexity [2]. Air traffic control task environments are supervisory-based and incorporate cognitive, visual, auditory, and speech workload components. Four participants with approximentaly twenty years of experience as air traffic control officers completed the experiment over five days. Each day consisted of seventy-five minutes with one of four scenarios: adaptation was not used, adaptation activated by task complexity, adaption activated by EEG, and adaptation activated by task complexity and EEG. Task complexity was determined by the algorithm developed by Sridhar et al. [111], which was dependent on traffic intensity, the speed, altitude, and aircraft heading, and the distance between aircraft. The adaptation activated by the 19 EEG channels was dependent on the ratio of the theta to beta power spectral densities. Adaptation was triggered when the task complexity and/or the EEG ratio was above a threshold and the level of adaptation was determined by the computational red teaming algorithm. The adaptive system was evaluated based on the task complexity scores determined by Sridhar et al's algorithm and responses to the NASA-TLX. The task complexity score was the lowest when adaptation was triggered by EEG alone, while the task complexity score was the highest when triggered by only task-complexity. The NASA-TLX scores show that adaptation incurred higher subjective frustration, mental, physical, and temporal demand, but the participant's rated performance was higher. The subjective scores differ from the same results with prior workload adaptive systems [119, 129], which show lower frustration and workload levels. The traffic control officers' experience and the need to audibly communicate the commands may have contributed to the higher subjective workload levels. The classification accuracy and epoch size are not known. Thus, the algorithm's **sensitivity** needs additional evidence. The algorithm's **diagnositicity** is conforming, as EEG correlates to only cognitive workload. The **suitability** is non-conforming, since no metric assesses the visual, auditory, or speech workload components of the task. The algorithm's **generalizabilty** is also non-conforming, since the algorithm is only capable of assessing cognitive workload.

Popovic et al. [95] developed PHYSIOPRINT based on IMPRINT's model of workload, to classify different tasks within each workload component using proprietary algorithms. Twenty-five participants completed computer based tasks for the auditory, speech, visual, and cognitive workload components. Treadmill and weight lifting based tasks were used for the physical workload component assessment. The physiological data consisted of EEG, electrocardiogram, electromyography, respiration, and head movement metrics with 1 second epochs. An individual proprietary classifier was created for each workload component within IMPRINT and classified events that corresponded to workload levels within each task. The classifiers were validated using the leave-one-subject-out cross validation schema. The speech component classifier used respiration rate and a speech envelope detector as features to classify breath holding, normal breathing, short speech, and long speech events, with an 88.7% accuracy. The fine motor component classifier used normalized electromyography levels and body/limb motion features to classify no physical activity, key press, keyboard typing, contour tracking, and driving events, with an 86.6% accuracy. The gross motor component classifier used X-, Y-, and Z-axis to classify sitting, walking, running, and push-up events, with an 89.3% accuracy. It is unclear what features were used in the audio component classifier. The audio component classified no activity, a beep, a bilateral beep, interpret speech, and interpret sound (e.g., a car honking) events, with a 75.8% accuracy. The visual component classifier achieved an 76.7% accuracy when classifying no activity, registering an image, detecting a difference between two images, reading a symbol, and scanning/searching events. The cognitive component classifier used EEG power spectral density to classify no activity, responding to a command (e.g., command to blink), memory recall, and calculation events, with a 72.5% accuracy. Although, the algorithm contains different classifiers for each workload component, the classification indexes are given based on the aggregated results. The algorithm's **sensitivity** is non-conforming, due to the classification accuracies and multiple workload levels assessed. However, classification accuracy may increase if a larger epoch size is used. The **diagnosticity** is conforming, as there is a separate classifier for each workload component. The algorithm's **suitability** is conforming, since the metrics used for each classifier correlates to the workload component assessed. The **generalizability** is rated as conforming, due to assessing each workload component.

A workload-index algorithmic approach was used to classify two cognitive workload levels in the Mixed Initiative eXperimental Test-bed [115]. The test-bed simulates an autonomous unmanned ground robot and two tasks: change detection and threat detection. The change detection task required participants to monitor and identify intelligence changes (target appeared, disappeared or moved) on an aerial map. Participants monitored the ground vehicle's video feed and marked targets (threats) in the threat detection task. Fifty-five participants experienced three workload levels in four scenarios: change detection task only, threat detection task only, both tasks with the change detection task's demand held constant, and both tasks with the threat detection task's demand held constant. EEG, HRV, fNIRS, and eye-tracking metrics were collected for use in a workload index-based binary classification algorithm. A workload index was computed by taking the ratio of the number of physiological markers observed in the both benchmark (baseline) and test scores over the number observed in the benchmark score. Essentially, the benchmark score consists of metrics that differed by more than 0.5 standard deviations between a low and high workload condition (training data), while the test score consists of metrics that differed between the low workload condition and current data set (testing data). If the computed workload index was lower than a threshold level (0.62), the 2-minutes of physiological data was classified as low workload; else, the data was classified as high workload. The algorithm achieved a classification accuracy of 81%; however, it appears that the algorithm was trained and tested on the same dataset, which inflates performance. Thus, the algorithm's **sensitivity** is classified as requires additional evidence, as the algorithmic performance was not sufficiently assessed. The algorithm does not conform to the **diagnosticity** criterion, but does conform to the **suitability** criterion. The participant-specific algorithm does not conform to the **generalizabilty** criterion.

Zhao et al. [143] used a support vector machine to classify cognitive workload in an anomaly detection task. Forty participants completed two task types: anomalous image detection and anomaly activity detection. The first task required identifying anomalous images from a set of distracting images (e.g., identify that a snow covered tree picture was different from the snow covered mountain pictures) across eight task difficulty levels. Data from the first task was used to train an support vector machine classifier, which classified cognitive workload in real-time for task 2. The second task required monitoring multiple videos of bidirectional pedestrian traffic and identifying abnormal objects (e.g, bikers or skaters) in three task difficulty levels. A few seconds of electrocardiogram, electrooculography, respiration-rate, galvanic skin-response, and performance (i.e., reaction time, miss rate, and false alarm rate) metric data were normalized prior to feature extraction, which extracted time-based and frequency features, while linear discriminant analysis was used for feature reduction. The reduced feature set was used by a support vector machine with a radial basis function to classify cognitive workload for each task, where classification accuracy was determined by leave-one-subject-out cross-validation. The algorithm achieved 95% accuracy when classifying workload for task one. However, performance substantially decreased (52% accuracy) when the algorithm was trained on task one data and tested in real-time for task two. Only the algorithmic performance for the first task is used to classify each evaluation criterion, due to the relatively poor performance for the second task data. The algorithm conforms with the **Sensitivity** criterion given the high classification accuracy for eight workload levels, but an insufficient epoch size was used. The algorithm does not conform to the **diagnosticity** criterion, due to classifying cognitive workload solely. The algorithm is **suitable** for the task environment (conforming), but is not **generalizable** across tasks (non-conforming). The authors did attempt to generalize across tasks in real-time, but achieved low performance.

Popovic et al.'s [95] algorithm achieves the highest ratings amongst the uncategorized algorithms. However, the algorithm is incapable of assessing overall workload, due to individual classifiers for each workload component.

2.3 Workload Assessment Algorithm Discussion

This dissertation developed a workload assessment algorithm that is intended to be included in a system that adapts its behaviors based on the human supervisor's workload levels (i.e., underload, normal load, and overload). None of the reviewed algorithms meet all the criteria necessary to achieve the goal of assessing all the workload components in real-time, due to assessing a limited set of the workload components and the algorithms' limitations in relation to the evaluation criteria: selectivity, diagnosticity, suitability, and generalizability.

All of the algorithms assess a limited set of the workload components, typically only cognitive. A comprehensive adaptive human-robot teaming system needs to assess each individual workload component and generalize across supervisory domains, in order to identify the primary contributors to the supervisor's workload state and to correctly adapt interactions and allocate tasks. Rusnock et al.'s [101] algorithm comes the closest to assessing overall workload; however, it is unable to separately assess the workload components. The existing algorithm also does not assess auditory and visual workload; although, it appears feasible to extend the algorithm to include the auditory and visual workload metrics.

Eleven algorithms conform with sensitivity, while nine algorithms do not conform. The primary limitations related to sensitivity were classification accuracy and detecting <3

workload levels. Four algorithm's do not achieve an 80% classification accuracy; although, the algorithms' accuracy may increase with longer epoch times. Epoch times impact an algorithm's sensitivity, as heart-rate and heart-rate variability require at least thirty seconds of data in order to be sensitive to workload. Five algorithms detect two workload levels, which is deemed insufficient for an adaptive workload system, while ten algorithms detect three workload levels, which is sufficient. Although, detecting ≥4 workload levels or outputting a continuous workload value is optimal in order for a system to determine the adaptation's magnitude and to (re)allocate tasks during overload and underload conditions. Two algorithms detect ≥4 workload levels, while seven output a continuous value. None of the algorithms detect the underload condition, which can be detrimental to performance.

Sixteen algorithms do not conform to diagnosticity, due to the algorithms not accounting for metrics' responses to unassessed workload components. An obvious solution is to choose metrics with high diagnosticity. Seven algorithms conform to diagnosticity; however, the algorithms are not viable for a robust adaptive workload system, as they are unable to identify the distinct workload component attributing to the human's workload level. The ability to identify the distinct contributors to the workload state is needed in order for the system to adapt the task allocations in a manner that adjusts the workload requirements appropriately. Popovic et al.'s [95] algorithm is capable of identifying all five distinct workload components via the five workload component-specific classifiers, but it does not aggregate the workload component classifiers into an overall workload value.

The majority of the algorithms do not assess the complete overall workload state demanded by a task. Thus, said algorithms are not completely suitable for assessing workload in the task environment. An adaptive workload system needs the complete overall workload state in order to adequately normalize an underloaded or overloaded state. If a system does not have the complete workload state, then the system's adaptation to normalize a workload state may be insufficient.

The ratings for the generalizability criterion tended to be lower than the other criteria.

These lower ratings exemplify the reviewed algorithms' primary limitations: individual differences and task generalizabilty. Eleven algorithms developed participant-specific classifiers in order to eliminate individual differences; however, a workload assessment algorithm for an adaptive system may need to generalize across the population. Participant-specific algorithms add a level of complexity to the system, since the system needs to maintain tracking of each human and their corresponding trained algorithm. Three algorithms incorporated feature smoothing or filtering techniques to account for individual differences, which attributed to a high generalized classification accuracy. Such techniques are worthwhile in an adaptive workload system.

Twenty-three of the algorithms were rated poorly for task generalizability, as the average number of workload components assessed was two, typically cognitive and visual. Real-world supervisory environments require assessment of at least the cognitive, visual, speech, and auditory workload components; although, more active domains require all five workload components (i.e., nuclear power plants). Popovic et al.'s [95] algorithm is the only algorithm capable of generalizing to tasks incorporating all five workload components; however, the algorithm does not assess the overall workload state.

Assessing workload algorithmically is also susceptible to an individual's day-to-day variances. Christensen et al.'s [25] approach is the only approach that attempts to address this issue. A small segment of physiological data collected at the beginning of each day is used to incrementally retrain a classifier. This approach had marginal success. Additional research is required to account for day-to-day variability.

The ratings indicate the need for a new workload assessment algorithm, as no algorithm conforms with every criterion. The ratings do not account for additional components to an adaptive workload system that may augment an algorithm's weaknesses. For example, an adaptive workload system may be able to use an activity recognition algorithm to determine the current task focus and choose a workload assessment algorithm tailored to the task, which reduces the need for task generalizability and increases the system's complexity.

Also, online or incremental learning may be deployed in order to tailor an algorithm to a specific participant. Such additional components have not been fully realized for an adaptive workload system, but represent a future research direction.

2.4 Adaptive System Architectures

Developing an adaptive human-robot teaming system requires not only a facet of the human's state to adapt to, such as workload, but also architectures to facilitate when and how an adaption takes place. These adaptive system architectures can be categorize as: adaptive, adaptable, or mixed-initiative [35]. Adaptive systems rely solely on an attribute of the human's state to change the systems or robot's interactions, while adaptable systems allow the human to determine what automation or adaption level the system uses. Mixed-initiative systems seek to combine adaptive and adaptable systems by allowing the human or system to determine the desired interaction methodology or degree of human control. The adaptive system architectures prescribe to the "perceive, select, act" cycle [125], where the system perceives some state variable, selects an action to perform based on the state variable, and then acts by implementing the chosen action.

The system must "perceive" some mechanism in order to invoke or disengage adaptions. These mechanisms can pertain to system, world, task, spatio-temporal, or human states [35]. The system state encompasses known system knowledge, such as current or predicted operation modes. The world state uses environmental measures, such as ambient light, to gain understanding of the surrounding environment. The task state corresponds to the current allocated task set, but can also be abstracted to mission variables (e.g., mission plan or human intent). Spatio-temporal states incorporate location information and time information, while human states may be determined by workload or engagement.

The "select" state may choose the adaptation types, which can be categorized as function allocation, task scheduling, interaction, and content [35]. Function allocation determines what tasks are allocated to which agents (human or system), while task scheduling can change when tasks are performed or the tasks' priority levels. A system may dynamically change its interactions by changing the interaction modality (e.g., auditory or visual) of a stimulus. The system can also vary the amount of content available to the human, such as reducing visual clutter in a high workload condition. Finally, the system "acts" or adapts once it selects an adaptation type(s), either by changing the system automation level or the system's interactions with the human.

The remainder of this section investigates research in relation to adaptive system architecture types and adaption strategies.

2.4.1 Imposed Aid

Adaptive systems seek to "close the loop" between the human and system by changing system components in order to achieve a desired performance level based on human input (e.g., human states or direct input). However, the loop may not need to be closed if optimal performance can be achieved by allowing the system to impose aid without direct human inputs. For example, the Society for Automation Engineers defines six levels of automation that range from 0 (full human control) to 5 (full system control) [62]. Ideally, autonomous vehicles will operate at level 5, while achieving optimal performance without human intervention. However, level 5 autonomous vehicles have yet to be realized and may not be realized in the near future. Thus, current autonomous vehicles must operate at lower autonomy levels with some human control, while maintaining human vigilance levels (keeping the human in the loop). This lower autonomy level operation is also needed for adaptive systems that do not achieve 100% performance with full autonomy.

Imposing aid is not limited to the autonomous vehicle domain. Teo et al. [116] examined the effectiveness of adaptive aid and imposed aid in a human-robot teaming paradigm. The same task environment, as described in Chapter 2.2.7 [115] was used, where the human monitored an unmanned ground vehicle and responded to changes in events and threats. The robotic aid was used solely in the event change task, where participants monitored an aerial map and had to push a button when an icon changed, moved, or disappeared. The robotic aid automatically detected a change and indicated this change using an auditory beep, where the human had to respond accordingly. Aid was triggered using the workload-index based algorithm described previously [115] when workload was determined to be high for at least 1.5 minutes (3 consecutive workload assessments) or was imposed. The adaptive aid achieved higher performance than the imposed aid during high workload conditions, but not during low workload conditions. This result may be due to the adaptive aid not triggering during the low workload condition, but the human had enough resources to allocate to the task in order to achieve high performance; thus, the imposed aid may not of been as beneficial in the low workload condition. Although the adaptive aid was beneficial, the study has several limitations. First, the workload model that the adaptive aid relied on only predicted cognitive workload, which limits the model's viability in other task domains. Additionally, the adaptive aid was only triggered after the human was determined to be overloaded for 1.5 minutes, which may be too long of a time-frame to appropriately adapt system interactions before severe performance decrements. Third, the system adaptations may not be realistic, due to the robot being able to perfectly detect system changes to which the human must respond. Human supervisors may heavily rely on this autonomous capability and suffer from reduced situational awareness. Further, if the robot can perfectly perform a task, then the human has limited utility performing the task. This study does have utility to adaptive system researchers due to demonstrating that an adaptive aid can outperform imposed aid in high workload conditions.

2.4.2 Adaptable Systems

Adaptable systems prescribe to the "perceive, select, act" cycle, but are limited to the human selecting an interaction strategy. This reliance on human selection subjects the system to psychological individual differences, where different humans want varying amounts of system control based on the human's perceived states (e.g., perceived workload). Thropp et al. [117] investigated how adaptable automation can be calibrated to an individual's attentional control. The task consisted of differentiating audio and visual stimuli with various levels of autonomy, where the system identified the stimuli with 90% accuracy at full autonomy. The participants were able to choose the level of autonomy for detecting the stimuli, where the preferred autonomy level offloaded some demand to the system while allowing the participant to remain engaged with the task. However, task performance was higher when the system had a fixed level of autonomy and the participant was unable to change the level; thus, allowing humans to choose a system's level of autonomy may not achieve optimal performance. This result was expected, as there is additional task demand associated with manually choosing the system's interaction strategy, humans not recognizing that an interaction strategy change is needed, or the human not having time to change strategies, despite the need to.

There is a dissociation between the human's perceived needed control level and control level needed to achieve optimal performance. Chavaillaz et al. [23] found that changing an autonomous system's reliability (60%, 80%, 100%) impacted negatively a human's trust in the system's capabilities (i.e., lower reliability levels resulted in lower trust), but the humans tended to keep the system on the same autonomy level. This result demonstrates that humans' may rely heavily on automation or adaptive systems despite system performance. The adaptation type may influence task performance and human trust. Ruff et al. [100] found that adaptable and mixed-initiative systems achieved higher performance than an adaptive system when using three levels of autonomy: manual control, shared-control, and full autonomy. The results did not properly support such a conclusion, as comparisons between the adaptable and adaptive conditions were either not provided or were not significant ($p \leq 0.05$). The paper did demonstrate that participants preferred the shared-control level of autonomy, but this preference may be an influence of the autonomy type and how an adaptation was determined to be needed, rather than actual system performance. The collaborative-decision making-based task required the participant and system to work together in order to decide the best possible option among a list of options in a supervisory-based task environment. The system suggested an option in the shared-control mode, while the system implemented the suggested option in the full autonomy mode. Participants were able to veto options the system chose in the full autonomy mode.

2.4.3 Adaptive Systems

Adaptive systems have mainly focused on adapting autonomy levels or adaptive automation. Adaptive automation frameworks typically rely on the human's workload state to allocate control to the system or human [19, 67, 109]. Higher levels of automation may elicit the underload state [73], while lower levels of automation may elicit the overload state. Thus, adaptive automation may use human workload estimates to prevent underload and overload states from occurring or mitigating them when they do occur. There are questions concerning how often to switch levels of autonomy, as switching frequently may create an "yo-yo effect", where the constant system change causes increased workload [2]. Switching infrequently may not improve task performance effectively. This question has yet to be investigated properly.

There is a question of when to revoke automation, once invoked. Rusnock and Geiger [102] investigated two revoking strategies: workload threshold and minimum duration in a simulation developed in IMPRINT Pro. The workload threshold strategy revoked automation as soon as the human's modeled workload level was less than a threshold: 5, 10, 15, 20, 22, 24. The minimum duration strategy used the workload threshold strategy, but only revoked automation once a minimum duration was met: 1, 2, 5, 10, 15, 20, 25, and 30 seconds. The 5-second minimum duration strategy produced the best results, which may indicate that human workload needs to be sampled at least every 5 seconds. The results also found that workload and situational awareness decrease together; thus, a workload threshold level needs to be set based on desired task performance and situational awareness.

Adaptive systems are not limited to adaptive automation frameworks. Task difficulty may be varied, rather than system autonomy, based on a human state. Bian et al. [11] varied the difficulty of a driving task based on human engagement and measured performance. Participants were more engaged when the task adapted to their engagement level, than when no adaptation occurred, illustrating that a desired engagement level may be obtained by varying task difficulty. Similarly, Walter et al. [122] manipulated task difficulty based on EEG-based workload measurements in a learning environment. Participants with the adaptive task difficulty had significant learning effects, demonstrating that manipulating task difficulty based on workload measurements is feasible.

Most adaptive systems have focused on a single human state construct to determine adaptations, but a multi-dimensional adaption scheme has been theorized [36, 107]. The "Real-Time Assessment of Multidimensional User State" system sought to assess workload, fatigue, and attentional focus in order to adapt a system's interactions intelligently, but is limited to high workload or high fatigue recognition. This multi-dimensional human state assessment was fed into an adaptation decision framework in order to determine how an adaptation occurs. The proof-of-concept system adapted by automating tasks (high workload), using a visual modality to show high priority tasks (incorrect attentional focus), or using an auditory modality (operator fatigue). Although the system seems promising, no performance data was presented; thus, the adaptations are not known to increase task performance. Further, it is unclear if the three participants actually experienced high fatigue levels during the 45-minute task or if the system can adapt to the underload workload state.

2.4.4 Mixed-Initiative Systems

Mixed-initiative systems seek to combine adaptive and adaptable systems in order allow for human-state based adaptations with the flexibility of incorporating human preferences. Hussein and Abbass [59] theorized a framework for human-swarm interaction that relied on human preferences, trust, situational awareness, and workload. These state variables may facilitate the system's interactions and level of autonomy, along with determining when the system is adaptable (when the human can directly change the system's adaptation strategy). Another theorized mixed-initiative system is task balancing between multiple humans [30]. The system displayed cognitive workload to two human pilots, where the pilots were able to allocate tasks to themselves or to the other pilot. The experiment used a confederate for one of the pilots and a workload model for the displayed cognitive workload; thus, no feedback loop was actually used. The results indicate that allowing human-driven task balancing can improve task performance, but the impact of system-driven task balancing on task performance is still unknown. The only known implemented mixed-initiative system was developed by Ruff et al. [100], as described in Chapter 2.4.1. The implemented system achieved similar performance to an adaptable and adaptive system; thus, the utility of mixed-initiative systems has yet to be validated.

2.5 Adaptive System Architectures Discussion

The objective is to develop an adaptive human-robot teaming system architecture that will use human state assessments to determine a system's or robot's interactions, but there is a question of what type of adaptive system architecture to deploy: imposed aid, adaptable, adaptive, or mixed-initiative. Simply imposing aid without human input will not allow a system to understand an interaction's impact effectively, due to the open-loop nature of imposing aid. The human may be able to determine the interaction strategy in adaptable systems, but humans may not use an optimal interaction strategy or have the resources available to change interaction strategies during high workload conditions. These limitations illustrate that there needs to be a non-invasive feedback loop to an adaptive human-robot teaming system architecture; thus, an adaptive or mixed-initiative adaptive system architecture is appropriate.

Adaptive system architectures seek to target adaptations intelligently to a state variable in order optimize performance. These state variables are classified as system, world, task, spatio-temporal, or human variables. However, the literature typically focuses on perceiving human state variables, either functional state (e.g., workload or engagement) or affective state (e.g., emotions). Knowing the other state variables may be beneficial for human-robot teaming systems. The robot teammate will already know system state variables, such as the current operation mode (level of autonomy), visual content being displayed, or current velocity. The world state contains environmental measures that may provide confidence information for a physiological metric. For example, knowing if the ambient light levels are changing may allow a system to infer that a change in pupil dilation is due to the environment, rather than workload. Activity recognition may be used to perceive the human's current task state. Additionally, knowing the mission plan or abstracted task state may allow a robot to predict the next task and when the task occurs correctly, such that the robot can provide aid to minimize task switching times. Finally, spatio-temporal information (e.g., the human's location or time completing the current task) will provide a robot with the necessary information to infer the current task in a non-stationary supervisory-based environment or how long an interaction needs to be postponed, if the current task is non-interruptable. The interactions between the state variables may be used to improve human-state recognition accuracy, such as using activity recognition to extract contextual information to improve human workload assessment accuracy. Developing methodologies that allow robots to perceive not only the human's state, but each state variable may allow for a more robust and generalizable adaptive human-robot teaming architecture, due to providing the system with richer data for more informative adaptations. The current state-of-the-art adaptive system architectures generally focus on a single state variable to adapt to, which may limit the system's viability in real-world scenarios.

Selecting and implementing an adaptation based on the state variables has generally been limited to function allocation (i.e., adaptive automation). Other adaptation types (i.e., task scheduling, interaction, and content) may prove beneficial in adaptive human-robot teaming architectures. Adaptive task scheduling may help manage human workload and ensure that tasks have minimal resource conflicts (i.e., tasks are not competing for the human's visual resources). There will be cases that task scheduling will be unable to mitigate resource conflicts. Adapting the tasks' interaction modality (e.g., one task uses a auditory modality while another task uses a visual modality) to further minimize resource conflicts may be appropriate in such cases. Some human-robot teaming domains may have scenarios where the human's attention is focused on a lower priority task. Adapting the content available to the human by using visual or auditory cues may redirect the human's attention to a higher priority task.

Adaptive systems have been theorized or implemented in a wide range of human-machine systems, but no system is capable of changing the system's or robot's interactions based on diagnostic human workload measurements. The adaptive system developed by Fuchs and Schwarz [36] is the closest to being able to adapt an interaction modality (auditory or visual) appropriately, but such an adaptation was not implemented for the same task (e.g, an alarm) and the system's performance was not validated. Further, the system relied on a non-diagnostic workload measure (did not assess the complete workload state), which is needed to determine how an interaction modality may impact a human.

2.6 Summary

The first part of this chapter examined how objective and subjective metrics vary in regards to changes in overall workload and its contributing components (i.e., cognitive, physical, auditory, visual, and speech) and how these metrics can be aggregated using machine-learning techniques to classify workload components. However, there were several limitations to the current workload aggregation algorithms. The algorithms do not assess overall workload and each workload component, do not detect the underload and overload workload conditions, are limited or evaluated in a single task domain, or do not generalize across a population. Chapter III introduces a workload assessment algorithm designed with these limitations in mind and the associated results to demonstrate how the algorithm overcomes these limitations [48, 51]. The remaining portion of this chapter reviewed state-of-the-art adaptive system architectures and the associated theoretical underpinnings. Currently available systems tend to be narrowly focused on adapting a system's level of autonomy using an incomplete human state assessment, such as only using cognitive workload. No work has examined how interactions may be adaptively scheduled or how the interaction modality can be changed based on a complete human workload state assessment. Chapter V presents a proof-of-concept adaptive human-robot teaming system that intelligently adapts a supervisory-based system's interactions using the developed diagnostic workload assessment algorithm.

Chapter 3

Diagnostic Workload Assessment Algorithm

An adaptive human-robot teaming system needs an algorithm to assess the human's complete workload state in order to adapt interactions and autonomy levels intelligently. The state-of-the-art workload assessment algorithms are not viable for an adaptive system, as they typically only assess a subset of the human's workload state; thus, the algorithms are unable to provide the adaptive system with the necessary workload information. This chapter introduces a diagnostic workload assessment algorithm capable of estimating each workload component and the overall workload state. The algorithm was validated using data from two human-subject evaluations.

The diagnostic workload assessment algorithm is designed to estimate overall workload and each workload component every thirty seconds using knowledge of the task being completed. This time threshold or window size is due to heart-rate and heart-rate variability requiring at least thirty seconds of data in order to be highly sensitive to workload variations [20, 22]. The desired estimates are based on IMPRINT Pro workload models [4], where the algorithm was supervisory trained with the workload models as the "true" values.

Prior work [49] developed a similar algorithm, which relied on physiological metrics to estimate cognitive and physical workload and subjective surveys for estimating the remaining workload components. Overall workload was calculated using a weighted aggregation of the workload components, where the weights were determined by IMPRINT Pro workload models. The previous algorithm is incapable of estimating workload in real-time, due to the reliance on subjective surveys. The current algorithm incorporates more sophisticated filtering techniques, does not rely on subjective metrics, and uses IMPRINT Pro workload models [4, 45] as the desired workload component estimates.

IMPRINT Pro creates models of complex task networks that designate start and stop times for each task and anchors each task to workload component values (i.e., a conversation is anchored to a speech workload component value of 4.0). The task networks and workload component values are used to derive continuous models across seven workload components: auditory, cognitive, visual, speech, gross motor, fine motor, and tactile. The presented models combine the gross motor, fine motor, and tactile components into the physical workload model. An overall workload model is generated by uniformly aggregating the workload component models.

Although IMPRINT Pro generates workload models, the workload models represent predicted workload outcomes, are static, and do not adjust in real-time to the current situation. Additionally, there is uncertainty between the IMPRINT Pro models and the human's actual task, which creates a mismatch between the IMPRINT Pro workload values and the human's actual workload. Using physiological metrics as the foundation of the workload assessment algorithm reduces the uncertainty and provides a more objective workload estimate. The IMPRINT Pro cognitive, physical, auditory, and speech workload models are simply used to train and validate the algorithm. The IMPRINT Pro visual workload model is used to estimate visual workload, due to a lack of objective data.

The developed algorithm was validated using data from two human-robot teaming evaluations: supervisory and peer. The peer-based evaluation was conducted by a prior PhD student (Caroline Harriott). This evaluation's design is presented in this dissertation for completion.

3.1 Algorithm Structure

The developed algorithm relies on the heart-rate (HR), heart-rate variability (HRV), respiration-rate (RR), posture-magnitude (PM), speech-rate (SR), voice pitch (VP), voice intensity (VI), and noise-level (NLVL) workload metrics to provide information regarding the human's workload state. HR, HRV, and NLVL are used to estimate cognitive workload, while HR, RR, and PM are used to estimate physical workload. Auditory workload is estimated using NLVL and speech workload is estimated using SR, VI, and VP.

The desired workload component estimates are based on IMPRINT Pro workload models. IMPRINT Pro models workload across seven components: auditory, cognitive, visual, speech, gross motor, fine motor, and tactile. The gross motor, fine motor, and tactile components are combined into the physical workload component model, as each of the three components are physical in nature. An overall workload value is generated by aggregating each workload component model.

The physiological metrics (HR, HRV, RR, and PM) are collected and calculated using the BioPac Bioharness™, while NLVL is captured using a Reed R8080 decibel meter. The speech-based metrics (SR, VP, and VI) are calculated from a 44100 Khz dual-channel audio signal captured by a Shure PGX1 microphone, where the captured signal is transformed into a mono-channel signal prior to metric calculation. Speech-rate is calculated by detecting syllables within the audio signal [60, 64]. This approach identifies voice intensity peaks that are preceded and followed by dips in intensity; however, some of the identified peaks are due to noise, rather than human speech. The audio signal's zero-crossing rate is used to determine if voice output or noise caused an intensity peak. A low zero-crossing rate represents voiced speech, while a high rate indicates noise. Peaks are discarded if the corresponding zero-crossing rate is higher than a threshold value (1,800 KHz), as used in SpeakRite [60]. The remaining peaks represent the number of syllables in the audio signal, which is divided by the signal's duration in order to calculate speech-rate. Pitch is the audio signal's dominant frequency over a time period. The signal is divided into one-second windows in order to determine how pitch varies across a time frame. Each window is transformed into the frequency domain using the fast Fourier transform, where the maximum power spectral density value's corresponding frequency represents the window's pitch.

A window is applied to each workload metric, where the window size and overlap between the windows are dependent on the user's application. Each windowed metric is filtered using Yin and Zhang's (2014) adaptive exponential smoothing technique in order to remove or reduce unwanted artifacts in a signal, such as noise. The smoothing function incorporates a tuning parameter, λ, which is adaptively chosen based on the data shifts,

$$s_k = (1-\lambda)s_0 + \lambda \sum_{i=0}^{k-1} (1-\lambda)^i x_{k-i}, \tag{3.1}$$

where s_k represents the new value for the respective workload metric at time k, and x_{k-i} represents the workload metric's time-series. λ is determined by:

$$\lambda = \begin{cases} \lambda_1, & if |x_k - s_{k-1}| \geq a\sigma \\ & \text{and } |x_{k+1} - s_{k-1}| \geq a\sigma \\ \lambda_2, & \text{otherwise.} \end{cases}$$

A large λ value (≥ 0.6) is chosen when there is a step-shift in the data represented by a constant (a) multiplied by the standard deviation (σ). Otherwise, a small λ value is chosen (≤ 0.3). The constant a is set to 1 based on the standard deviation of the data segment.

Four time-based features are extracted from the filtered metrics: mean, standard deviation, average gradient, and the slope of the signal. Means and standard deviations capture the metrics' response to workload variations, but do not capture a metric's directional shift, (i.e., the metric is increasing over the time window). The average gradient and slope features are extracted to capture this directional change. Slope is the linear change over the window, while the gradient is the average change between each second in the window.

Relying solely on workload metric-based features may be sufficient in well-known and highly constrained environments, when a single workload component impacts overall workload; however, dynamic environments contain time-varying contributions from multiple workload components. Contextual features capture these time-varying workload contributions. Currently, there are three contextual features: cognitive task composition, physical task composition, and auditory task composition, where task composition represents how much the respective workload component contributes to the human's overall workload. Each task composition is calculated using the corresponding IMPRINT Pro workload component value at time t and dividing the value by the overall workload value at time t, (i.e., $CognitiveTaskComposition = CognitiveModel(t)/OverallModel(t)$). Speech task composition is not included as a contextual feature, due to using voice activity detection to determine if the human is speaking or not.

The cognitive (W_C), physical (W_P), speech (W_S), and auditory (W_A) workload components are estimated using neural networks, where each network uses corresponding workload component features. For example, the cognitive workload component uses HR, HRV, NLVL, and cognitive task composition. Each network contains five layers, where the input layer's number of neurons equals the number of features and the output layer has one neuron. The three hidden layers have 128 neurons, where the number of hidden neurons was chosen by incrementing the hidden neuron number by powers of two until satisfactory performance was achieved. Rectified linear units were used as the activation functions for each input and hidden layer, while the output layer used regression. Each network was trained using the ADAM optimizer [69] and a mean-squared error loss function. Five percent of the training data was used as a validation set, which determined when training stopped in order to prevent overfitting. Overfitting occurs when the algorithm is unable to generalize (perform well) for scenarios on which it was not trained. Visual workload, W_V, is estimated using the respective IMPRINT Pro model, as none of the incorporated workload metrics correlate with visual workload. Speech workload (W_S) is determined to be zero if no syllables are detected within the thirty second window. If a syllable is detected, then a neural network is used to estimate speech workload.

The uniformed aggregation of the individual workload components results in the overall workload, W_O estimate, as IMPRINT Pro aggregates the individual component models into an overall model. The overall workload estimate is mapped to one of n workload states, based on thresholds. Each threshold is tailored to a specific task environment and determined by the corresponding IMPRINT Pro workload models used to train the algorithm.

For example, if three IMPRINT Pro workload models representing the underload, normal load, and overload workload states are used to train the algorithm, then the potential overall workload thresholds may be 20 and 60. Any value less than 20 represents the underload state, any value greater than 60 represents the overload state, and any other value is the normal load state. Different scenarios will require a different number of workload states (i.e., only low and high workload states).

The metric filtering, feature extraction, and workload estimation are combined into Algorithm 1. The algorithm cycles through the while loop (lines 3 - 14) until no workload estimate is required. A workload estimate is generated by filtering the metrics, using the *FilterMetrics* function (lines 16 - 20), which applies a 30 second window and the adaptive exponential smoothing formula. The corresponding features are extracted, *ExtractFeatures*, from each filtered metric. The speech-based metrics are calculated from an audio signal using the *VoiceActivityDetection* function. If the voice activity detection function finds at least one spoken syllable in the audio signal, then speech workload is estimated from the corresponding neural network, otherwise, speech workload is determined to be 0. The remaining workload components are calculated using neural networks (cognitive, auditory, and physical) or an IMPRINT Pro workload model (visual). An overall workload value is the uniform aggregate of the workload component estimates. Generating an overall workload value allows for understanding the human's complete workload state (i.e., underloaded), while the workload components provide information about why the human is in the current workload state.

ALGORITHM 1: Workload Assessment Algorithm

```
Output: WorkloadValues = $[W_C, W_P, W_A, W_S, W_V, W_O]$
WorkloadMetrics = $[HRV, HR, ST, NLVL, RR, PM, SR, VP, VI]$
TaskCompositions = $[Task_{Cognitive}, Task_{Physical}, Task_{Auditory}]$
while Task != Complete do
    if WorkloadEstimateRequired then
        FilteredMetrics = FilterMetrics(WorkloadMetrics)
        $W_C = CognitiveNNET.Predict(ExtractFeatures([HRV, HR, NLVL, Task_{Cognitive}]))$
        $W_A = AuditoryNNET.Predict(ExtractFeatures([NLVL, Task_{Auditory}]))$
        $W_P = PhysicalNNET.Predict(ExtractFeatures([HR, ST, RL, PM, Task_{Physical}]))$
        if VoiceActivityDetection(Audio) then
            $W_S = SpeechNNET.Predict(ExtractFeatures([SR, VI, VP]))$
        else
            $W_S = 0$
        $W_V = VisualWorkloadModel$
        $W_O = W_C + W_A + W_P + W_V + W_S$
End Algorithm
Function FilterMetrics(WorkloadMetrics)
    for each Metric in WorkloadMetrics do
        WindowMetric = Metric(time - WindowSize: time)
        FilteredMetric = $(1-\lambda)s_0 + \lambda \sum_{i=0}^{k-1}(1-\lambda)^i WindowMetric(k-i)$
    return FilteredMetrics
Function ExtractFeatures(FilteredMetrics)
    for each Metric in FilteredMetrics do
        Mean = CalculateAverage(Metric)
        Variance = CalculateVariance(Metric)
        Slope = CalculateSlope(Metric)
        Gradient = CalculateGradient(Metric)
    return Features
Function VoiceActivityDetection(Audio)
    Audio = CalculateAverage(Audio[0], Audio[1])
    VoiceIntensity = AbsoluteValue(Audio)
    PowerSpecturm = FastFourierTransform(Audio)
    Pitch = ArgMax(PowerSpectrum)
    Energy = RootMeanSquareEnergy(Audio)
    EnergyPeaks = FindPeaks(Energy)
    ZCR = FindZeroCrossingRate(Audio)
    for Peak in EnergyPeaks do
        if ZCR[Peak] ¡ ZCRThreshold then
            NumberOfSyllables++
    return True if NumberOfSyllables $\geq 1$
    return False
```

3.2 Hypotheses

The remainder of this chapter analyzes the metric data from each day of the supervisory-based evaluation and validates the developed workload assessment algorithm's ability to estimate workload correctly for the supervisory-based and peer-based evaluations. The evaluations are detailed in Chapters 3.4 and 3.7, respectively. The analyses are broken into four sections: Supervisory Day 1, Supervisory Day 2, Human-Robot Teaming Generalizability, and Peer-Based Task Generalizability. Several hypotheses were formed for these analyses. An overview of the hypotheses is presented in Table 3.1 and they are explained in more detail throughout the remainder of this chapter.

Each hypothesis was formed after the evaluations were completed, but before analyzing the workload assessment algorithm's performance. The classification accuracy threshold of 80% was chosen to reflect the sensitivity criterion's threshold from Chapter 2.2. The 70% classification accuracy threshold was chosen for the hypotheses that tested task or human-robot teaming generalizability, as classification accuracy was expected to decrease.

3.3 Experimental Design Overview for the Supervisory and Peer Evaluations

Physiological data from two human-machine teaming evaluations were used to analyze the workload assessment algorithm post-hoc. Chapter IV discusses the real-time algorithm analysis. Both evaluations manipulated workload and the independent variables were modeled using IMPRINT Pro. Each workload condition was verified post-hoc using performance and subjective metrics.

A five-minute break occurred between each evaluation's tasks and trials. Five minutes is a common time period used with physiological measures [98], since within 5 minutes of the highest workload task, heart rate drops significantly to within 0.6 beats per minute of baseline values. Demographic data, such as age, gender, caffeine consumption, education level, and subjective stress levels, were collected at the start of each evaluation and no

Table 3.1: Chapter 3 Hypotheses

Analysis	**Hypothesis**	
Supervisory Day 1	$\mathbf{H_1^{WL}}$	The algorithm's estimates will be within a standard deviation of the corresponding IMPRINT Pro workload model values.
	$\mathbf{H_2^{WL}}$	The algorithm will classify correctly overall workload and each workload component at least 80% of the time.
	$\mathbf{H_3^{WL}}$	The algorithm's estimates will positively and significantly correlate with the corresponding IMPRINT Pro workload models.
Supervisory Day 2	$\mathbf{H_4^{WL}}$	The algorithm's classification accuracy for the supervisory evaluation's second day will be within 5% of the corresponding classification accuracies for the evaluation's first day.
Human-Robot Teaming Generalizability	$\mathbf{H_5^{WL}}$	The algorithm will generalize across populations by classifying workload at least 80% of the time for an unseen participant.
	$\mathbf{H_6^{WL}}$	The algorithm will generalize across human-robot teaming paradigms by classifying workload correctly at least 70% of the time, when not trained on the teaming relationship specific data.
	$\mathbf{H_7^{WL}}$	The algorithm's estimates will positively and significantly correlate with the corresponding IMPRINT Pro workload models for an unseen human-robot teaming paradigm.
Peer-Based Task Generalizability	$\mathbf{H_8^{WL}}$	The algorithm will classify workload correctly at least 70% of the time for peer-based tasks of which the algorithm was not trained on.
	$\mathbf{H_9^{WL}}$	The algorithm's estimates will significantly and positively correlate with the corresponding IMPRINT Pro workload models for each peer-based task.

relevant interactions were identified. A power analysis determined the number of required participants. There is a difference in sample sizes between the evaluations, but the difference is accounted for in the algorithmic analysis, as the same number of data points per evaluation were use for testing and evaluating the algorithm.

The experimental design for the supervisory-based evaluation is presented in Chapter 3.4, followed by the associated results and discussion. The peer-based evaluation's experimental design is provided in Chapter 3.7, which is followed by the associated results and discussion.

3.4 Supervisory-Based Evaluation Experimental Design

The NASA MATB-II was developed to study human performance in multi-tasking workload scenarios and has been used to evaluate workload assessment algorithms (e.g., [31, 129]. The NASA MATB-II simulated a supervisory-based human-machine team, where the human was supervising a remotely-piloted aircraft.

The within-subjects evaluation with workload level (i.e., underload, normal load, and overload) as the independent variable spanned two days for each participant, where the average time between experiment days was 3.48 days (Std. Dev. = 2.13). Each day manipulated workload (the independent variable) and collected objective workload, performance, and subjective workload metrics as the dependent variables. The first evaluation day required each participant to complete a consent form, a demographic questionnaire, and 10-minutes of training using the NASA MATB-II [27] prior to completing three 15-minute trials. The participants completed four simultaneous tasks during each trial: tracking, system management, resource management, and communication monitoring. These tasks are described further in Chapter 3.4.3.

Each first day 15-minute trial corresponded to either the underload, normal load, or overload condition, with trials counterbalanced to negate ordering effects. A 5-minute break occurred after the training session and between each first day trial in order to allow the participant's physiological signals to return to their resting state levels. The second day emulated real-world conditions in which workload transitioned between levels (e.g., underload (UL) to overload (OL) to normal load (NL)). Each participant completed one 35-minute trial, which contained seven consecutive 5-minute workload conditions. There were three condition orderings:

- UL-NL-OL-UL-OL-NL-UL
- NL-OL-UL-OL-NL-UL-NL
- OL-UL-OL-NL-UL-NL-OL

The orderings were chosen, such that each workload condition transition occurred once. The workload conditions were not randomized, as the second day focused on workload transitions, rather than the conditions themselves. Additionally, the 5-minute time-frame per condition was chosen to reflect the time that physiological signals need to identify a workload transition. Longer time-frames may be chosen, but it was desired that the total trial time was less than 45 minutes. Future research will randomly order and choose time-frames for each condition in order to better simulate a real-world environment.

3.4.1 Environment

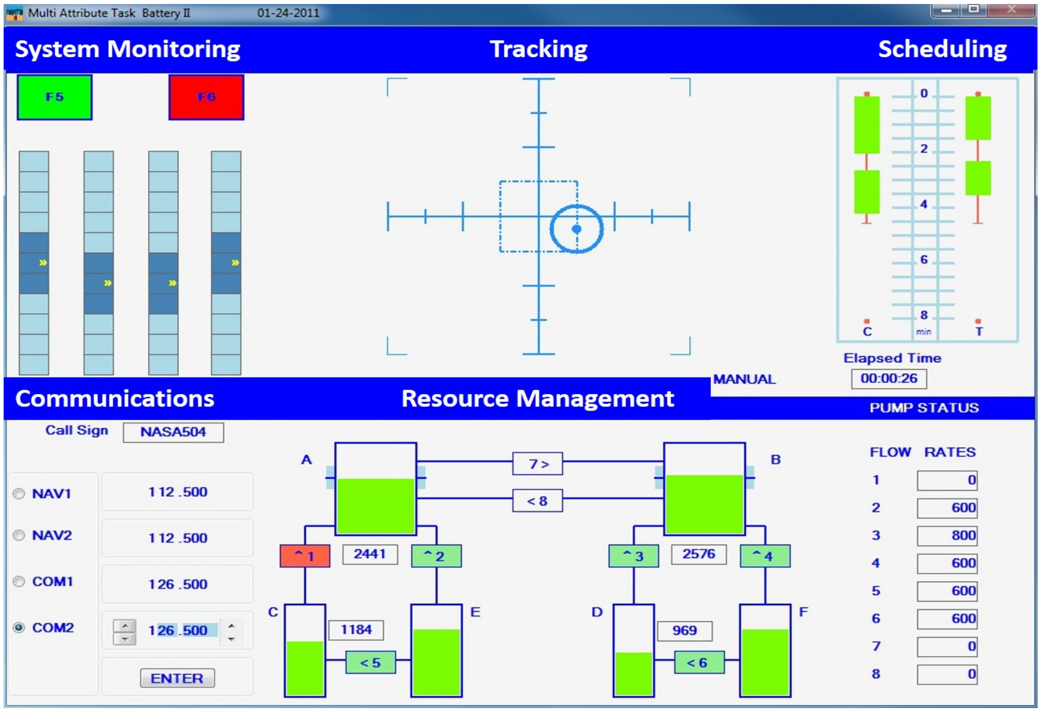
The evaluation occurred in a faculty office on Vanderbilt University's campus. The participants sat in front of a single computer monitor, using a mouse and a joystik to interact with the NASA MATB-II. The experimenter sat behind and to the side of the participant.

3.4.2 Apparatus

The NASA MATB-II ran on a laptop connected to a computer monitor. The participants in front of the monitor were free to ask questions during the training session, but were not allowed to ask questions during each evaluation trial.

3.4.3 Procedure

The participants completed a consent form and a demographic questionnaire upon arrival on the first day. The participants were fitted with the BioPac Bioharness and a mi-



Elapsed Time 00:00:26

Call Sign NASA504

○ NAV1	112.500
○ NAV2	112.500
○ COM1	126.500
● COM2	126.500

FLOW	RATES
1	0
2	600
3	800
4	600
5	600
6	600
7	0
8	0

Figure 3.1: The NASA MATB-II Task Environment

crophone, before reading a script to stage the scenario. A 10-minute training session occurred before starting the first trial. The second day consisted of the same equipment set-up protocol as the first day, before participants completed one 35-minute trial on the NASA MATB-II. The four concurrent tasks are incorporated into the NASA MATB-II: tracking, system monitoring, resource management, and communication monitoring.

The tracking task, depicted in the top, center of Figure 3.1, required participants to keep the circle with a blue dot in the middle of the cross-hairs using a joystick. Workload was manipulated by setting the tracking mode: automatic (low) and manual (high). The underload condition used the automatic mode, with no input, for the entire trial, while the overload condition required the manual mode, or full control, for the entire trial. The normal workload condition switched between manual and automatic modes approximately every 2.5 minutes, which was determined using IMPRINT Pro. The text in the lower right corner of the tracking task area indicated task mode.

The system monitoring task required monitoring two colored buttons and four gauges, shown in the upper left of Figure 3.1. If the green button turned grey (off) or the other button turned red (on), the value was out of range and required resetting by selecting the button. The four gauges had a randomly moving indicator, up and down, that typically remained in the middle. Participants clicked on a gauge if it was out of range (i.e., the indicator was too high or too low). The underload condition had one out of range instance per minute, overload had twenty instances per minute, and normal load had five instances per minute.

The resource management task included six fuel tanks (A-F) and eight fuel pumps (1-8), shown in the bottom center of Figure 3.1. The arrow by the fuel pump's number indicated the direction fuel was pumped. Participants were to maintain the fuel levels of Tanks A and B by turning the fuel pumps on or off. Fuel Tanks C and D had finite fuel levels, while Tanks E and F had an infinite fuel supply. A pump turned red when it failed. Zero pumps failed during the underload condition, while two or more pumps failed during the overload condition. The normal load condition switched from zero pumps failing to one or two pumps failing every minute.

The communications task required listening to air-traffic control requests for radio changes. A communication request may be "NASA 504, please change your COM 1 radio to frequency 127.550." The original MATB communications task required no speech, but a required verbal response was added. A response may be "This is NASA 504 tuning my COM 1 radio to frequency 127.550." Participants were to change the specified radio to the specified frequency by selecting the desired radio and using arrows to change the radio's frequency, as depicted in the lower left of Figure 3.1. Communications not directed to the participants' aircraft, as indicated by the call sign, were to be ignored. The underload condition contained $\leq$ two requests, the overload contained $\geq$ eight, and normal load contained two to eight requests per minute.

3.4.4 Workload Models

IMPRINT Pro was used to develop continuous workload models for each trial, prior to conducting the evaluation. Each workload component model was built by designating start/stop times for each task and linking each task to workload component values. IMPRINT Pro provides anchors to help choose the correct workload component value (i.e., a conversation is anchored to a speech workload component value of 4.0). The workload component values for each NASA MATB-II task were chosen based on IMPRINT Pro's anchors. The chosen values are shown in Table 3.2. The communication task was split into two separate tasks, communication and communication response, in order to model the communication itself and the participant's verbal response, if needed. A visual search task was added to model the constant monitoring of each NASA MATB-II task. The values represent each task instance (i.e., whenever the task was active). The system monitoring and communication tasks were active for ten seconds once a task event occurs (i.e., a communication request was instantiated). The resource management task was active for the entire trial, while the tracking task was only active when the tracking mode was set to manual.

Table 3.2: Workload Component Values for each NASA MATB Task Instance

Task	**Auditory**	**Cognitive**	**Physical**	**Speech**	**Visual**	**Overall**
Tracking	0.0	1.0	4.6	0.0	4.4	10.0
System Monitoring	0.0	1.0	4.2	0.0	1.0	6.2
Resource Management	0.0	3.0	0.0	0.0	6.0	9.0
Communication	6.0	1.0	0.0	0.0	0.0	7.0
Communication Response	0.0	3.0	2.6	4.0	0.0	9.6
Visual Search	0.0	1.0	0.0	0.0	3.00	4.0

The last item needed to build the workload models was the timing of each NASA MATB-II task. Task timings were chosen such that the correct workload condition was elicited, based on the task descriptions in Chapter 3.2.1.2. For example, the timings for the system monitoring tasks were chosen for the overload condition, such that there are approximately twenty instances per minute, where either a light or gauge goes out of range.

The number of tasks per minute by each NASA MATB-II task for each workload condition are shown in Tables 3.3 to 3.5. The resource management task is not shown, as the task is continuous, lasting the entire trial time for each workload condition.

Table 3.3: Number of Tasks Per Minute for the Underload Condition. Note: Vertical Bold Line Designates when an In-Situ Workload Rating was assessed (4:30, 9:00, 13:00 minutes) and TRCK = tracking, SYSM = system management, COM = communication, and RESP = response to the communication

Task	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
TRCK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SYSM	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0
COM	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1
RESP	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
Total	0	0	0	1	0	2	0	2	0	0	0	0	1	0	1

Table 3.4: Tasks Per Minute for the Normal Load Condition

Task	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
TRCK	1	1	0	1	0	1	1	0	0	1	1	1	0	1	1
SYSM	4	4	3	1	4	3	3	2	1	4	4	3	3	2	2
COM	2	2	1	2	0	0	1	1	1	1	1	1	1	1	1
RESP	1	2	0	1	0	0	1	1	1	1	1	0	1	1	0
Total	8	9	4	5	4	4	6	4	3	7	7	5	5	4	4

Table 3.5: Tasks Per Minute for the Overload Condition

Task	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
TRCK	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SYSM	19	18	15	15	21	17	18	20	21	23	18	19	22	20	12
COM	3	4	5	4	3	4	4	3	5	5	4	5	3	4	3
RESP	2	2	4	3	2	3	3	2	4	3	3	2	3	2	1
Total	25	25	25	23	27	25	26	26	31	32	26	27	29	27	17

The IMPRINT Pro workload models for the supervisory evaluation's second day were built in a similar manner to the first day's models. However, each workload condition within the second day's 35-minute trial lasted 5-minutes. The number of tasks per minute for each 5-minute workload condition were the same as the first 5-minutes of the corresponding workload condition for the evaluation's 15-minute first day trial. The IMPRINT

Pro workload modeling results are presented with the algorithm estimates in Chapter 3.5.

3.4.5 Participants

The thirty participants (18 female and 12 male) had a mean age of 25.70 (Standard Deviation (St. Dev.) = 8.65), with an age range from 18 to 62. Video game experience may impact participants' performance when using the NASA MATB-II, due to the multi-tasking nature of the task environment. Twenty-five participants played video games for three or less hours per week. Participants indicated that they exercise a mean of 3.86 (St. Dev. = 1.59) times per week. Seventeen participants did not drink any caffeine the day of the experiment, while six drank $\leq$ 16 oz. and seven drank $\geq$ 16 oz.

The participants slept an average of 6.58 (St. Dev. = 1.57) hours the night before the first day of the experiment and an average of 6.78 (St. Dev. = 1.85) hours two nights prior. The participants' stress levels rated on a Likert scale (1-little to 9-extreme) and were rated on an average of 2.90 (St. Dev. = 1.76), while fatigue levels were rated on average of 2.83 (St. Dev. = 1.32) on the same scale. A Kruskall-Wallis test found no significant effect of sleep, fatigue, or stress on workload. Demographic information was not collected on the evaluation's second day.

3.4.6 Metrics

Objective and subjective workload measures were collected throughout the evaluation. An overview of all of the collected metrics is provided in Table 3.6. The BioPac BioHarness 3 portable measurement device was attached to a flexible strap that fastens around the participants' ribs against the skin, much like an athletic heart rate monitor. This device captured the heart rate, heart-rate variability, respiration rate, skin temperature, body activity, and posture objective workload metrics. The other objective workload metrics, such as noise level and speech-rate, were collected using a REED R8080 decibel meter and a Shure PGX1 microphone head-set, respectively.

Table 3.6: The Objective and Subjective Metrics for the Supervisory Evaluation.

Metric Type	**Metric**
Algorithm Metrics	Heart-Rate Heart-Rate Variability Respiration-Rate Posture Noise Level Speech-Rate Pitch Voice Intensity
Other Objective	Body Activity Skin Temperature Tracking Task: Tracking Error System Monitoring Task: Reaction Time System Monitoring Task: Failure Rate Resource Management Task: Time-in-Range Communications Task: Failure Rate
Subjective	In-Situ Workload Ratings NASA-TLX

The NASA MATB-II automatically collected the tasks' performance metrics. The tracking task's performance was measured by the error in pixels between the center of the cross-hairs and the center of the object, which was collected every fifteen seconds. The system monitoring task's performance was determined by using response time and failure rate. Response time was the number of seconds a participant took to click on a light or gauge once the respective light or gauge went out of range, while the failure rate represented the number of out of range lights and gauges that were not corrected within fifteen seconds, which is the default threshold for the NASA MATB-II. The resource management task's performance was determined by the amount of time the fuel Tanks A and B were out of range (i.e., the fuel levels were not between 2,000 and 3,000 units). The fuel levels of each tank were collected every thirty-seconds. The number of failed communication requests (i.e., the participant failed to respond or the number of times the radio was tuned to the wrong frequency) determined the communications task performance.

The subjective workload measures consisted of verbal in-situ workload ratings and the

NASA-TLX. The verbal ratings were administered three times during each task, which equates to one rating every four and a half minutes. The NASA-TLX was completed after each task. The in-situ workload ratings required the participant to verbally rate six demand channels, (auditory, visual, speech, motor, tactile, cognitive) from 1 (little to no demand) to 5 (extreme demand) [45]. The NASA-TLX is a standard subjective measure of overall workload [47], where the participant rated six demand channels, (mental, physical, temporal, performance, effort, and frustration), and completed fifteen pairwise comparisons between the channels in order to determine the weights for each channel.

3.5 Supervisory Evaluation's Day 1 Results

Several analyses were conducted for supervisory evaluation's first day: Objective Metrics, Subjective Metrics, and Algorithm Analysis. The objective metrics analysis focused on determining if significant differences existed between the three workload conditions for the physiological and performance metrics described in Chapter 3.2.1.5. Likewise, the subjective metrics analysis focused on if significant differences existed for the three workload conditions for the NASA-TLX and In-Situ workload ratings. The last analysis focused on determining how accurate the workload assessment algorithm's estimates were for the first day of the Supervisory Evaluation.

3.5.1 Objective Metrics

The objective measures pertaining to physiological signals will differ from person to person (i.e., resting heart-rates are different). A baseline measurement is subtracted from each physiological metric in order to capture the change in the metric, rather than the raw values. The baseline values are participant-specific, which allows for comparisons. For example, if heart-rate is measured as 80 and 85 for participants 1 and 2, respectively and the respective baseline measurements are 60 and 80. Participant 1 experienced the higher workload than participant 2, as participant 1's heart-rate increased more. However, a comparison of the raw values indicates that participant 2 experienced more workload, which is incorrect.

Means and standard deviations pertaining for each condition were calculated for each objective metric (see Table 3.6). The best value for each group is bolded for each table, where the best value may indicate the highest value for particular metrics (e.g., heart-rate, noise-level) or the lowest value for other metrics (e.g., heart-rate variability, respiration rate). A two-way MANOVA was used to determine if significant differences existed between the three workload conditions, between the participant groups, and if there was a significant interaction between the conditions and groups. The corresponding orderings for each participant group are provided in Table 3.7.

Table 3.7: Task Ordering for Each Group

Group	**Task 1**	**Task 2**	**Task 3**
1	Underload	Normal Load	Overload
2	Underload	Overload	Normal Load
3	Normal Load	Underload	Overload
4	Normal Load	Overload	Underload
5	Overload	Underload	Normal Load
6	Overload	Normal Load	Underload

Heart-Rate

Heart-rate increases with cognitive and physical workload. The associated descriptive statistics are provided in Table 3.8. The overload condition's heart-rates are higher than the other workload conditions for each group, other than Group 4. The heart-rates for normal load were lower than the underload condition for Groups 1, 2, and 5, which is unexpected, as heart-rate increases with workload. The errors with Group 4 and Groups 1, 2, and 5 are attributed to ordering effects, as Group 4 completed the normal load condition first and Groups 1 and 2 completed the underload condition first. A two-way MANOVA determined that there is a significant effect of workload (F(2,23) = 14.29, $p < 0.01$), but there is no significant effect by group. There is a significant interaction between group and workload condition (F(10,48) = 3.06, $p < 0.01$).

Table 3.8: Heart-Rate Descriptive Statistics. Note: Mean (Std. Dev)

Group	Underload	Normal Load	Overload
1	8.83 (2.30)	7.30 (1.86)	**9.13 (3.69)**
2	11.19 (2.29)	6.70 (1.85)	**12.03 (3.04)**
3	7.59 (2.55)	8.11 (2.15)	**8.77 (2.35)**
4	8.95 (3.67)	**13.13 (8.33)**	13.04 (6.94)
5	10.02 (3.73)	8.42 (2.90)	**14.05 (4.26)**
6	9.50 (3.71)	12.73 (6.07)	**14.31 (5.77)**
Overall	9.35 (3.01)	9.39 (4.88)	**9.61 (4.81)**

Respiration Rate

The means and standard deviations for respiration rate by group and workload condition are provided in Table 3.9. The underload condition had the lowest values for Groups 3 - 6, which is unexpected, as respiration rate decreases with workload. The low values for the underload conditions may be attributed to respiration rate not significantly differing for workload, group, or interaction between workload and group.

Table 3.9: Respiration Rate Descriptive Statistics.

Group	Underload	Normal Load	Overload
1	11.38 (5.09)	11.08 (4.91)	8.24 (1.37)
2	10.12 (4.77)	10.04 (5.02)	11.52 (4.21)
3	10.04 (1.58)	10.98 (2.44)	10.17 (1.51)
4	9.34 (2.96)	11.44 (3.58)	10.74 (4.02)
5	8.79 (1.59)	9.19 (2.24)	10.58 (2.74)
6	8.93 (2.42)	10.33 (2.28)	10.25 (3.57)
Overall	9.76 (3.20)	10.51 (3.38)	10.25 (3.02)

Skin Temperature

There was no clear pattern in the skin temperature's descriptive statistics across workload conditions, as shown in Table 3.10. The condition the participant completed first had the lowest skin-temperature for the group, which indicates a task ordering effect. A two-way MANOVA determined that there is no significant effect of workload or group. There is a significant interaction between workload condition and group (F(10,48) = 6.47, p < 0.01).

Table 3.10: Skin Temperature Descriptive Statistics Note: **Bold** represents the lowest value in each group

Group	**Underload**	**Normal Load**	**Overload**
1	0.57 (0.32)	0.91 (0.36)	1.12 (0.57)
2	0.20 (0.05)	0.68 (0.30)	0.63 (0.26)
3	0.54 (0.21)	0.27 (0.09)	0.49 (0.23)
4	0.57 (0.51)	0.29 (0.07)	0.64 (0.31)
5	0.60 (0.26)	0.62 (0.28)	0.43 (0.12)
6	0.49 (0.22)	0.37 (0.14)	0.25 (0.08)
Overall	0.50 (0.30)	0.52 (0.32)	0.59 (0.39)

Posture Magnitude

There were no clear patterns in the posture magnitude means and standard deviations, as seen in Table 3.11. This result was to be expected, as the participants are sitting during the entire experiment. There was no significant effect of workload condition, group, or interaction between workload and group.

Table 3.11: Posture Magnitude Descriptive Statistics

Group	**Underload**	**Normal Load**	**Overload**
1	-8.71 (10.52)	-7.31 (9.04)	-8.90 (10.59)
2	-12.24 (7.99)	-3.17 (23.23)	-12.98 (7.55)
3	-10.61 (9.59)	-8.22 (11.52)	-1.38 (3.48)
4	-18.84 (8.44)	-14.72 (15.18)	-12.02 (3.48)
5	-11.65 (5.32)	-9.70 (6.29)	-6.49 (6.21)
6	-12.99 (8.09)	-8.48 (4.70)	-8.66 (7.09)
Overall	-12.51 (8.35)	-8.60 (12.51)	-8.41 (10.38)

Noise Level

The noise level's descriptive statistics are presented in Table 3.12. The noise level for the overload condition was higher than normal load, which was higher than underload. This result was expected, as the higher the workload level, the higher number of communication requests that occurred. A two-way MANOVA determined the there was a significant effect for workload condition (F(2,23) = 1009, $p < 0.01$). There was no significant effect for group or interaction between group and workload condition.

Heart-Rate Variability

Table 3.12: Noise Level Descriptive Statistics

Group	**Underload**	**Normal Load**	**Overload**
1	2.77 (0.52)	6.89 (0.80)	**18.94 (1.85)**
2	3.00 (0.56)	6.41 (0.37)	**17.47 (1.33)**
3	3.19 (1.15)	7.25 (1.11)	**19.67 (2.27)**
4	3.24 (0.68)	7.43 (0.88)	**19.08 (2.22)**
5	2.59 (0.50)	6.35 (1.04)	**18.49 (2.57)**
6	2.60 (0.68)	6.55 (1.45)	**17.71 (2.07)**
Overall	2.90 (0.71)	6.81 (1.00)	**18.56 (2.05)**

The descriptive statistics for heart-rate variability are presented in Table 3.13. The overload condition elicited lower heart-rate variability values than the underload and normal load conditions, which is expected as heart-rate variability decreases with workload. However, Groups 1, 2, and 5 had higher heart-rate variability for the normal load condition than underload. The same groups had a similar pattern with heart-rate, which indicates a training effect due to the task ordering. A two-way MANOVA determined that there is a significant effect for workload condition (F(2,23) = 16.61, $p < 0.01$), while there is no significant effect for group. The interaction between group and workload condition is significant (F(10,48) = 3.71, $p < 0.01$), which may indicate individual differences between the groups.

Table 3.13: Heart-Rate Variability Descriptive Statistics Note: **Bold** represents the lowest value in each group

Group	**Underload**	**Normal Load**	**Overload**
1	0.36 (0.15)	0.37 (0.17)	**0.35 (0.15)**
2	0.37 (0.15)	0.41 (0.15)	**0.35 (0.15)**
3	0.51 (0.15)	0.50 (0.16)	**0.49 (0.15)**
4	0.42 (0.26)	0.37 (0.21)	**0.36 (0.21)**
5	0.53 (0.09)	0.54 (0.08)	**0.47 (0.07)**
6	0.47 (0.17)	0.44 (0.16)	**0.42 (0.14)**
Overall	0.44 (0.17)	0.44 (0.16)	**0.41 (0.15)**

Speech-Rate

The participants were expected to speak faster as workload increased. The descriptive statistics for speech-rate are provided in Table 3.14. The results correspond to only when the participant was speaking. The participants tended to speak the fastest during the overload condition, while there was a significant main effect on workload (F(2,23) = 700.05, p < 0.01) and group (F(5,23) = 141.14, p < 0.01). There was also a significant interaction between workload condition and group (F(10,48) = 16.59, p < 0.01), which was attributed to individual differences in speaking rate.

Table 3.14: Speech-Rate Descriptive Statistics Note: **Bold** represents the highest value in each group

Group	**Underload**	**Normal Load**	**Overload**
1	0.77 (0.72)	1.16 (1.0)	**1.32 (1.01)**
2	0.59 (0.69)	0.96 (0.95)	**1.49 (1.12)**
3	0.76 (0.67)	1.26 (1.08)	**1.48 (1.12)**
4	0.48 (0.44)	0.87 (0.82)	**1.05 (0.84)**
5	0.88 (0.8)	**1.61 (1.27)**	1.57 (1.15)
6	0.58 (0.54)	0.93 (0.92)	**1.34 (1.08)**
Overall	0.7 (0.68)	1.17 (1.07)	**1.38 (1.07)**

Voice Intensity

Voice intensity was expected to increase as workload increases. The average voice intensities by workload condition and group are provided in Table 3.15. The highest intensities typically occurred during the overload condition, except for groups 5 and 6. These groups had the highest voice intensities overall as well. A two-way MANOVA found a significant main effect on workload (F(2,23) = 103.82, p < 0.01) and group (F(5,23) = 271.37, p < 0.01). Additionally, there was a significant interaction between the groups and workload (F(10, 48) = 42.33, p < 0.01), which may be attributed to individual differenes or microphone placement.

Pitch Lastly, pitch increases with an increase in workload. The resulting descriptive statistics are provided in Table 3.16. Pitch was the highest in the overload condition for each group. Similar to the previous speech-based metrics, there were significant differences for workload condition (F(2,23) = 265.64, p < 0.01) and group (F(5,23) = 259.52, p < 0.01), along with a significant interaction between the two (F(10,48) = 14.75, p < 0.01).

Table 3.15: Voice Intensity Descriptive Statistics Note: **Bold** represents the highest value in each group

Group	Underload	Normal Load	Overload
1	82.26 (124.55)	168.74 (204.07)	**180.44 (196.09)**
2	79.69 (121.74)	111.64 (131.89)	**253.78 (256.64)**
3	147.31 (175.26)	125.9 (134.1)	**312.12 (350.41)**
4	156.98 (343.87)	155.01 (274.82)	**202.79 (246.11)**
5	361.64 (602.76)	**502.15 (948.99)**	322.09 (627.45)
6	221.85 (262.8)	**566.1 (959.36)**	430.68 (450.99)
Overall	178.24 (317.35)	**317.06 (675.34)**	289.06 (403.39)

Table 3.16: Pitch Descriptive Statistics Note: **Bold** represents the highest value in each group

Group	Underload	Normal Load	Overload
1	162.88 (64.07)	176.69 (97.96)	**223.61 (132.62)**
2	138.5 (116.22)	139.54 (51.54)	**174.53 (119.98)**
3	137.04 (78.54)	183.17 (89.35)	**211.02 (171.93)**
4	139.76 (63.39)	145.81 (39.75)	**162.67 (103.71)**
5	149.09 (103.17)	226.65 (220.57)	**256.37 (281.23)**
6	126.16 (49.69)	127.42 (55.92)	**145.01 (99.55)**
Overall	140.82 (81.53)	170.96 (127.81)	**198.41 (172.07)**

3.5.1.1 Performance Measures

It was expected that the participants will perform higher during the normal load condition than the other conditions, as underload and overload can impact task performance. Thus, the best performance value was bolded for each table.

Tracking Task Performance

Task performance for the tracking task was determined using the average RMSE between the center of the cross-hairs and center of the object to be tracked. The resulting descriptive statistics are presented in Table 3.17. The underload statistics are not provided, as that condition does not require the participant to track the object. The normal load RMSEs across the groups are lower than the overload RMSEs, which was to be expected. There was a significant main effect of workload condition (F(2,23) = 347.07, $p < 0.01$), while there was no significant effect of group. A two-way MANOVA also determined that there is no significant interaction between workload condition and group.

Table 3.17: Tracking Task Performance Descriptive Statistics for Average Root-Mean Squared Error. Note: Lower is Better.

Group	**Normal Load**	**Overload**
1	**42.78 (2.04)**	61.92 (5.04)
2	**37.68 (8.33)**	71.39 (33.98)
3	**44.13 (7.05)**	81.05 (25.39)
4	**36.86 (10.53)**	53.00 (15.02)
5	**38.17 (8.24)**	53.56 (13.00)
6	**39.51 (10.72)**	63.41 (21.81)
Overall	**39.85 (8.06)**	64.05 (21.71)

Resource Management Task Performance

The time in seconds that the fuel tanks were out of range determined task performance for the resource management task, where the higher the value represents poorer performance. The descriptive statistics are presented in Table 3.18. Every group's performance, other than group 1 decreased with increased workload. Group 1 performed better on the normal load task, than the underload task. A two-way MANOVA determined that there is a significant main effect of workload condition (F(2,23) = 107.27, $p < 0.01$), but there was no significant effect of group or interaction between group and workload condition.

Table 3.18: Resource Management Task Performance (%) Descriptive Statistics for Time out of Range in Seconds. Note: Lower is Better.

Group	**Underload**	**Normal Load**	**Overload**
1	168.0 (326.31)	**78.0 (86.42)**	696.0 (172.85)
2	**60.0 (76.48)**	150.0 (186.14)	798.0 (130.07)
3	**246.0 (363.97)**	528.0 (270.49)	690.0 (176.2)
4	**36.0 (65.03)**	198.0 (345.72)	636.0 (115.02)
5	**48.0 (62.21)**	48.0 (75.29)	606.0 (152.08)
6	**198.0 (337.81)**	330.0 (406.93)	768.0 (126.57)
Overall	**126.0 (239.79)**	222.0 (289.75)	699.0 (150.54)

System Monitoring Task Performance

The system monitoring task contained two task performance metrics: mean reaction time and failure rate. The descriptive statistics for mean reaction time are presented in

Table 4.22. There was no significant effect between workload conditions or groups. There was also no significant interaction between condition and group.

Table 3.19: System Monitoring Task Performance Descriptive Statistics for Mean Reaction Time in Seconds. Note: Lower is Better.

Group	Underload	Normal Load	Overload
1	**3.45 (1.47)**	3.81 (0.73)	4.11 (0.81)
2	4.37 (2.22)	**4.22 (1.14)**	4.35 (0.57)
3	**4.11 (2.30)**	4.42 (1.17)	4.46 (0.78)
4	**3.40 (0.81)**	3.93 (1.03)	3.92 (0.56)
5	**2.92 (1.38)**	3.59 (0.33)	4.25 (0.68)
6	4.40 (2.13)	**3.47 (0.91)**	4.09 (0.55)
Overall	3.77 (1.72)	3.91 (0.91)	4.20 (0.63)

The failure rate was lower in the underload condition than normal load condition for Groups 1 and 3 (as shown in Table 3.20). Groups 1, 3, and 6 had four participants fail to click $<$ 1 out of the four out of range lights/gauges for the underload condition, which accounts for the groups' lower failure rates. There was a significant effect of workload condition (F(2,23) = 5.19, $p < 0.01$). No significant effect was found by groups or the interaction between workload conditions and groups.

Table 3.20: System Monitoring Task Performance Descriptive Statistics for Failure Rate (%). Note: Lower is Better.

Group	Underload	Normal Load	Overload
1	**15.00 (22.36)**	22.10 (24.30)	19.11 (13.23)
2	25.00 (25.00)	**21,67 (32.83)**	32.83 (22.79)
3	**25.00 (17.67)**	26.36 (6.14)	36.75 (15.07)
4	25.00 (25.00)	**22.08 (15.01)**	21.83 (12.51)
5	35.00 (28.50)	**21.36 (19.05)**	32.75 (23.91)
6	20.00 (20.91)	**16.36 (20.66)**	32.38 (26.70)
Overall	24.16 (22.24)	**21.42 (17.33)**	29.28 (19.21)

Communications Task Performance

The communication task performance was determined by failure rate, specifically the number of failed communication requests divided by the total number of communication requests. The descriptive statistics by workload condition and group are provided in Table

3.21. The groups performed exceptionally well during the underload conditions, which was due to there being only two communication requests. The failure rate for group 3's underload condition is not as high of a failure rate as it seems, as there were two communication requests and five people per group. Thus, failing one communication request was a failure rate of 50%. There was a significant effect by workload condition (F(2,23)=9.90, $p<0.01$), while no significant effect was found for group or the interaction between workload conditions and groups.

Table 3.21: Communication Task Performance Descriptive Statistics for Failure Rate (%). Note: Lower is Better.

Group	**Underload**	**Normal Load**	**Overload**
1	**10.00 (22.36)**	32.95 (24.65)	43.54 (17.92)
2	**0.00 (0.00)**	23.60 (15.45)	27.71 (18.23)
3	30.00 (27.38)	**22.42 (10.37**)	37.74 (23.22)
4	**0.00 (0.00)**	16.61 (10.33)	21.29 (8.86)
5	**20.00 (27.38)**	30.58 (22.13)	39.67 (20.86)
6	**30.00 (44.72)**	30.36 (22.37)	41.93 (20.11)
Overall	**15.00 (26.79)**	26.09 (17.79)	35.32 (18.93)

3.5.2 Subjective Measures

The highest value for each table is bolded, as the value represents the highest perceived workload level.

In-Situ Workload Ratings

The in-situ workload ratings subjectively assessed workload across six dimensions: auditory, visual, speech, motor, tactile, and cognitive. Each rating ranges from 1 (little to no demand) to 5 (extreme demand).

The auditory workload rating's descriptive statistics are provided in Table 3.22. A two-way MANOVA determined that there were significant differences for workload condition (F(2,23) = 264.75, $p < 0.01$), but there were no significant differences for group or the interaction between condition and group.

Table 3.22: Descriptive Statistics for In-Situ Auditory Workload Ratings.

Group	**Underload**	**Normal Load**	**Overload**
1	1.07 (0.15)	2.13 (0.73)	**4.53 (0.38)**
2	1.33 (0.00)	2.73 (0.64)	**4.40 (0.36)**
3	1.13 (0.18)	2.80 (1.04)	**4.66 (0.33)**
4	1.06 (0.15)	3.06 (1.25)	**4.20 (1.07)**
5	1.00 (0.15)	1.66 (1.25)	**4.27 (0.27)**
6	1.66 (1.31)	2.26 (1.36)	**3.80 (0.51)**
Overall	1.21 (0.55)	2.44 (1.00)	**4.31 (0.58)**

The means and standard deviations by group and condition for the visual workload ratings are presented in Table 3.23. There was a significant effect for workload condition (F(2,23) = 61.17, $p < 0.01$). There was no significant effect for group or for the interaction between workload condition and group.

Table 3.23: Descriptive Statistics for In-Situ Visual Workload Ratings.

Group	**Underload**	**Normal Load**	**Overload**
1	1.93 (1.58)	3.13 (1.42)	**4.60 (0.36)**
2	2.20 (0.84)	3.33 (0.24)	**4.40 (0.72)**
3	2.40 (0.60)	4.07 (0.89)	**4.46 (0.50)**
4	1.40 (0.43)	3.80 (3.13)	**4.46 (0.44)**
5	1.80 (1.12)	2.33 (2.13)	**3.67 (1.13)**
6	2.13 (1.55)	2.93 (1.25)	**4.07 (1.09)**
Overall	1.97 (1.06)	3.26 (1.08)	**4.27 (0.77)**

The speech workload ratings by group and condition are presented in Table 3.24. A two-way MANOVA determined that there was a significant effect of workload condition (F(2,23) = 170.60, $p < 0.01$) and no significant effect for the group or the interaction between workload condition and group.

The motor in-situ rating is a subcomponent of physical workload, where the descriptive statistics are provided in Table 3.25. There was a significant difference between workload conditions (F(2,23) = 41.47, $p <0.01$), but no significant difference between the groups or the interaction between the workload condition and group.

Tactile workload is a subcomponent of physical workload, while the means and stan-

Table 3.24: Descriptive Statistics for In-Situ speech workload ratings.

Group	Underload	Normal Load	Overload
1	1.13 (0.30)	1.80 (077)	**3.87 (0.61)**
2	1.27 (0.15)	2.60 (0.86)	**4.07 (0.43)**
3	1.13 (0.30)	2.60 (0.72)	**4.47 (0.30)**
4	1.00 (0.00)	2.53 (0.86)	**4.27 (0.50)**
5	1.00 (0.00)	1.26 (0.28)	**3.13 (1.04)**
6	1.66 (1.31)	2.27 (1.36)	**3.66 (0.78)**
Overall	1.20 (0.56)	2.17 (0.93)	**3.91 (0.74)**

Table 3.25: Descriptive Statistics for In-Situ motor workload ratings.

Group	Underload	Normal Load	Overload
1	1.86 (1.09)	2.50 (1.41)	**3.87 (0.73)**
2	1.93 (0.89)	3.00 (0.53)	**4.00 (0.52)**
3	1.93 (0.72)	3.47 (0.90)	**4.27 (0.79)**
4	1.00 (0.00)	3.13 (1.53)	**3.53 (1.26)**
5	1.26 (0.59)	2.13 (0.76)	**3.33 (1.11)**
6	2.06 (1.47)	2.73 (1.25)	**4.00 (0.75)**
Overall	1.67 (0.92)	2.82 (1.12)	**3.83 (0.87)**

dard deviations for the tactile ratings are presented in Table 3.26. A two-way MANOVA found a significant difference for workload condition (F(2,23) = 14.35, $p < 0.01$) and no significant differences for group or the interaction between group and workload condition.

Table 3.26: Descriptive Statistics for In-Situ Tactile workload ratings.

Group	Underload	Normal Load	Overload
1	1.73 (1.16)	2.06 (1.46)	**2.53 (1.76)**
2	1.46 (0.51)	2.33 (0.66)	**3.00 (0.78)**
3	1.93 (0.86)	2.47 (0.98)	**3.33 (1.33)**
4	1.00 (0.00)	1.60 (0.81)	**2.00 (1.24)**
5	1.20 (0.45)	1.33 (2.07)	**1.46 (0.65)**
6	1.73 (1.30)	2.33 (0.88)	**3.20 (0.77)**
Overall	1.51 (0.83)	2.02 (0.98)	**2.58 (1.25)**

The last in-situ rating is cognitive workload, where the descriptive statics by workload condition and group are provided in Table 4.32. There was a significant effect for workload condition (F(2,23) = 48.55, $p < 0.01$). There was no significant effect of group or the interaction between workload condition and group.

Table 3.27: Descriptive Statistics for In-Situ Cognitive workload ratings.

Group	**Underload**	**Normal Load**	**Overload**
1	2.13 (1.50)	2.93 (0.79)	**4.33 (0.62)**
2	2.40 (1.14)	3.53 (0.50)	**4.53 (0.45)**
3	2.47 (0.87)	3.73 (0.92)	**4.80 (0.18)**
4	1.13 (0.30)	3.40 (1.10)	**4.33 (0.62)**
5	1.53 (1.02)	2.07 (0.98)	**3.40 (1.06)**
6	2.13 (1.43)	2.73 (0.92)	**4.13 (0.93)**
Overall	1.96 (1.12)	3.06 (0.99)	**4.25 (0.75)**

The overall in-situ workload rating is the aggregate of the individual ratings and the descriptive statistics are provided in Table 3.28. Group 5 tended to rate workload generally lower across all conditions than the other groups, while Group 6 had the largest standard deviations. A two-way MANOVA determined that there is a significant difference between workload conditions (F(2,23) = 111.28, $p < 0.01$). No significant effect was found for group or the interaction between group and workload condition.

Table 3.28: Descriptive Statistics for In-Situ workload ratings.

Group	**Underload**	**Normal Load**	**Overload**
1	9.86 (5.68)	14.53 (5.79)	**23.73 (2.91)**
2	10.60 (2.80)	17.53 (2.35)	**24.40 (0.72)**
3	11.00 (3.01)	19.13 (3.47)	**25.73 (1.87)**
4	6.60 (0.43)	17.75 (4.42)	**22.80 (3.48)**
5	7.80 (3.11)	10.80 (3.92)	**19.46 (3.91)**
6	11.40 (7.95)	15.13 (6.44)	**23.46 (2.75)**
Overall	9.54 (4.48)	15.77 (5.02)	**23.2 (3.24)**

NASA-TLX Scores

The NASA-TLX uses a weighted aggregation of six channels: mental, physical, temporal, performance, effort, and frustration. The presentation of the individual NASA-TLX channels represents the unweighted score, while the overall NASA-TLX results represents the weighted aggregate.

The mental demand ratings' means and standard deviations by group and workload condition are provided in 3.29. There was a significant difference by workload condition

(F(2,23) = 65.9, $p < 0.01$) and group (F(5,24) = 2.64, p = 0.04). There was no significant interaction between workload condition and group.

Table 3.29: Descriptive Statistics for the NASA-TLX Mental Demand Scores.

Group	**Underload**	**Normal Load**	**Overload**
1	31 (31.89)	50 (28.50)	**89 (11.40)**
2	53 (16.04)	74 (9.62)	**92 (6.51)**
3	38 (28.85)	75 (17.67)	**94 (6.57**)
4	12 (8.36)	58 (13.12)	**80 (11.72)**
5	24 (34.35)	34 (27.70)	**74 (14.74)**
6	26 (24.34)	47 (23.81)	**85 (13.69)**
Overall	30.66 (26.67)	56.46 (25.47)	**85.66 (12.44)**

The descriptive statistics for the physical demand scores are presented in Table 3.30. A two-way MANOVA determined that there were significant differences between workload condition (F(2,23) = 46.83, $p < 0.01$), but not by group or the interaction between group and workload condition.

Table 3.30: Descriptive Statistics for the NASA-TLX Physical Demand Scores.

Group	**Underload**	**Normal Load**	**Overload**
1	18 (14.41)	27 (28.42)	**82 (26.80)**
2	28 (20.79)	58 (16.81)	**70 (11.18)**
3	26 (26.55)	30 (22.08)	**73 (35.98)**
4	11 (5.47)	56 (20.31)	**88 (10.36**)
5	10 (11.18)	29 (24.59)	**43 (29.07**)
6	23 (20.49)	42 (29.71)	**75 (12.74)**
Overall	19 (17.92)	36.83 (24.44)	**66.5 (24.32)**

The temporal demand corresponds to the time pressure the participant felt and the descriptive statistics by group and condition are provided in Table 3.31. There was a significant difference between workload conditions (F(2,23) = 85.7, $p < 0.01$), but no significant effect was found for group or the interaction between workload conditions and group.

The performance scores rate how well the participant felt they performed the task. The means and standard deviations are presented in Table 3.32. A two-way MANOVA determined that there were significant differences between workload conditions (F(2,23) = 35.6,

Table 3.31: Descriptive Statistics for the NASA-TLX Temporal Demand Scores.

Group	Underload	Normal Load	Overload
1	27 (41.02)	46 (24.84)	**93 (8.36)**
2	29 (22.19)	60 (20.00)	**89 (7.42)**
3	23 (21.67)	53 (12.04)	**78 (38.34)**
4	11 (8.94)	56 (23.82)	**88 (10.37)**
5	9 (8.94)	18 (21.10)	**73 (10.37)**
6	18 (13.50)	42 (25.15)	**78 (24.90)**
Overall	19.5 (21.71)	45.83 (24.17)	**83.16 (19.71)**

$p < 0.01$) and groups (F(5,24) = 8.51, $p < 0.01$). There was no significant interaction between the workload conditions and groups.

Table 3.32: Descriptive Statistics for the NASA-TLX Performance Scores.

Group	Underload	Normal Load	Overload
1	16 (8.94)	36 (19.17)	**67 (23.34)**
2	69 (10.24)	57 (23.87)	**88 (9.08)**
3	27 (30.94)	48 (23.87)	**78 (14.40)**
4	13 (10.36)	48 (19.23)	**58 (24.64)**
5	9 (8.94)	12 (10.95)	**44 (25.34)**
6	19 (16.35)	39 (32.86)	**52 (17.88)**
Overall	25.50 (25.37)	40.00 (25.18)	**64.50 (23.79)**

The descriptive statistics for the effort scores by group and workload condition are provided in Table 3.33. There was a significant effect for workload condition (F(2,23) = 79.47, $p < 0.01$) and group (F(5,24) = 4.29, $p < 0.01$). The interaction between group and workload condition was found to be not significant.

Table 3.33: Descriptive Statistics for the NASA-TLX Effort Scores.

Group	Underload	Normal Load	Overload
1	24 (24.08)	43 (16.81)	**89 (9.62)**
2	52 (15.25)	70 (16.58)	**91 (8.21)**
3	25 (28.28)	72 (10.37)	**95 (5.00)**
4	9 (6.52)	52 (26.36)	**78 (9.75)**
5	18 (29.07)	26 (30.08)	**79 (14.32)**
6	17 (13.96)	48 (24.14)	**77 (22.52)**
Overall	24.16 (23.67)	51.83 (25.51)	**64.50 (23.79)**

The frustration scores' means and standard deviations are presented in Table 3.34. A two-way MANOVA found a significant effect for workload condition (F(2,23) = 120.68, $p < 0.01$) and group (F(5,24) = 2.99, p = 0.03). There was no significant effect for the interaction between group and workload condition.

Table 3.34: Descriptive Statistics for the NASA-TLX Frustration Scores.

Group	**Underload**	**Normal Load**	**Overload**
1	14 (10.84)	36 (25.35)	**82 (17.53)**
2	35 (20.00)	46 (22.75)	**80 (18.71)**
3	18 (18.57)	44 (17.10)	**90 (7.91)**
4	11 (6.52)	39 (29.03)	**73 (13.51)**
5	6 (2.24)	9 (6.52)	**69 (25.84)**
6	13 (7.58)	34 (31.30)	**68 (10.36)**
Overall	16.16 (14.83)	51.83 (25.51)	**77.00 (17.15)**

The NASA-TLX overall weighted scores' descriptive statistics by group and condition are presented in Table 3.35. Group 2 rated the underload condition much higher than the other groups (51.46, while 26.26 is the second highest). These ratings may be attributed to Group 2 undergoing the underload condition first, as Group 1 completed the underload condition first and had the second highest ratings for underload. A two-way MANOVA determined that the NASA-TLX scores significantly differ between workload conditions (F(2,23) = 101.87, $p < 0.01$). The NASA-TLX scores also significantly differ between groups (F(5,24) = 3.97, p = 0.01). The interaction between workload condition and group was found to be not significant.

Table 3.35: Descriptive Statistics for the Overall Weighted NASA-TLX Scores.

Group	**Underload**	**Normal Load**	**Overload**
1	26.26 (26.40)	44.93 (23.16)	**83.39 (11.19)**
2	51.46 (10.41)	65.57 (14.97)	**89.85 (6.40)**
3	26.06 (23.10)	59.39 (8.38)	**90.86 (5.77)**
4	11.99 (8.17)	49.65 (19.04)	**77.05 (7.49)**
5	16.33 (23.14)	24.79 (23.42)	**72.33 (9.56)**
6	20.12 (17.05)	42.65 (25.72)	**73.73 (12.19)**
Overall	25.38 (21.80)	47.83 (22.52)	**81.21 (11.15)**

Overall, each subjective measure significantly differed between workload conditions.

The subjective measures also significantly differed between groups, which was unexpected. Heart-rate, heart-rate variability, and noise-level significantly differed between workload conditions; thus, the developed workload assessment algorithm may be able to distinguish the workload conditions.

3.5.3 Algorithm Analysis

The algorithm analysis evaluates the algorithm's ability to classify workload by using two-fold cross-validation on data from the supervisory evaluation's first day. Cross-validation is a common machine learning technique and produces a more accurate representation of how the algorithm will perform in unseen scenarios. The first training fold contains the first ten consecutive minutes of each workload condition, while the testing fold contains the remaining five minutes of each workload condition. The second training-fold contains the last ten consecutive minutes of each workload condition, while the testing fold contains the first five minutes. Using sixty-six percent of the data for training and the remaining for testing creates a variance in the testing set approximately equal to the variance in the training set. There is no set standard for the percentage split between the testing and training sets, as the split depends highly on the entire data set's size. What is important is that the testing and training sets have similar characteristics. No participant's data was left out of this analysis in order to limit the affect of individual differences.

The algorithm was trained in a supervised fashion, where the inputs were the workload metric data (see Chapter 3.1) from time $(t - 30)$ to time (t) and the expected output was the corresponding IMPRINT Pro workload model value. No subjective ratings were used in the algorithm's training and validation. The algorithms' estimates were compared to the corresponding IMPRINT Pro model workload values using descriptive statistics, while the Kruskal-Wallis test determined if the algorithms' estimates differed between workload conditions. Classification accuracy determines if the algorithm can accurately detect workload states and was calculated by dividing the number of correctly classified data points by the total number of data points. Pearson's correlation analysis analyzed the algorithms' ability to track workload changes within and across workload conditions.

Thresholds for classifying the workload states were extrapolated from the IMPRINT Pro workload models and differ by evaluation. The supervisory evaluation thresholds were the mid-points between the maximum and minimum values for the condition pairs, underload-normal load and normal load-overload, respectively. The thresholds are provided in Table 3.37. Thresholds for classifying speech workload are not provided, due to the trinary nature of the speech workload models (i.e., speech workload is either 0, 2, or 4). These values are insufficient for determining if a human is underloaded, overloaded, or at a normal level.

Table 3.36: Model Ranges and Thresholds by Workload Condition and Component, Evaluation, and Overall Workload. Note: UL = Underload, NL = Normal Load, OL = Overload.

Evaluation	**Condition**	**Cognitive**	**Physical**	**Auditory**	**Overall**
	UL	1.00 - 2.03	0.00 - 0.70	0.00 - 6.00	4.00 - 9.27
Supervisory	NL	4.93 - 11.63	4.04 - 6.86	0.00 - 6.00	13.80 - 39.19
	OL	21.20 - 22.76	11.20 - 12.50	0.00 - 6.00	59.20 - 66.81
Threshold	UL-NL	3.48	2.37	1.5	11.53
	NL-UL	16.42	9.03	2.9	49.20

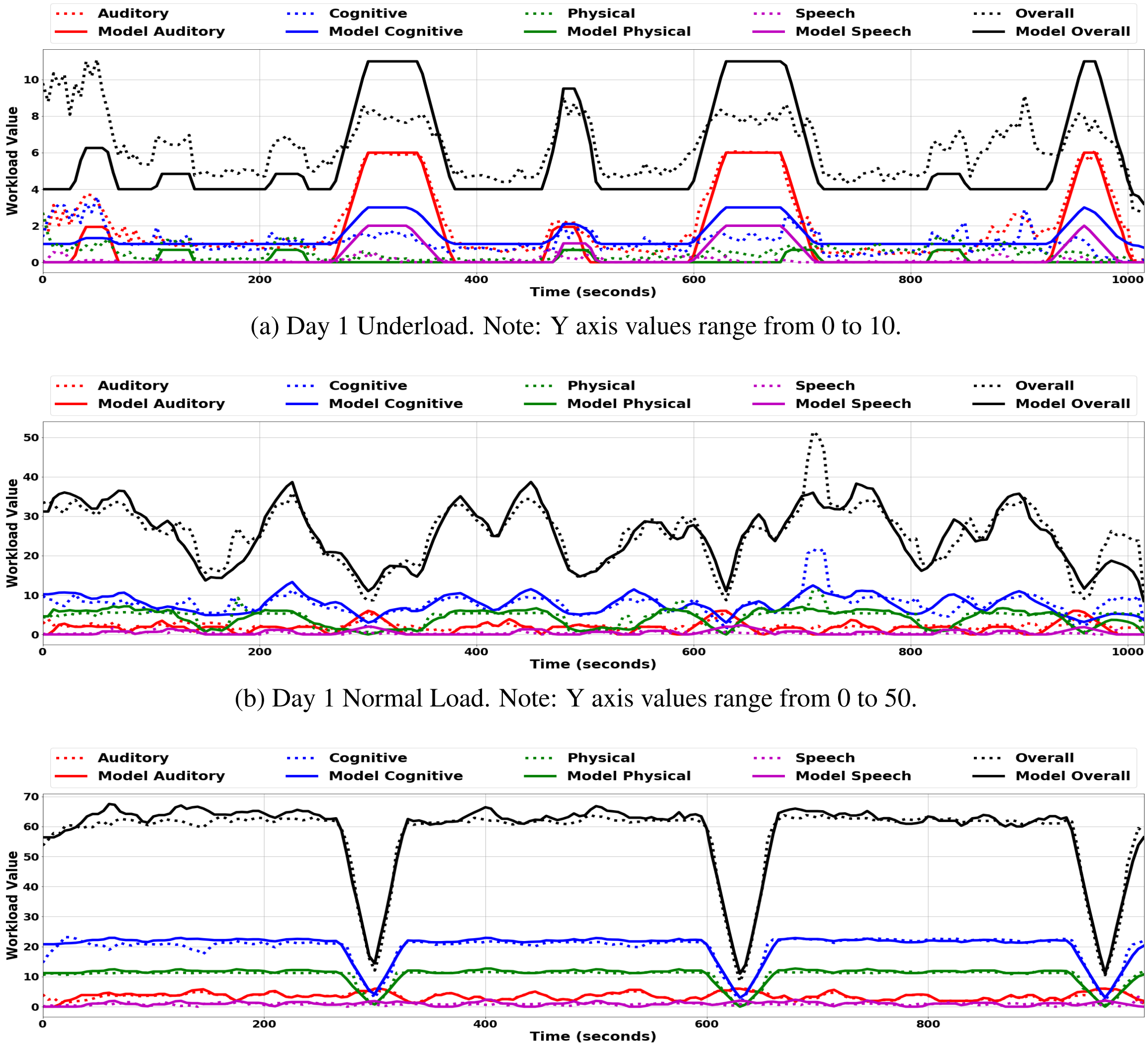
Three hypotheses were formed to determine if the algorithm generalizes across similar workload conditions. It was expected that the algorithm will accurately estimate workload; thus, Hypothesis $\mathbf{H_1^{WL}}$ predicted that the algorithm's estimates will be within a standard deviation of the corresponding IMPRINT Pro workload models. Classification accuracy was used to analyze the algorithm's ability to discriminate between workload conditions. Hypothesis $\mathbf{H_2^{WL}}$ predicted that overall workload and each workload component will be classified correctly at least 80% of the time. An adaptive workload teaming system may use the workload trend to determine if and how an adaptation is to occur. The algorithm's ability to track workload trends within and across workload conditions is analyzed using Pearson's Correlation Coefficient. Hypothesis $\mathbf{H_3^{WL}}$ predicted that the algorithm's estimates will correlate positively and significantly with the corresponding IMPRINT Pro models.

The developed algorithm estimates the cognitive, auditory, speech, physical, and overall workload components every 5 seconds. The algorithm's average estimates and IMPRINT Pro model values are plotted in Figure 3.2 by workload condition. The visual workload component is not included, as the the IMPRINT Pro models are used to estimate visual workload. The IMPRINT Pro's modeled workload values and the algorithm's estimates by workload component and condition are provided in Table 3.37. The algorithm tends to overestimate cognitive workload for the underload condition, while the auditory, physical, speech, and overall workload estimates were within a standard deviation of the IMPRINT Pro model values. The Kruskal-Wallis test determined that the IMPRINT Pro model's values and the algorithm's estimates differed significantly between workload conditions.

Table 3.37: Workload Generalizability: Algorithm Estimated and IMPRINT Pro Modeled Workload Descriptive Statistics and Kruskal-Wallis Significance by Workload Component and Condition.

Workload	**Algorithm**	**UL**	**NL**	**OL**	χ^2
Auditory	Model	1.32 (2.18)	2.12 (1.62)	3.31 (1.13)	63.0*
	Algorithm	2.12 (2.12)	2.4 (1.61)	3.2 (1.21)	32.7*
Cognitive	Model	1.43 (0.72)	8.19 (2.52)	21.93 (0.79)	80.7*
	Algorithm	1.19 (1.85)	8.02 (3.57)	21.28 (2.16)	80.4*
Physical	Model	0.11 (0.24)	4.3 (2.13)	11.73 (0.57)	80.4*
	Algorithm	0.53 (1.08)	4.3 (2.57)	11.3 (0.82)	79.2*
Speech	Model	0.41 (0.72)	0.61 (0.61)	0.95 (0.59)	63.0*
	Algorithm	0.12 (0.61)	0.39 (1.06)	0.85 (1.44)	63.0*
Overall	Model	5.76 (2.58)	26.67 (7.73)	62.85 (2.74)	79.9*
	Algorithm	6.44 (2.88)	26.56 (8.47)	61.57 (3.09)	79.1*

It is important that the algorithm estimate workload accurately in order to discern the human's current workload state for unseen similar workload instances. The algorithm's classification accuracies for each workload component and condition are presented in Table 3.38. Each workload state was correctly classified at least 85% of the time for overall workload and each workload component. The lowest accuracy occurred when classifying physical workload in the normal load condition. There is no corresponding classification



(a) Day 1 Underload. Note: Y axis values range from 0 to 10.

(b) Day 1 Normal Load. Note: Y axis values range from 0 to 50.

(c) Day 1 Overload. Note: Y axis values range from 0 to 70.

Figure 3.2: Workload Generalizability: Algorithm Estimates vs IMPRINT Pro Workload Models.

accuracy for the speech workload component, due to the speech model's values not reflecting the underload, normal load, and overload conditions accurately.

Pearson's Correlation Coefficients were used to analyze the algorithm's ability to track workload variations within and across the three workload conditions. A significant positive correlation coefficient signifies that the algorithm's workload estimates changed in a similar manner to the corresponding IMPRINT Pro workload model. The correlation coefficients between the algorithm's estimates and workload models, within and across

Table 3.38: Workload Generalizability: Algorithm's Classification Accuracy (%) by Workload Component and Condition. **Note:** Speech workload classification accuracies are not provided due to the trinary nature of the IMPRINT Pro speech workload model

Workload	**UL**	**NL**	**OL**
Auditory	91.57	90.02	92.01
Cognitive	100	99.28	100
Physical	99.56	86.67	100
Overall	100	99.83	100

Table 3.39: Workload Generalizability: Pearson's Correlation Coefficients for Within and Across Workload Conditions. Note: * Indicates $p < 0.05$.

Workload	**Within**			**Across**
	UL	**NL**	**OL**	
Auditory	0.85*	0.63*	0.68*	0.77*
Cognitive	0.09*	0.65*	0.12*	0.96*
Physical	0.20*	0.78*	0.09*	0.96*
Speech	0.08*	0.11*	0.03*	0.15*
Overall	0.35*	0.86*	0.41*	0.99*

workload conditions, are provided in Table 3.39. The within column's results indicate that the algorithm was able to track workload variations for auditory and overall workload for each workload condition, but had trouble tracking underload and overload variations for cognitive, physical, and speech workload. Additionally, the algorithm's speech workload estimates were marginally correlated with the speech workload model values, which is due to the trinary speech workload model values and the potential mismatch between when the model believes the participant is speaking and when the participant actually speaks. The Across column demonstrates that the algorithm had excellent tracking of workload variations across workload conditions for each component, besides speech workload, given the strong positive and significant correlations.

3.5.3.1 Discussion

There will be instances in which an adaptive teaming system is deployed in similar workload conditions to the conditions to which the workload assessment algorithm was trained. The algorithm needs to be able to achieve high performance in such cases; thus, Hypothesis $\mathbf{H_1^{WL}}$ evaluated the algorithm's ability to estimate workload in unforeseen, but similar workload conditions by stating that the algorithm's estimates will be within a standard deviation of the corresponding workload model values. The hypothesis was fully supported for overall workload and each workload component.

The algorithm needs to discriminate between workload conditions. Hypothesis $\mathbf{H_2^{WL}}$ stated that the algorithm will correctly classify workload at least 80% of the time and was fully supported. This supported hypothesis demonstrates that the algorithm generalizes across similar workload conditions.

The ability to track workload trends will allow a system to trigger an adaptation to prevent or mitigate the occurrence of an overload or underload state. The third hypothesis ($\mathbf{H_3^{WL}}$) evaluated the algorithm's ability to track workload trends and stated that the algorithm's estimates will correlate significantly and positively with the corresponding IMPRINT Pro workload models. The hypothesis was supported for auditory and overall workload, but was only partially supported for cognitive, speech, and physical workload. Poor correlations were produced for the underload and overload conditions, which was due to the corresponding IMPRINT Pro models having low variance. This low variance was difficult to replicate in the algorithm's estimates, as it relies on physiological signals that do not remain stationary and create inherent variance in the algorithm's estimates. Additional filtering techniques may reduce this variance, but may also reduce the algorithm's performance inadvertently in the normal load condition.

Overall, the algorithm estimated and classified workload accurately in unseen similar workload conditions, as long as the algorithm was trained on prior data from the human.

3.6 Supervisory Evaluation's Day 2 Results

The supervisory evaluation's second day consisted of one 35-minute trial, where workload transitioned between levels: underload, normal load, and overload. Participants completed the supervisory evaluation's second day in one of three workload condition orders. These orders are provided in Table 3.40.

Table 3.40: Supervisory Evaluation's Second Day Workload Condition Orderings.

Group	**Order**
1	UL-NL-OL-UL-OL-NL-UL
2	NL-OL-UL-OL-NL-UL-NL
3	OL-UL-OL-NL-UL-NL-OL

The evaluation's second day was analyzed from three perspectives: objective metrics, subjective metrics, and algorithm analysis. The objective and subjective metrics were analyzed in a similar manner to the Supervisory Day 1 analyses.

3.6.1 Objective Metrics

Heart-Rate

The associated descriptive statistics for heart-rate are provided in Table 3.41. Heart-rate was expected to increase with an increase in workload, but the highest average heart-rate did not occur in the overload condition for orders 1 and 3. This result is attributed to the transitions between workload conditions. A two-way MANOVA determined that there is a significant effect of workload (F(2,26) = 3.54, $p < 0.01$), but there was no significant effect by order. There is a significant interaction between order and workload condition (F(10,48) = 9.74 $p < 0.01$).

Table 3.41: Heart-Rate Descriptive Statistics. The highest values for each order are in **Bold**.

Order	**Underload**	**Normal Load**	**Overload**
1	**17.58 (24.51)**	16.47 (21.71)	17.14 (21.53)
2	20.83 (26.28)	21.07 (28.18)	**21.51 (26.51)**
3	11.65 (6.79)	**12.72 (6.08)**	12.29 (7.94)
Overall	16.74 (21.69)	**17.93 (23.51)**	16.14 (19.0)

Heart-Rate Variability

It was expected that the lowest average heart-rate variability will occur in the overload workload condition, as the metric decreases with an increase in workload. The descriptive statistics for heart-rate variability by order and workload condition are presented in Table 3.42. Each workload condition produced similar descriptive values for heart-rate variability. A two-way MANOVA (F(2,26)= 7.94, $p < 0.01$) determined that the overall values differed significantly between the three workload conditions. The was no significant effect on the workload orderings, but there was a significant interaction between order and workload condition (F(10,48) = 16.30, $p < 0.01$).

Table 3.42: Heart-Rate Variability Descriptive Statistics. The lowest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	0.46 (0.35)	**0.45 (0.27)**	**0.45 (0.23)**
2	0.33 (0.15)	0.34 (0.34)	**0.32 (0.15)**
3	**0.37 (0.23)**	**0.37 (0.23)**	**0.37 (0.24)**
Overall	0.39 (0.27)	**0.38 (0.30)**	**0.38 (0.22)**

Respiration-Rate

Respiration-rate decreases as workload increases. The descriptive statistics for this metric are provided in Table 3.43. The lowest average respiration-rate values typically occurred during the underload and overload conditions; however, the overall values differed significantly (F(2,26)= 19.58, $p < 0.01$) between the three workload conditions. There was no significant effect on the trial orderings or on the interaction between the workload conditions and orderings.

Table 3.43: Respiration-Rate Descriptive Statistics. The lowest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	9.92 (5.08)	9.95 (3.94)	**9.63 (3.93)**
2	**10.82 (3.88)**	11.58 (4.55)	11.35 (4.31)
3	10.03 (4.01)	**9.48 (3.32**)	10.00 (4.03)
Overall	**10.23 (4.44)**	10.62 (4.24)	10.31 (4.15)

Skin-Temperature

An increase in workload typically elicits a decrease in skin-temperature. These associated responses are provided in Table 3.44. The lowest skin-temperatures occurred during the normal load condition for orders 2 and 3, while the lowest value occurred during the underload condition for order 1. A two-way MANOVA determined that skin-temperature differed significantly between workload conditions (F(2,26)= 49.7, $p < 0.01$). There was also a significant effect on the workload condition orderings (F(2,26)= 20.7, $p < 0.01$) and the interaction between the orderings and conditions (F(10, 48)= 202.8, $p < 0.01$).

Table 3.44: Skin-Temperature Descriptive Statistics. The lowest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	**2.43 (1.14)**	2.63 (1.08)	2.67 (1.05)
2	2.66 (0.93)	**2.43 (1.08**)	2.48 (0.98)
3	2.56 (1.22)	**2.45 (1.03)**	2.60 (1.34)
Overall	2.54 (1.11)	**2.50 (1.08**)	2.58 (1.18)

Posture Magnitude

The largest posture magnitudes occurred during the normal load and overload conditions, as seen in Table 3.45. This metric differed significantly between the workload conditions (F(2,26)= 51.81, $p < 0.01$), but not by the condition ordering. Additionally, there was a significant interaction (F(10, 48)= 19.05, $p < 0.01$) between the workload conditions and orderings.

Table 3.45: Posture Descriptive Statistics. The highest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	-26.91 (13.8)	**-24.93 (13.88)**	-26.22 (14.3)
2	-33.34 (18.94)	-33.11 (17.1)	**-32.63 (18.09)**
3	-19.76 (18.42)	**-17.91 (21.07)**	-19.42 (18.69)
Overall	-26.66 (17.77)	-27.46 (17.85)	**-24.91 (18.47)**

Noise-Level

Noise-level was expected to increase as workload increases. The descriptive statics for noise-level are provided in Table 3.46 by workload condition and ordering. The largest noise-levels occurred during the overload condition for each ordering. Additionally, noise-level differed significantly between workload conditions (F(2,26)= 2146, $p < 0.01$) and orders (F(2,26)= 34.67, $p < 0.01$). There was also a significant interaction (F(10, 48)= 44.07, $p < 0.01$) between the workload conditions and orderings.

Table 3.46: Noise-Level Descriptive Statistics. The highest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	8.42 (9.72)	11.22 (10.79)	**12.22 (11.11)**
2	8.83 (10.26)	10.1 (10.86)	**12.67 (11.4)**
3	7.79 (8.93)	9.53 (9.53)	**12.47 (10.19)**
Overall	8.35 (9.67)	10.34 (10.58)	**12.47 (10.77)**

Speech-Rate

The participants were expected to speak faster as workload increased. The descriptive statistics for speech-rate are provided in Table 3.47. Speech-rate was the highest during the overload condition, while there was a significant main effect on workload (F(2,26) = 196.66, $p < 0.01$) and order (F(2,26) = 40.55, $p < 0.01$). There was also a significant interaction between workload condition and group (F(10,48) = 7.81, $p < 0.01$).

Table 3.47: Speech-Rate Descriptive Statistics Note: **Bold** represents the highest value in each group

Order	Underload	Normal Load	Overload
1	1.02 (0.99)	1.32 (1.02)	**1.42 (1.08)**
2	0.84 (0.82)	1.06 (0.94)	**1.36 (1.02**)
3	1.0 (0.96)	1.35 (1.09)	**1.42 (1.09)**
Overall	0.98 (0.95)	1.22 (1.01)	**1.40 (1.07)**

Voice intensity increases as workload increases. The average voice intensities by workload condition and group are provided in Table 3.48. The highest intensities typically occurred during the overload condition for each ordering. A two-way MANOVA found a significant main effect on workload (F(2,26) = 106.86, $p < 0.01$) and ordering (F(2,26) = 61.71, $p < 0.01$). Additionally, there was a significant interaction between the orders and workload conditions (F(10, 48) = 9.56, $p < 0.01$).

Table 3.48: Voice Intensity Descriptive Statistics Note: **Bold** represents the highest value in each group

Order	Underload	Normal Load	Overload
1	163.71 (241.44)	153.72 (147.57)	**202.53 (192.73)**
2	113.68 (166.66)	144.97 (201.08)	**200.71 (224.48)**
3	144.16 (188.08)	211.02 (300.3)	**247.07 (320.37)**
Overall	147.56 (215.7)	164.57 (217.6)	**223.67 (270.31)**

Lastly, pitch increases with an increase in workload. The resulting descriptive statistics are provided in Table 3.49. Pitch was the highest in the overload condition for each order, except order 3. Similar to the previous speech-based metrics, there were significant differences for workload condition (F(2,26) = 36.73, $p < 0.01$) and group (F(2,26) = 10.25, $p < 0.01$), along with a significant interaction between the two (F(10,48) = 8.41, $p < 0.01$).

Table 3.49: Pitch Descriptive Statistics Note: **Bold** represents the highest value in each group

Order	Underload	Normal Load	Overload
1	172.81 (92.81)	182.15 (121.83)	**206.83 (204.6)**
2	158.81 (84.98)	164.55 (103.83)	**199.73 (227.87)**
3	183.35 (198.71)	**202.89 (207.58)**	197.65 (179.19)
Overall	171.66 (121.82)	180.18 (143.24)	**200.44 (199.67)**

Workload Metrics by Workload Transition

The supervisory-based evaluation's second day emulated real-world conditions by having workload transitions between states (i.e., from underload to normal load). It was expected that the objective workload metrics will be sensitive to these workload transitions, where the metric's sensitivity is analyzed using the Spearman correlation coefficient between the metric values and the overall IMPRINT Pro workload model. These correlations are provided in Table 3.50. The Spearman coefficient analysis was chosen, due to the differing ranges between the overall workload model and each metric.The workload transitions contained 120 seconds of data from each participant, meaning that a transition from Underload to Overload contained the last 60 seconds of data from the underload condition and the first 60 seconds from the overload condition. 60 seconds for each condition was chosen to ensure that the workload metrics had enough time to transition between workload states.

Table 3.50: The Spearman Correlation Coefficients between Each Workload Metric and the IMPRINT Pro Overall Workload Model by Workload Transition. Note: * indicates p < 0.05.

Metric	**UL-NL**	**UL-OL**	**NL-UL**	**NL-OL**	**OL-UL**	**OL-NL**
Heart-Rate	0.15*	-0.13*	0.04	-0.09	0.16*	0.51*
Heart-Rate Variability	0.00*	-0.15*	-0.01	-0.03	-0.12*	-0.54*
Noise Level	0.44*	0.74*	0.55*	0.53*	0.76*	0.29*
Posture Magnitude	0.06	-0.24*	0.03	-0.07	0.12*	0.16*
Respiration-Rate	0.78*	-0.62*	-0.27*	-0.04	-0.22*	0.36*
Speech-Rate	0.04	0.21*	0.24*	-0.02	0.35*	0.21*
Pitch	0.13*	-0.18*	-0.39*	0.00	-0.44*	-0.28*
Voice Intensity	0.04	0.11*	0.27*	-0.09	0.32*	0.19*

Significant correlations occurred between each workload metric and the overall workload model for the UL-OL and OL-UL workload transitions. These correlations also indicated that the metrics changed as expected. For example, heart-rate variability decreases as workload increases; thus, negative correlations will occur between the metric and the overall workload model. This result indicates that the metrics trended as expected when there were large changes in workload.

The majority of correlations that were not significant occurred when workload transitioned between the normal load to the overload conditions; however, significant correlations occurred when workload transitions from overload to normal load. This result may be attributed to a physiological “red line” [39], meaning that an increase in workload does not elicit a response from a workload metric, as the participant had no more resources to allocate to the tasks’ demands.

Noise Level and respiration-rate tended to produce the largest correlations for each workload transitions, which is attributed to noise level being a direct task demand measure. The correlations associated with respiration-rate is attributed to respiration rate being sensitive to multi-tasking scenarios.

The metrics and corresponding correlations for the workload transitions varied in their significance, which may mean that the developed workload assessment algorithm may have difficulty capturing these workload transitions. However, the algorithm relies on the combination of the workload metrics, rather than a single metric; thus, the algorithm may depend on specific metrics more than others based on the metric's sensitivity to the workload transitions.

The collected workload metrics are sensitive to overall workload and specific workload components, which permits analysing the metric's correlation to these components. This correlation analysis focused on the workload components that the algorithm estimates via neural networks and the metrics used in this estimation. The Spearman correlation coefficients between noise-level and the IMPRINT Pro auditory workload model are provided in Table 3.51. Noise-level significantly and positively correlated with the workload model when workload transitioned from the normal load state to the underload or overload states. However, noise level was inversely correlated during the underload to normal load transition, which was unexpected. This inverse correlation and the not significant correlations may be attributed to inaccuracies within the auditory workload model.

Table 3.51: The Spearman Correlation Coefficients between Each Workload Metric and the IMPRINT Pro Auditory Workload Model by Workload Transition. Note: * indicates $p < 0.05$.

Metric	**UL-NL**	**UL-OL**	**NL-UL**	**NL-OL**	**OL-UL**	**OL-NL**
Noise Level	-0.29*	0.02	0.49*	0.12*	0.09	0.03

The developed workload assessment algorithm relied on the heart-rate, heart-rate variability, and noise-level metrics to estimate cognitive workload. The Spearman's correlation coefficients between these metrics and the IMPRINT Pro cognitive workload model are presented in Table 3.52. Each cognitive workload metric significantly correlated with the workload model for the OL-UL and OL-NL workload transitions, while noise-level produced significant correlations for each transition. Heart-rate and heart-rate variability did not significantly correlate with the cognitive workload model for the UL-OL and NL-OL transitions.

The resulting Spearman's correlation coefficients between the physical workload met-

Table 3.52: The Spearman Correlation Coefficients between Each Workload Metric and the IMPRINT Pro Cognitive Workload Model by Workload Transition. Note: * indicates p < 0.05.

Metric	**UL-NL**	**UL-OL**	**NL-UL**	**NL-OL**	**OL-UL**	**OL-NL**
Heart-Rate	-0.14*	0.02	0.09	-0.07	0.22*	0.33*
Heart-Rate Variability	0.03	-0.07	-0.03	-0.0	-0.18*	-0.39*
Noise Level	0.2*	0.7*	0.62*	0.54*	0.48*	0.21*

rics and IMPRINT Pro physical workload model are presented in Table 3.53. Significant correlations occurred for each workload metric when workload transitioned from the overload to the normal load conditions, but no significant correlations occurred for the normal load to the underload transition. Posture magnitude had larger correlations than heart-rate did, which was unexpected, as participants were seated throughout the evaluation. This result may be attributed to cognitive workload confounding heart-rate's sensitivity to physical workload.

Table 3.53: The Spearman Correlation Coefficients between Each Workload Metric and the IMPRINT Pro Physical Workload Model by Workload Transition. Note: * indicates p < 0.05.

Metric	**UL-NL**	**UL-OL**	**NL-UL**	**NL-OL**	**OL-UL**	**OL-NL**
Heart-Rate	-0.17*	0.0	0.08	-0.06	0.09	0.29*
Posture Magnitude	0.07	-0.17*	0.01	-0.12*	0.23*	-0.25*
Respiration-Rate	-0.03	0.65*	0.02	-0.01	-0.1	0.3*

It was expected that the speech-based metrics will correlate with the speech workload model, where the resulting correlations are provided in Table 3.54. Significant correlations occurred for the UL-NL, NL-UL, and OL-UL transitions for each speech workload metric. Speech-rate tended to produce the largest correlations, which was attributed to speech-rate being sensitive to human speech. If the participant was not speaking, then participant's speech-rate was zero. The pitch and voice intensity are not as sensitive as speech-rate, as the metrics are sensitive to non-verbal artifacts (i.e., the air-traffic control messages).

The workload metrics were sensitive to each workload component, although the majority of the correlations were weak. The developed workload assessment algorithm may still

Table 3.54: The Spearman Correlation Coefficients between Each Workload Metric and the IMPRINT Pro Speech Workload Model by Workload Transition. Note: * indicates p < 0.05.

Metric	**UL-NL**	**UL-OL**	**NL-UL**	**NL-OL**	**OL-UL**	**OL-NL**
Speech-Rate	0.25*	0.11*	0.63*	0.10	0.14*	0.19*
Pitch	-0.15*	-0.02	0.18*	0.00	0.11*	0.02
Voice Intensity	0.13*	0.08	0.59*	0.04	0.17*	0.2*

produce accurate workload component metrics, as the algorithm relies on a combination of the metrics. Additionally, features were extracted from the raw metric data, which may increase the algorithm's accuracy.

3.6.2 Performance Measures

Task performance decreases when workload is too low (underload) or too high (overload); thus, it was expected that task performance will be the highest during the normal load condition. The best performance value was bolded for each of the following tables.

Tracking Task Performance

Task performance for the tracking task was determined using the average root-mean squared error between the center of the cross-hairs and the center of the object to be tracked. These average errors are presented in Table 3.55 by ordering and workload condition. There are no corresponding results for the underload condition, as the tracking task was never active during the condition. The highest performance was achieved during the normal load condition for each ordering. There was a significant main effect of workload (F(2,23) = 17.89, p < 0.01) and a significant effect on the ordering (F(2,23) = 8.67, p < 0.01). There was no significant interaction between the workload condition and ordering.

Resource Management Task Performance

The average amount of time the resource management task was in range by workload condition and ordering are provided in Table 3.56. The participants typically performed the best during the underload condition, except for order 2, where the highest performance

Table 3.55: Tracking Task Performance Descriptive Statistics for the Average Root-Mean Squared Error.

Order	Normal Load	Overload
1	**53.82 (30.75)**	63.16 (39.36)
2	**49.03 (22.8)**	56.0 (30.64)
3	**43.93 (17.61)**	56.61 (32.31)
Overall	**49.06 (24.49)**	58.38 (34.22)

was during the normal load condition. This result was attributed to the order 2 having more instances of the normal load condition than the other orderings, as the performance was similar to order 2's underload condition. A two-way MANOVA determined that there was a significant main effect on workload (F(2,23) = 49.52, $p < 0.01$) and ordering (F(2,23) = 2071, $p < 0.01$). There was also a significant interaction between the condition and ordering (F(10, 48) = 10.23, $p < 0.01$).

Table 3.56: Resource Management Task's Time in Range (%) Descriptive Statistics for the Average Root-Mean Squared Error.

Order	Underload	Normal Load	Overload
1	**0.73 (0.44)**	0.66 (0.47)	0.67 (0.47)
2	0.73 (0.44)	**0.75 (0.43)**	0.59 (0.49)
3	**0.81 (0.39)**	0.79 (0.40)	0.64 (0.48)
Overall	**0.75 (0.43)**	0.73 (0.44)	0.63 (0.48)

System Monitoring Task Performance

There were two task performance metrics for the system monitoring task: reaction time and failure rate. The descriptive statistics for reaction time are provided in Table 3.57. The lowest average reaction times occurred during the normal load condition, except for order 2, where the underload condition produced the lowest reaction times. There was a significant main effect on workload (F(2,23) = 3.28, $p = 0.04$), but no significant effects for the workload condition ordering or the interaction between the workload conditions and orderings.

The system monitoring task's failure rates by workload condition and ordering are provided in Table 3.58. The participants were the most successful during the underload con-

Table 3.57: System Monitoring Task's Reaction Time's Descriptive Statistics.

Order	Underload	Normal Load	Overload
1	3.80 (3.03)	**3.63 (2.49)**	3.82 (2.6)
2	**3.39 (2.34)**	3.68 (2.53)	4.01 (2.66)
3	4.11 (2.74)	**3.65 (2.35)**	3.84 (2.55)
Overall	3.76 (2.75)	**3.66 (2.47)**	3.88 (2.6)

dition for each workload condition ordering. A two-way MANOVA determined that there were significant differences between the workload conditions (F(2,23) = 47.87, $p < 0.01$) and orderings (F(2,23) = 2.05, $p < 0.01$). There was no significant interaction between the workload conditions and orderings.

Table 3.58: System Monitoring Task's Failure Rate (%) Descriptive Statistics.

Order	Underload	Normal Load	Overload
1	**14 (35)**	20 (04)	22 (39)
2	**03 (18)**	17 (37)	25 (41)
3	**11 (32)**	24 (42)	24 (41)
Overall	**07 (25)**	19 (39)	24 (41)

The participants monitored and responded to air-traffic control requests. A failure occurred when the participant failed to respond or responded incorrectly. The descriptive statistics for these failures are provided in Table 3.59. The lowest failure rates occurred during the normal load condition for each ordering, while there was also a significant main effect of workload (F(2,23) = 112.71, $p < 0.01$) and the workload ordering (F(2,23) = 9.18, $p < 0.01$). Additionally, there was a significant interaction between the workload conditions and orderings (F(10, 48) = 1.22, $p = 0.03$).

Table 3.59: Communications Task's Failure Rate (%) Descriptive Statistics.

Order	Underload	Normal Load	Overload
1	17 (29)	**01 (06)**	19 (24)
2	12 (25)	**05 (14)**	19 (24)
3	28 (26)	**08 (19)**	22 (25)
Overall	22 (26)	**05 (15)**	20 (25)

3.6.3 Subjective Ratings

The highest value for each table is bolded, as the value represents the highest perceived workload level.

In-Situ Workload Ratings

The In-Situ workload ratings were administered every 4.5 minutes during the supervisory evaluation's second day's trial. The auditory rating's descriptive statistics are provided in Table 3.60. The highest ratings occurred during the overload condition for each workload condition ordering, while there was a significant main effect on workload (F(2,23) = 151.99, $p < 0.01$), but not on the condition orderings or the interaction between the orderings and conditions.

Table 3.60: Descriptive Statistics for In-Situ Auditory Workload Ratings.

Order	Underload	Normal Load	Overload
1	1.23 (0.43)	2.30 (0.57)	**3.90 (0.64)**
2	1.35 (0.93)	2.23 (1.01)	**3.05 (0.89)**
3	1.20 (0.52)	2.40 (0.88)	**3.63 (0.81)**
Overall	1.26 (0.63)	2.30 (0.86)	**3.54 (0.85)**

The means and standard deviations for the visual workload ratings are provided in Table 3.61. Similar to the auditory ratings, the highest visual ratings occurred during the overload condition of each ordering. A two-way MANOVA determined that there was a significant effect on workload conditions (F(2,23) = 48.54, $p < 0.01$) and the workload condition orderings (F(2,23) = 14.86, $p < 0.01$). There was no significant interaction between the workload conditions and orderings.

Table 3.61: Descriptive Statistics for In-Situ Visual Workload Ratings.

Order	Underload	Normal Load	Overload
1	1.80 (0.81)	2.45 (0.76)	**3.50 (0.61)**
2	1.75 (1.02)	2.30 (1.06)	**2.85 (1.09)**
3	2.25 (0.91)	3.20 (0.62)	**3.87 (0.86)**
Overall	1.91 (0.91)	2.60 (0.94)	**3.47 (0.96)**

The participants were required to verbally respond to communication requests. The associated speech ratings for these responses are provided in Table 3.62. Again, the highest ratings occurred during the overload condition and there was a significant effect between workload conditions (F(2,23) = 119.60, p $<$ 0.01). A significant effect also occurred between the workload condition orderings (F(2,23) = 2.51, p $<$ 0.01) and the interaction between the orderings and workload conditions (F(2,23) = 4.02, p $<$ 0.01).

Table 3.62: Descriptive Statistics for In-Situ Speech Workload Ratings.

Order	**Underload**	**Normal Load**	**Overload**
1	1.07 (0.25)	2.20 (0.70)	**3.65 (0.88)**
2	1.35 (0.93)	1.90 (1.03)	**2.65 (0.75)**
3	1.2 (0.52)	1.90 (0.55)	**3.20 (0.89)**
Overall	1.19 (0.60)	1.99 (0.83)	**3.17 (0.92)**

A sub-component of physical workload is the motor component, where the descriptive statistics for the motor ratings are presented in Table 3.63. The overload workload condition elicited the largest motor ratings for each workload condition ordering. A two-way MANOVA determined a significant main effect on workload (F(2,23) = 74.77, p $<$ 0.01), but not for the orderings and the interaction between the orderings and workload conditions.

Table 3.63: Descriptive Statistics for In-Situ Motor Workload Ratings.

Order	**Underload**	**Normal Load**	**Overload**
1	1.57 (0.68)	2.15 (0.75)	**3.65 (0.67)**
2	1.50 (0.95)	2.30 (1.02)	**2.75 (1.02)**
3	1.45 (0.60)	2.35 (0.81)	**3.37 (0.93)**
Overall	1.51 (0.74)	2.27 (0.88)	**3.27 (0.95)**

Another sub-component of physical workload is the tactile component. The means and standard deviations by order and workload condition are provided in Table 3.64. The largest tactile ratings occurred during the overload condition for each ordering, while there was a significant difference between the workload conditions (F(2,23) = 16.81, p $<$ 0.01). A two-way MANOVA found no significant difference between the workload orderings and for the interaction between the orderings and workload conditions.

The last In-Situ workload rating is the cognitive ratings. The associated descriptive

Table 3.64: Descriptive Statistics for In-Situ Tactile Workload Ratings.

Order	Underload	Normal Load	Overload
1	1.2 (0.41)	1.45 (0.69)	**2.35 (1.31)**
2	1.55 (1.00)	1.77 (1.07)	**1.85 (1.09)**
3	1.2 (0.41)	1.85 (0.75)	**2.37 (1.00)**
Overall	1.3 (0.64)	1.70 (0.89)	**2.21 (1.13)**

statistics are provided in Table 3.65. The participants tended to rate cognitive workload the highest during the overload condition. A two-way MANOVA found a significant main effect on the workload conditions (F(2,23) = 79.79, $p < 0.01$) and orderings (F(2,23) = 2.37, p = 0.01). There was no significant interaction between the workload conditions and orderings.

Table 3.65: Descriptive Statistics for In-Situ Cognitive Workload Ratings.

Order	Underload	Normal Load	Overload
1	1.67 (0.8)	2.45 (0.76)	**3.80 (0.83)**
2	1.60 (0.99)	2.47 (0.94)	**3.05 (0.83)**
3	1.80 (0.7)	2.65 (0.81)	**3.60 (0.81)**
Overall	1.69 (0.83)	2.51 (0.85)	**3.50 (0.86)**

The uniform aggregate of each in-situ workload rating resulted in an overall rating. The mean and standard deviations for the overall ratings are presented in Table 3.66. The participants' overall workload ratings were the highest during the overload condition, while a significant effect (F(2,23) = 126.49, $p < 0.01$) between the conditions occurred. There was also a significant effect between the condition orderings (F(2,23) = 3.71, p = 0.03) and the interaction between the orderings and conditions (F(2,23) = 3.18, p = 0.01).

Table 3.66: Descriptive Statistics for In-Situ Overall Workload Ratings.

Order	Underload	Normal Load	Overload
1	8.53 (2.57)	13.00 (3.24)	**20.85 (2.87)**
2	9.10 (5.43)	12.97 (5.35)	**16.20 (4.27)**
3	9.10 (2.29)	14.35 (3.05)	**20.03 (3.24)**
Overall	8.86 (3.52)	13.37 (4.23)	**19.17 (3.93)**

NASA-TLX Scores

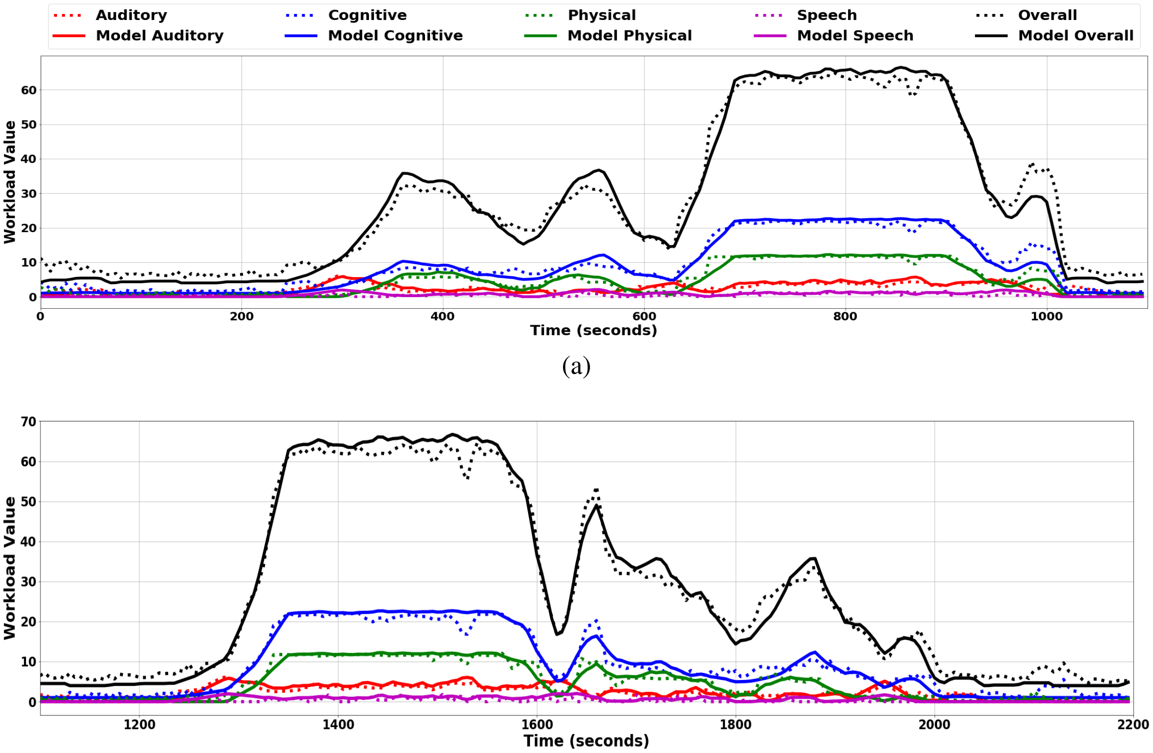
The presentation of each NASA-TLX score represents the unweighted score, while the overall NASA-TLX results represent the weighted aggregate. The NASA-TLX was administered after the single trial; thus, the results cannot be broken down by workload condition. The descriptive statistics for each NASA-TLX scale and the associated ANOVA result are provided in Table 5.11. The participants in Order 3 tended to rate each scale higher than the other participants, which was attributed to Order 3 having the most instances of the overload condition.

Table 3.67: NASA-TLX Workload Ratings by Workload Condition Ordering.

Order	**Effort**	**Frustration**	**Mental**	**Performance**	**Physical**	**Temporal**	**Overall**
1	45.62 (22.43)	35.0 (18.71)	53.12 (22.03)	40.0 (22.04)	41.88 (19.07)	48.75 (19.78)	48.5 (14.55)
2	51.0 (25.47)	18.0 (13.78)	52.0 (19.89)	33.5 (31.18)	41.88 (19.07)	44.0 (22.83)	47.37 (18.1)
3	69.0 (13.5)	42.0 (19.03)	75.5 (6.85)	57.0 (23.59)	54.0 (24.36)	74.0 (7.75)	68.8 (7.84)
ANOVA F(2, 25)	3.18, n.s.	5.09, $p = 0.01$	5.77, $p < 0.01$	2.12, n.s.	1.42, n.s.	8.00, $p < 0.01$	7.07, $p < 0.01$

3.6.4 Algorithm Analysis

This analysis evaluated the algorithm's ability to classify workload in emulated real-world conditions by training the algorithm on all of the supervisory evaluation's first day's data and testing it using all of the second day's data. The evaluation's second day mimicked real-workload conditions, as there were transitions between the workload conditions. It was expected that the algorithm's performance on the second day data will be similar to the previous algorithmic performance, as the workload conditions were similar. Hypothesis $\mathbf{H_4^{WL}}$ predicted that the algorithm's classification accuracy for the second day data will be within 5% of the workload generalizability accuracies in Table 3.38.



(a)

(b)

Figure 3.3: Emulated Real-World Conditions: Algorithm Estimates and IMPRINT Pro Model Values for the First Order. Top Graph is from Time = 0 to 1050 and Bottom Graph is from Time = 1050 to 2100.

The evaluation's second day emulated real-world conditions by incorporating workload transitions. There were three potential workload condition orderings for the second day task and a Kruskal-Wallis test determined that there were no significant differences across the orderings; thus, the results are presented as an aggregate of the three workload condition orders. The algorithm's estimates and IMPRINT Pro workload model values are plotted against time for only the first-order (UL-NL-OL-UL-OL-NL-UL) in Figure 3.3. The descriptive statistics for the algorithm's estimates and the IMPRINT Pro workload model values are presented in Table 3.68. The algorithm's estimates were within a standard deviation of the workload model values for overall workload and each workload component. A Kruskal-Wallis test determined that the algorithm's estimates differed significantly between the workload conditions for overall workload and each workload component.

Table 3.68: Emulated Real-World Conditions: Algorithm Estimated and Modeled Workload Descriptive Statistics.

Workload	**Algorithm**	**UL**	**NL**	**OL**	χ^2
Auditory	Model	1.16 (2.05)	2.14 (1.4)	4.02 (1.14)	67.5*
	Algorithm	1.89 (2.1)	2.26 (1.4)	3.58 (1.14)	30.7*
Cognitive	Model	1.40 (0.72)	8.29 (3.01)	22.03 (0.89)	81.2*
	Algorithm	2.05 (1.64)	8.70 (3.76)	21.16 (2.14)	79.0*
Physical	Model	0.14 (0.28)	4.29 (2.39)	11.73 (0.55)	78.6*
	Algorithm	0.76 (0.88)	4.61 (2.97)	11.49 (1.30)	77.3*
Speech	Model	0.32 (0.69)	0.75 (0.72)	0.96 (0.55)	33.0*
	Algorithm	0.07 (0.49)	0.29 (0.93)	0.79 (1.4)	33.0*
Overall	Model	5.61 (2.37)	26.74 (9.2)	63.69 (3.13)	79.7*
	Algorithm	7.38 (2.67)	27.14 (9.7)	61.97 (4.0)	79.1*

The distinct differences between each workload condition, as seen in Table 3.68, indicate that the algorithm may discriminate the conditions. The algorithm's classification accuracies by workload component and condition are provided in Table 3.69. The algorithm classified cognitive, auditory, and overall workload for each workload condition correctly at least 80% of the time. Physical workload was classified correctly at least 90% of the time for the underload and overload conditions, but was only classified correctly 75% of the time in the normal load condition. The underload classification accuracies for each workload component besides cognitive workload were within 5% of the corresponding workload generalizability classification accuracies (see Table 3.38). The normal cognitive, auditory, and physical workload accuracies were not within 5% of the workload generalizability accuracies, but were within 10% and were above 80%. The overload accuracies were within 5% of the workload generalizability accuracies for each workload component.

Table 3.69: Emulated Real-World Conditions: Algorithm's Classification Accuracy (%) by Workload Component and Condition. Delta (Δ) from Workload Generalizability Results. **Note:** Speech workload classification accuracies are not provided due to the trinary nature of the IMPRINT Pro speech workload model

Workload	**UL**	**NL**	**OL**	**Δ UL**	**Δ NL**	**Δ OL**
Cognitive	91.86	92.17	97.65	8.14	7.11	2.35
Auditory	90.80	84.10	88.4	0.77	5.92	3.61
Physical	96.21	76.00	96.43	3.35	10.67	3.57
Overall	95.29	93.11	99.89	4.71	6.72	0.11

Table 3.70: Emulated Real-World Conditions: Pearson's Correlation Coefficients for Within and Across Workload Conditions. Note: * Indicates $p < 0.05$.

Workload	**Within**			**Across**
	UL	**NL**	**OL**	
Auditory	0.90*	0.60*	0.67*	0.81*
Cognitive	0.43*	0.71*	0.14*	0.96*
Physical	0.38*	0.77*	0.10*	0.95*
Speech	0.10*	0.09*	0.08*	0.16*
Overall	0.58*	0.86*	0.48*	0.99*

Analysis of the Pearson's Correlation Coefficients between the algorithm's estimates and the corresponding IMPRINT Pro workload models determined the algorithm's ability to track workload variations within and across workload conditions, the correlation coefficients are presented in Table 3.70. Each correlation was positive and significant for each workload component and condition. The speech workload estimates were weakly correlated with the IMPRINT Pro workload models, which is again attributed to the trinary nature of the workload models. Smaller correlation coefficients tended to be produced in the overload condition in respect to the correlations for the same workload component.

3.6.5 Discussion

Humans transition between workload states in real-world domains, which requires the algorithm to accurately capture such transitions. Hypothesis $\mathbf{H_4^{WL}}$ focused on the algorithm's ability to classify workload under emulated real-world conditions and stated that the algorithm's classification accuracy for each workload component will be within 5% of the workload generalizability classification accuracies. The hypothesis was supported for all workload components, except cognitive for the underload condition, and for all components for the overload condition. The hypothesis was not supported for the normal load condition.

There is a delay between when a workload transition begins and when the algorithm's estimates reflect that particular transition. The delay appears to be approximately fifteen seconds when workload changes from a higher state to a lower state (e.g., overload to underload). However, the delay is less than ten seconds when the change transitions from a lower workload state to a higher state. These delays are due to the moving window size and the physiological response time. A thirty second window size was used based on relevant literature, but performance may increase if a different window size is used (see Appendix B.).

3.7 Peer-Based Evaluation Experimental Design

The peer-based evaluation analyzed differences in workload and task performance in human-robot peer-based teams [44, 49], where the evaluation was conducted by a prior PhD student: Caroline Harriott. The experimental design is repeated here for completeness, prior to presenting the algorithmic analysis. The peer-based evaluation results were suitable for analysis of the workload assessment algorithm. The peer-based evaluation scenario required training the participants as a civil support team member, and focused on identifying suspicious items, searching for hazardous materials, and collecting samples of both liquid and solid hazardous materials. The tasks were ordered in the general order of response steps to a disaster event and partnered a robot with a human [58].

The repeated measures design used the responder partner (human and robot) and the tasks as within-subjects elements. The participants completed four tasks with both a human partner (H-H) and a robot partner (H-R). Two sessions were completed, one with each partner, in which, all four tasks were completed.

3.7.1 Environment

All tasks were located close to one another on the same floor of an academic building at Vanderbilt University. Training took place in a small office with minimal distractions, in which the photo search task also occurred. The item search task occurred in the hallway where random people were able to walk through the environment. Sound traveled into the area from nearby offices, classrooms, and laboratories. The solid contaminant sampling task occurred in an engineering laboratory, isolated from foot traffic, and contained engineering equipment, lab benches, tables and tools. The liquid contaminant sampling task occurred in a virtual reality laboratory with two tables on which the task area was focused.

3.7.2 Apparatus

During the H-H condition, a second experimenter acted as the human teammate. A script dictated the verbal interactions between the participant and experimenter. The same male experimenter, who wore a reflective vest, was partnered with all participants.

Participants were instructed that the robot moved and spoke autonomously. The Pioneer 3-DX robot partner's navigation was controlled by an experimenter using line of sight teleoperation and a web-cam streaming video to the experimenter's laptop, unbeknownst to the participants. The robot spoke using a digital voice through on-board speakers. Participants donned a wireless microphone headset to amplify their voices when communicating with the robot. The same experimenter that controlled the robot, heard participant questions and responses, and used this knowledge to advance the robot's pre-programmed speech script. The robot's speech script was identical to the H-H condition's verbal script. When participants asked questions, the experimenter chose from a set of pre-programmed responses, repeated the robot's last statement, or wrote in a custom response.

3.7.3 Procedure

Participants completed the evaluation sessions (H-H and H-R) on different days. The mean days between participant sessions was 13.25 (St. Dev. = 10.77), where some second sessions were delayed due to cancellations and scheduling restrictions.

Upon arrival for the first session, participants completed a consent form and demographic questionnaire. The participants received an evaluation task briefing and were shown a 3 minute 40 second training video. Participants donned the Bioharness heart rate monitor, a Looxcie wearable video camera, the Shure microphone headset, a walkie talkie with ear piece, the Fitbit activity monitor, the Scosche Rhythm+ heart rate monitor, and a reflective vest. Participants were introduced to the responder partner, either human or robot, and began the first task, the photo search task.

The *photo search task* required identifying suspicious items in photographs of an area taken by a surveillance team. The participant was told that a team of robots previously entered the building to photograph rooms and areas that may contain victims, hazardous chemicals, suspicious items, or nothing to investigate. The participant used an Google Nexus 7 tablet computer running the Android mobile operating system to view, search, and edit the photographs. The incident commander, a remotely located experimenter, was responsible for sending photograph folders to the participant's tablet and notes from the investigation team to the responder.

The small office space included two adjacent tables and two chairs. During the H-H teaming condition, the human partner sat at the table on the right, while the participant sat at the table to the left. During the H-R teaming condition, the robot drove to a spot near the

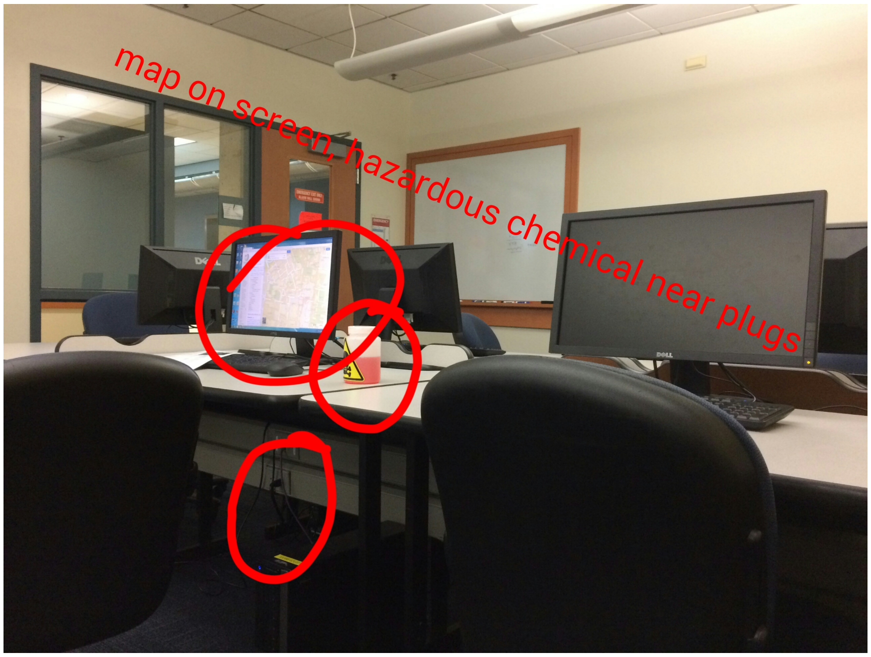


Figure 3.4: Post-edited Photograph

table on the right. Participants held the tablet during the task without constraint. No stand was used to prop the tablet on the table. The participant was able to swivel the chair to face the partner, but no movement outside of controlling the tablet was required.

The participant's tablet computer received the transmitted folders of photographs via a document sharing service, Box. Periodically, new folders containing three photographs each, appeared to be investigated. Once the participant reviewed the photograph and identified something suspicious, it was their job to edit the photograph in the photo-editing application, Aviary, by circling the item or adding a note to describe why the item must be investigated by a follow-up team. An instruction sheet provided information regarding how to use Box and perform the Aviary photo-editing steps.

Participants performed a training session using test photographs prior to starting the fifteen minute task and used the instruction sheet to learn how to use the tablet, Box application, and photo-editing system without time pressure. The participant took as long as he or she desired to train on three test photos and to ask questions during the training.

Once the fifteen minute task began, two sets of three photographs were searched during the low workload condition and four sets of three photographs were searched during the high workload condition. The photographs in each folder showed different angled perspectives of rooms, that included areas of the building in which the evaluation took place, such as study areas, classrooms, and a computer laboratory. The folders were presented in the same order for all participants within a workload condition, but different and comparable sets of photographs were used for the H-H and H-R sessions. An example photograph with post-participant annotations is provided in Figure 3.4.

A remotely located evaluator uploaded the photographs to the participant's folder at predefined times. During both workload conditions, the first photo set was uploaded immediately before the task began. During the low workload condition, the second set of photographs was uploaded seven minutes and thirty seconds into the task, while in the high workload condition, the four folders arrived every three minutes and forty-five seconds.

Some participants took longer to finish examining the folders. If a participant was not finished with Folder 1 by the time Folder 2 arrived, they simply opened Folder 2 when they finished Folder 1. If the team finished evaluating Folder 1 before Folder 2 arrived, the responder explained that the incident commander uploads folders as soon as photographs become available. An audible beep tone was used (the same tone in all tasks) to indicate when a new folder arrived. Teams viewed the photographs individually; thus, the beeps indicated time pressure for teams that were completing the task slowly. The teams who finished before the next folder arrived were able to know precisely when the new folder arrived and start the investigation immediately.

The second task, the *item search task*, required conducting an exhaustive search of an environment for potentially hazardous items, while gathering environmental context and air samples. The goal was to identify all hazardous items, while search-

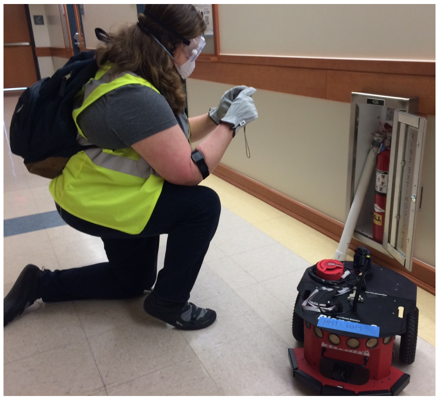


Figure 3.5: Participant performing the item search task.

ing as much of the assigned area within the time limit. The participant's role was to locate and photograph the items. The participants wore equipment to simulate personal protective equipment, including safety gloves, goggles, a dust mask, and a 10-pound backpack. The dust mask and backpack represented a civil support team's re-breather. The participant in Figure 3.5 is wearing the backpack, dust mask, goggles, and gloves.

This higher physical activity task involved walking around a hallway. The participants searched areas above the robot's sensors' field of view, while the human or robot partner scanned for hazards near the ground, collected air samples, and alerted the participant if any hazards or high air samples were detected. The team collaborated by discussing whether the detected items were suspicious.

Four items were investigated in the low workload condition, while eight items were investigated in the high workload condition. Each item investigated in the first session was identical, regardless of the participant's partner. A similar item set, placed in different locations in the same environment, was used for all participants during the second session, independent of the assigned partner. The first session items are depicted in Figure 3.6, where the starred items were only used in high workload tasks. The second session's items are depicted in Figure 3.7 with the same denotation for the additional high workload items.

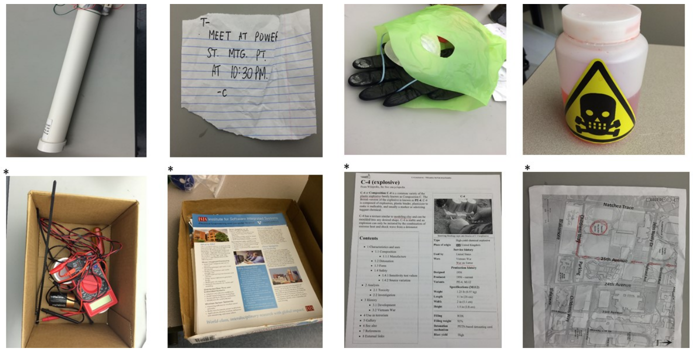


Figure 3.6: The first session item search task items. Starred items (*) were only used in the high workload condition. From top left, clockwise: a pipe bomb, cryptic note, bag with gloves and dust mask, hazardous liquid, suspiciously marked map of Vanderbilt campus, papers regarding C4 explosive use, box of advertisements (not suspicious), and a cardboard box filled with wires and suspicious material.

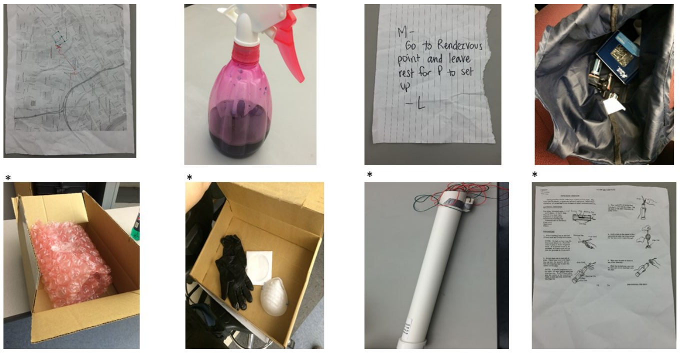


Figure 3.7: The search task's second session items. Starred items (*) represent the high workload condition. From top left, clockwise: suspiciously marked map of Nashville, suspicious liquid in spray bottle, cryptic note, bag filled with batteries and nails, papers instructing the fabrication of pipe bombs, pipe bomb, box with gloves suspicious envelope with white powder, and bubble wrap (not suspicious).

Either teammate determined whether the team needed to stop to investigate an item, though the responder teammate only did so when the participant missed an item that required investigation.

During the task, the responder paused the search in order to send current information to incident command and waited for a response indicating approval to continue the search. Incident command's messages were indicated with an audible beep at pre-defined times (i.e., 3:45, 7:30, 11:15, and 15:00). Teams completed the investigation in varying time lengths; thus, the beep created time pressure for the slower teams. The teams who finished before the next message arrived were able to know when the new notification was received.

The *solid contaminant sampling task*, the third task, required collecting samples from potentially hazardous solids in a room. The participant donned safety gloves and goggles. This task had a lower physical activity level, because the items were in close proximity and the participant did not wear the weighted backpack. The participants collected samples of the solids stored in various containers using a sterile collection kit by following guidelines audibly provided by the responder partner. These guidelines included detailed procedures that required strict compliance for maintaining safe and sterile sampling procedures. These evaluation steps were based on published government standards for the bulk sample collection of visible powders and suspected biological agents [61]. The partner indicated which solid to sample based on a message from the incident commander.

The partner provided hazard collection instructions and requested information from the participant regarding each potential hazard. Figure 3.8 depicts a participant completing this task.

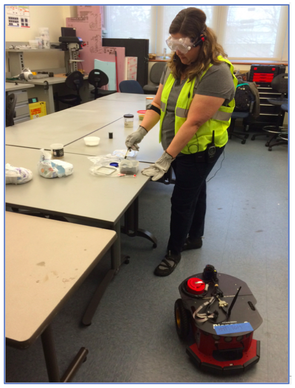


Figure 3.8: An H-R team sampling solid contaminants.

The team entered the room with hazards visible on a table. The hazards were containers (e.g., clear plastic storage container, glass jar, and a film canister) filled with unknown, colored solids (colored sand and baby powder). The participants used sampling kits, transported in a large gardening wagon, to collect small samples from subset of the hazards in the order requested by the incident commander. Each kit contained two sandwich-sized zip-lock plastic bags, one 4-ounce glass sample jar, one stainless steel scoopula, and one alcohol wipe, all placed in a gallon-sized storage bag and wrapped in a diaper to maintain sterility and protection from breakage. The wagon contained mailing labels to seal the bags and one permanent marker for writing the time on the seal. The participants were free to move the wagon, which was stationed at the room's entrance in the same place for each session.

Two samples were assigned in the low workload condition, and four samples were assigned in the high workload condition. The collection steps are presented in Figure 3.9. The messages regarding additional samples were marked by an audible beep., which added time pressure for slower teams.

The *liquid contaminant sampling task* required following hazardous material sample collection procedures that dictate conducting the collection from the least hazardous materials to the most hazardous materials [61]. The participants sampled liquids, while wearing safety gloves, goggles, a dust mask, and the 10-pound weighted backpack, simulating the

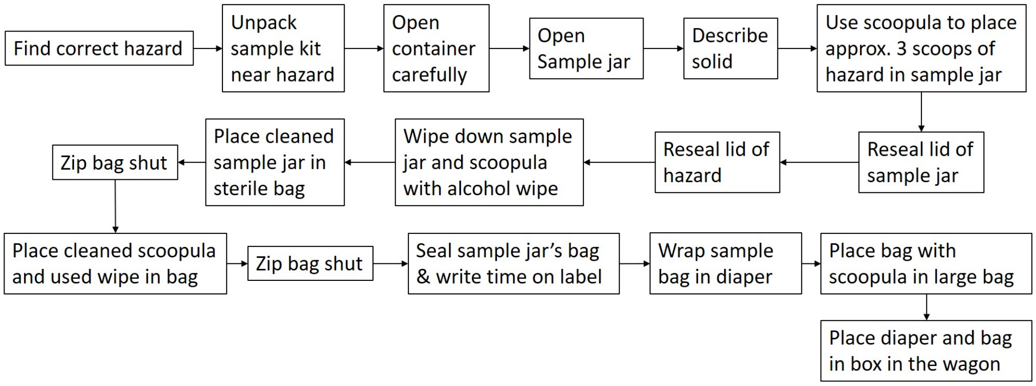

Find correct hazard
Unpack sample kit near hazard
Open container carefully
Open Sample jar
Describe solid
Use scoopula to place approx. 3 scoops of hazard in sample jar
Reseal lid of sample jar
Reseal lid of hazard
Wipe down sample jar and scoopula with alcohol wipe
Place cleaned sample jar in sterile bag
Zip bag shut
Place cleaned scoopula and used wipe in bag
Zip bag shut
Seal sample jar's bag & write time on label
Wrap sample bag in diaper
Place bag with scoopula in large bag
Place diaper and bag in box in the wagon

Figure 3.9: The steps completed for each contaminant sample collected in the solid contaminant sampling task.

cumbersome personal protective gear (see Figure 3.5). This task had higher physical activity, as it required walking around a larger laboratory space between the samples, and wearing the backpack. The gloves and mask added extra physical workload by increasing the task difficulty, such as opening plastic bags.

The sampling steps were similar to the solid contaminant sampling task. The highly structured protocol, based on government requirements, ensured sterile and safe collection of the “toxic liquid hazards”. The partner provided these structured guidelines and gathered information from the participant.

The participants entered the room with two tables, each with containers (e.g., sports water bottle, glass jar, Pepsi bottle) containing liquids (water dyed with various hues of food coloring). Nine containers were set out for all tasks in the same configuration for each evaluation during both sessions. The first session assigned a different subset of the nine containers than the second session.

The collection kits contained two sandwich-sized zip-lock plastic bags, one four-ounce glass sample jar, one plastic pipette, one drop cloth cut from plastic sheeting (approximately 2 feet by 1 foot), and one alcohol wipe. The drop cloth was placed under the sampling area to catch potential spillage. The kits were stored in a gallon-sized plastic bag and wrapped in a diaper to maintain sterility and protection from breakage. The kits were stored in the gardening wagon stationed near the room's entrance. The wagon also held the mailing labels for sealing the sample and a marker for labeling it. The participants were able to move the wagon at their discretion.

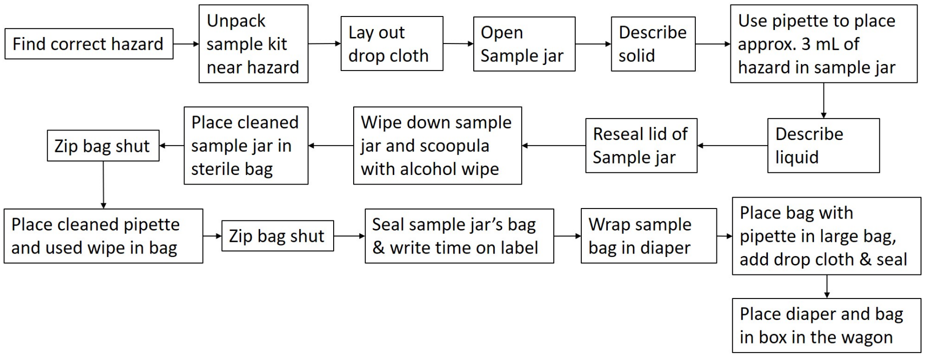

Find correct hazard
Unpack sample kit near hazard
Lay out drop cloth
Open Sample jar
Describe solid
Use pipette to place approx. 3 mL of hazard in sample jar
Describe liquid
Reseal lid of Sample jar
Wipe down sample jar and scoopula with alcohol wipe
Place cleaned sample jar in sterile bag
Zip bag shut
Place cleaned pipette and used wipe in bag
Zip bag shut
Seal sample jar's bag & write time on label
Wrap sample bag in diaper
Place bag with pipette in large bag, add drop cloth & seal
Place diaper and bag in box in the wagon

Figure 3.10: The steps completed for each liquid contaminant sample.

The liquid contaminant collection steps are presented in Figure 3.10. The participant and responder roles were similar to that in the solid contaminant sampling task. Two samples were assigned in the low workload condition, while four samples were assigned in the high workload condition. Beeps indicated when a new sample request arrived and added time pressure to slower teams.

After all four tasks, the participants returned for the second session, completing a questionnaire regarding their prior night's sleep, donning the equipment, and completing the evaluation in the identical task order as the first session, independent of responder partner.

3.7.4 Workload Modeling

The IMPRINT Pro workload models for the peer-based evaluation were developed by the prior PhD Student [44, 45] in a similar manner to the approach described in Chapter 3.4.4.

3.7.5 Participants

The eighteen participants completed the evaluation. Four male and five female participants worked with the human partner first (human partner first group) and five male and four female participants were paired with the robot first (robot partner first group).

The mean age of the human partner first group was 22.63 (St. Dev. = 6.16), with a range from 18 to 39, and the robot first group mean age was 21.89 (St. Dev. = 5.04), with a range from 18 to 34. The human partner first group rated their search and rescue experience on a Likert scale from 1 (little or no experience) to 9 (very experienced) as a median of 1, with a range from 1 to 9, while the robot partner first group rated their experience with a median of 4.5 with a range from 1 to 7. The human partner first group rated their robotics experience on the same scale as a median of 1, with a range from 1 to 9, while the robot partner first group rated their experience as a median of 3 with a range from 1 to 8. No significant difference existed between the two groups.

The human partner first group slept a median of 6.5 hours (range: 3 to 8) the night before their first session, while the robot partner first group slept 8 hours (range: 7 to 10) and slept significantly more hours, (Wilcox test, U = 8.5, p = 0.012). The human partner first group was awake for a median of 7.13 hours (range: 1 to 12.5) before the first session, while the robot partner first group was awake for 7.63 hours (range: 0.25 to 12.45). No significant differences existed.

The human partner first group slept a median of 6 hours (range: 3 to 9) the night before their second session, while the robot partner first group slept 7.5 hours (range: 3 to 9). The human partner first group was awake for a median of 8 hours (range: 5 to 13) before the second session, while the robot partner first group was awake for 8 hours (range: 2 to 10). No significant difference existed.

3.7.6 Metrics

The objective metrics included physiological responses (i.e., heart rate, respiration rate, heart rate variability, vector magnitude, steps taken), performance measures (i.e., subtask time, primary task response time, primary task reaction time, primary task failure rate, secondary task failure rate, a memory recall task), and the calculation of task density. The NASA-TLX and In-Situ workload ratings were collected for each task. An overview of the metrics collected during the peer-based evaluation is provided in Table 3.71. These metrics are not analyzed in this dissertation, as they have been validated previously [44].

Table 3.71: The Objective and Subjective Metrics for the Peer-Based Evaluation.

Metric Type	**Metric**
Algorithm	Heart-Rate Heart-Rate Variability Respiration-Rate Posture Noise Level Speech-Rate Pitch Voice Intensity
Other Objective	Skin Temperature Body Activity Steps Taken Vector Magnitude Task Density Subtask time Primary Task Response Time Primary Task Failure Rate Secondary Task Failure Rate Memory Recall Task
Subjective	In-Situ Workload Ratings NASA-TLX

The participants were assigned a secondary task to monitor a walkie-talkie for incoming messages from the Incident Commander for their team, Team 10. An example of a message was, “Incident Command to Team 10: there is a suspicious person running south on Anderson Road.” The participant was responsible for recognizing that the message was directed at Team 10, and repeating it to his or her partner when the communication arrived. The failure rate counted each time the participant did not report the message to his or her partner, and a partial failure was identified when the participant was able to respond to the Team 10 message, but relayed the message content incorrectly. During low workload tasks, there were eight total messages; two messages were for Team 10. During high workload tasks, six of the twenty-four messages were directed to Team 10.

3.8 Human-Robot Teaming Generalizability Analysis

Three different data sets were used to train separately the algorithm for the human-robot teaming generalizability analysis, resulting in three trained algorithm variants: the peer-based evaluation's data (**PEER**), the supervisory-based evaluation's data (**SUP**), and a combination of both evaluations' data (**BOTH**). The **PEER** data set consisted of twelve participants from only the human-robot teaming condition, (70% of the peer-based human-robot data), and the **SUP** data-set consists of fourteen participants, (50% of the supervisory-based data from day 1). The participants for each set were chosen at random. The number of participants differed due to equating the total number of data points used to train the algorithm variants, as more training data potentially increases performance and creates an unbalanced comparison. The **BOTH** data set aggregated the **PEER** and **SUP** data sets, which allowed for determining if a single trained algorithm may be used in both teaming paradigms.

Each trained algorithm was evaluated using the respective test data set, which included four peer evaluation participants, (30% of the peer-based data), and six supervisory evaluation participants, (16% of the supervisory-based data). The workload assessment algorithm was also trained on twenty supervisory-based participants and tested on the other ten participants, creating a 70/30 split for training and testing datasets.

Classification accuracy determines how well the workload assessment algorithm discriminates between workload conditions and generalizes across populations. The thresholds to determine the pariticipant's workload state are provided in Table 3.72. Hypothesis $\mathbf{H_5^{WL}}$ predicted that the algorithm will correctly classify workload states for each workload component and overall workload at least 80% of the time. This hypothesis tested the algorithm's population generalizability, as the testing set contains data from participants that are not in the training set. Classification accuracy is used to examine the algorithm's ability to generalize across human-robot teaming paradigms; however, it is expected that accuracy will decrease when the algorithm is not trained on the teaming relationship specific data. Thus, hypothesis $\mathbf{H_6^{WL}}$ predicted that the algorithm will correctly classify workload states for each workload component and overall workload at least 70% of the time, when not trained on the human-robot teaming relationship's specific data. The algorithm's ability to track workload shifts is an important factor for workload prediction and estimation, particularly for a system intended to adapt to the human's current and projected workload state. Hypothesis $\mathbf{H_7^{WL}}$ predicted that the algorithm's estimates will positively and significantly correlate with the IMPRINT Pro workload models, as a positive significant correlation indicates tracking of a workload shift.

Table 3.72: Model Ranges and Thresholds by Workload Condition and Component, Evaluation, and Overall Workload. Note: UL = Underload, NL = Normal Load, OL = Overload.

Evaluation	**Condition**	**Cognitive**	**Physical**	**Auditory**	**Overall**
Peer	Low (L)	0.00 - 9.35	0.00 - 11.71	-	0.00 - 32.25
	High (H)	0.46 - 8.78	0.93 - 11.77	-	10.97 - 30.35
Threshold	L-H	4.39	5.88	-	16.12
Supervisory	UL	1.00 - 2.03	0.00 - 0.70	0.00 - 6.00	4.00 - 9.27
	NL	4.93 - 11.63	4.04 - 6.86	0.00 - 6.00	13.80 - 39.19
	OL	21.20 - 22.76	11.20 - 12.50	0.00 - 6.00	59.20 - 66.81
Threshold	UL-NL	3.48	2.37	1.5	11.53
	NL-UL	16.42	9.03	2.9	49.20

The results for the three algorithm variants (**SUP**, **PEER**, **BOTH**) are presented by teaming evaluation. Only cognitive, physical, speech, and overall workload estimates were analyzed, as the peer-based evaluation did not capture workload metrics sensitive to auditory workload.

3.8.1 Peer-Based Evaluation Results

The peer-based evaluation contains low and high workload conditions. The IMPRINT Pro's modeled workload values and the algorithms' estimates for each workload condition are presented in Table 3.73. The **PEER** and **BOTH** trained algorithms' estimates are within a standard deviation of the IMPRINT Pro workload model values for each workload component. The **SUP** algorithm generally overestimated workload, due to the supervisory workload model values having a wider range, from minimum to maximum. The Kruskal-Wallis test determined that each trained algorithm's workload estimates significantly differed between workload conditions, except for the SUP algorithm's physical workload estimates.

Table 3.73: Workload Modeled and Algorithm Estimated Value's Descriptive and Kruskall-Wallis Statistics for the Peer Evaluation. Note: * Indicates P $<$ 0.001.

Workload	**Training**	**Workload Condition**		
		Low	High	χ^2
Cognitive	Model	3.58 (2.84)	6.14 (1.59)	61.75*
	SUP	9.27 (7.76)	14.38 (7.85)	49.12*
	PEER	3.77 (2.82)	5.95 (1.40)	35.92*
	BOTH	4.33 (3.22)	7.74 (3.62)	59.75*
Physical	Model	4.17 (2.63)	5.56 (2.14)	17.85*
	SUP	7.07 (4.93)	8.55 (3.34)	2.44
	PEER	4.28 (2.50)	5.35 (1.89)	10.94*
	BOTH	4.36 (2.54)	5.95 (2.02)	22.58*
Speech	Model	1.09 (0.91)	0.98 (0.67)	17.85*
	SUP	0.78 (1.4)	1.05 (1.5))	10.94*
	PEER	0.77 (1.4)	0.95 (1.5)	10.94*
	BOTH	0.72 (1.3)	1.19 (1.6)	10.94*
Overall	Model	14.33 (7.83)	21.31 (3.36)	56.26*
	SUP	25.24 (16.56)	32.86 (8.99)	45.24*
	PEER	14.64 (7.62)	21.31 (3.24)	45.60*
	BOTH	15.27 (7.61)	20.35 (2.92)	76.86*

A workload assessment algorithm must be able to detect various workload states in order to be viable for adaptive teaming systems. The algorithms' classification accuracy for low and high workload conditions is presented in Table 3.74. The **PEER** trained algorithm achieves $>$ 90% classification accuracy for each workload component. The **BOTH** and **SUP** algorithms classify cognitive and overall workload correctly at least 80% of the time, but both algorithms achieve lower physical workload accuracy. The **SUP** algorithm's physical workload accuracy is $<$ 70%, due to the algorithm receiving no prior training on the peer-based data. There is a negligible difference between the **PEER** and **BOTH** trained algorithms when classifying overall workload. It is interesting that the **PEER** algorithm achieves the lowest cognitive workload accuracy for the high workload condition, which is due to the **PEER** algorithm underestimating high cognitive workload. Underestimating high workload conditions means that the algorithm's estimates will be closer to the threshold between low and high workload, which decreases the classification accuracy.

Table 3.74: Classification Accuracy (%) for the Peer Evaluation.

Workload	**Training**	**Workload Condition**	
		Low	High
Cognitive	SUP	80.58	95.79
	PEER	**97.05**	94.52
	BOTH	92.12	**98.31**
Physical	SUP	67.90	58.88
	PEER	**90.50**	**90.35**
	BOTH	83.78	79.00
Overall	SUP	84.27	96.60
	PEER	**94.78**	96.55
	BOTH	93.84	**97.04**

Bold represents highest accuracy per column
Note: Speech workload accuracy is not shown

The peer evaluation workload conditions can be decomposed by task ($T1-T4$) and workload condition (L or H). The classification accuracy by peer evaluation task and workload condition is provided in Table 3.75. The **PEER** algorithm achieves the highest physical workload classification accuracy for each task-workload condition pair, while achieving the highest classification accuracy for six of the cognitive workload task-workload condition pairs and seven of the overall workload task-workload condition pairs. The **BOTH** algorithm achieves the second highest classification accuracy for each workload component, which is expected. The **PEER** and **BOTH** trained algorithms achieve $> 80\%$ accuracy when classifying cognitive or overall workload, with a negligible difference between the algorithms' overall workload classification accuracy for five tasks. The **SUP** algorithm achieves $> 80\%$ accuracy for six of the cognitive and all overall task-workload condition pairs. The **SUP** algorithm has poor physical workload classification.

Table 3.75: Classification Accuracy (%) by Peer-based Task.

Workload	Training	Peer Evaluation Task							
		$T1_L$	$T1_H$	$T2_L$	$T2_H$	$T3_L$	$T3_H$	$T4_L$	$T4_H$
Cognitive	SUP	93	89	84	96	71	96	72	**100**
	PEER	**100**	**93**	**95**	94	**96**	**100**	**96**	89
	BOTH	98	**93**	92	**100**	86	**100**	90	**100**
Physical	SUP	78	60	61	38	65	61	64	70
	PEER	**96**	**78**	**93**	**89**	**88**	**100**	**82**	**93**
	BOTH	**96**	68	88	71	78	83	71	91
Overall	SUP	94	89	89	96	77	**100**	75	**100**
	PEER	**98**	**93**	**92**	92	**88**	**100**	**100**	**100**
	BOTH	**98**	**93**	**92**	**98**	84	96	**100**	**100**

Note: Speech workload accuracy is not shown

Table 3.76: Peer Evaluation Correlation Coefficients for Within and Across Workload Conditions.

Workload	Training	Within		Across
		Low	High	
Cognitive	SUP	0.73*	0.58*	0.73*
	PEER	0.95*	0.85*	0.94*
	BOTH	0.79*	0.55*	0.74*
Physical	SUP	0.80*	0.63*	0.74*
	PEER	0.91*	0.92*	0.92*
	BOTH	0.87*	0.64*	0.79*
Speech	SUP	0.01	0.02	0.00
	PEER	0.09*	0.10*	0.09*
	BOTH	-0.15*	0.12	-0.07
Overall	SUP	0.84*	0.54*	0.82*
	PEER	0.97*	0.91*	0.97*
	BOTH	0.93*	0.44*	0.83*

The algorithms' tracking of workload shifts is analyzed using the Pearson's Correlation Coefficient. The correlation coefficients between the algorithms' estimates and workload models within and across the workload conditions for the entire evaluation are presented in Table 3.76. The Across column shows that each trained algorithm produced significant correlations across workload conditions, meaning that each algorithm can track large workload shifts (e.g., a shift from low to high workload). The significant correlations in the Within columns indicate that each algorithm can also track small workload shifts (i.e., a change in workload within the low workload condition). The **PEER** trained algorithm produced the highest correlations, followed by the **BOTH** algorithm.

It is not surprising that the **PEER** trained algorithm produced the best results. The **BOTH** algorithm also achieved high performance, even when incorporating the supervisory training data. The **SUP** algorithm performance is acceptable for the cognitive and overall workload assessments, but is unable to accurately assess physical workload.

3.8.2 Supervisory-Based Evaluation Results

The IMPRINT Pro modeled workload values and the algorithms' estimates for each supervisory condition (underload (UL), normal load (NL), and overload (OL)) are presented in Table 3.77. The **SUP** and **BOTH** trained algorithms' estimates are within one standard deviation of each IMPRINT Pro workload model for each workload component. The **PEER** trained algorithm overestimates the underload condition and underestimates the normal load and overload conditions for each workload component. The Kruskal-Wallis test determined that the algorithms' estimates significantly differed between conditions for each workload component.

The workload assessment algorithm was designed to classify various workload levels. The algorithms' supervisory classification accuracies are presented in Table 3.78. Each trained algorithm achieves a 100% classification accuracy for the underload condition and has $\geq 85\%$ accuracy for the normal load condition for each workload component. The **SUP**

Table 3.77: Workload Modeled and Algorithm Estimated Descriptive Statistics and Kruskal-Wallis for the Supervisory Evaluation.

Workload	Training	Workload Condition			
		UL	NL	OL	χ^2
Cognitive	Model	1.43 (0.72)	8.19 (2.52)	21.93 (0.79)	80.7*
	SUP	1.19 (1.85)	8.02 (3.57)	21.28 (2.16)	80.4*
	PEER	2.35 (0.11)	5.52 (1.09)	6.95 (0.03)	76.57*
	BOTH	0.46 (0.43)	7.68 (2.52)	20.22 (1.68)	78.91*
Physical	Model	0.11 (0.24)	4.30 (2.13)	11.73 (0.57)	80.4*
	SUP	0.53 (1.08)	4.30 (2.57)	11.30 (0.82)	79.2*
	PEER	0.18 (0.45)	3.38 (1.22)	3.91 (0.07)	58.75*
	BOTH	0.25 (0.59)	4.53 (2.15)	11.18 (0.68)	77.23*
Speech	Model	0.41 (0.72)	0.61 (0.61)	0.95 (0.59)	63.0*
	SUP	0.12 (0.61)	0.39 (1.06)	0.85 (1.44)	80.46*
	PEER	0.10 (0.44)	0.28 (0.74)	0.58 (0.96)	58.75*
	BOTH	0.11 (0.69)	0.31 (1.11)	0.69 (1.59)	77.23*
Overall	Model	4.53 (1.97)	27.21 (9.36)	63.19 (4.73)	80.00*
	SUP	4.04 (0.91)	27.66 (6.08)	63.25 (1.39)	79.12*
	PEER	5.87 (0.85)	23.19 (4.56)	40.01 (1.15)	79.12*
	BOTH	4.05 (1.30)	26.51 (5.64)	60.66 (2.58)	79.12*

and **BOTH** algorithms achieve ≥ 90% classification accuracy for the overload condition, while the **PEER** algorithm is unable to classify the condition due to the PEER training data's maximum value being lower than the threshold (49.20, Table 3.36) between the normal load-overload task condition pair. The supervisory evaluation cannot be easily decomposed into separate tasks that permits analysis of the individual tasks for the classification accuracy; thus, such results are not presented.

The algorithms' ability to track workload shifts is analyzed using the Pearson's correlation coefficient analysis. The correlations for within and across workload conditions are presented in Table 3.79. The significant correlations in the Across column indicate that the algorithms can track large workload shifts; however, each trained algorithm has difficulty tracking small workload shifts in the underload and overload conditions for the cognitive workload component. The negative cognitive workload correlations are due to the static nature of the IMPRINT Pro workload models. A human's physiological signals typically

Table 3.78: Classification Accuracy (%) for the Supervisory Eval.

Workload	**Training**	**Workload Condition**		
		UL	NL	OL
Cognitive	SUP	**100**	93.74	**100**
	PEER	**100**	89.37	0
	BOTH	**100**	**94.96**	98.33
Physical	SUP	**100**	95.53	**100**
	PEER	**100**	**96.64**	0
	BOTH	**100**	93.85	**100**
Overall	SUP	**100**	**100**	**100**
	PEER	**100**	**100**	0
	BOTH	**100**	**100**	**100**

oscillate, resulting in oscillating workload estimates; thus, the estimates oscillate around the static workload model values, resulting in negative correlations.

Table 3.79: Supervisory Evaluation's Correlations for Within and Across Workload Conditions.

Workload	**Training**	**Within**			**Across**
		UL	NL	OL	
Cognitive	SUP	-0.07	0.45*	-0.04	0.99*
	PEER	-0.59*	-0.02	-0.44	0.86*
	BOTH	0.16	0.17	0.67*	0.97*
Physical	SUP	0.92*	0.85*	0.19	0.98*
	PEER	0.91*	0.86*	0.65*	0.83*
	BOTH	0.91*	0.84*	0.02	0.98*
Speech	SUP	0.10*	0.09*	0.08*	0.16*
	PEER	0.06*	0.08*	0.03	0.11*
	BOTH	0.10*	0.12*	0.02	0.15*
Overall	SUP	0.96*	0.89*	0.53*	0.99*
	PEER	0.95*	0.93*	0.96*	0.98*
	BOTH	0.83*	0.84*	0.79*	0.99*

The **SUP** algorithm achieved the highest performance, followed by the **BOTH** algorithm for the supervisory-based relationship. The **PEER** algorithm performs well, except when classifying the overload state.

3.8.3 Human-Robot Teaming Generalizability Discussion

Real-world dynamic domains, such as first response, contain multiple tasks with varying workload compositions. It is essential that a workload assessment algorithm classify workload for each task within the domain in order to be viable in real-world dynamic domains, which requires assessing overall workload and each workload component. The state-of-the-art algorithms typically only assess cognitive workload [50]; thus, limiting the algorithms to cognitively focused tasks. The presented workload assessment algorithm overcomes this limitation by estimating overall workload and each workload component.

$\mathbf{H_5^{WL}}$ evaluates the algorithms' ability to classify cognitive, physical, and overall workload by stating that each algorithm will correctly classify workload states $\geq$ 80% of the time when trained and tested on data from the same evaluation. The hypothesis was upheld for the **SUP** and **PEER** algorithms, when the algorithms are classifying data from their corresponding evaluation (i.e., the **PEER** algorithm classifying the peer evaluation's data). The hypothesis was not upheld for the **BOTH** algorithm for the peer evaluation's physical workload component, which indicates that the supervisory data negatively impacts the classification accuracy. However, the **BOTH** algorithm achieves a 79% physical workload classification accuracy, which is slightly below the threshold.

The long term objective is to develop an adaptive workload system that generalizes across human-robotic interaction paradigms; thus, $\mathbf{H_6^{WL}}$ evaluates the algorithms' classification accuracy across interaction paradigms. The hypothesis was partially supported. The peer evaluation's **SUP** algorithm achieves >80% classification accuracy when classifying cognitive and overall workload, but achieves poor classification for physical workload. Likewise, the supervisory evaluation's **PEER** algorithm achieves high classification accuracy for the underload and normal load conditions for overall workload and each workload component, but cannot classify the overload condition due to the peer training data's maximum value being below the threshold value separating the normal load and overload conditions. A limitation is that the algorithm must be trained on a similar set of workload conditions in order to generalize across human-robotic interaction paradigms.

$\mathbf{H_7^{WL}}$ evaluated the ability to track workload shifts and was supported for all algorithms when using the peer evaluation data set. The supervisory data set analysis only provided partial support, as large workload shifts were tracked, but not small shifts in the underload and overload conditions. The underload and overload models are typically static (low variance), due to limits on the models (e.g., the underload model cannot go below zero). However, a human's physiological signals can oscillate; thus, the algorithms' estimates fluctuate around the static models, resulting in low correlations.

The developed workload assessment algorithm typically achieves the highest accuracy when using data from the same interaction paradigm. However, the **BOTH** algorithm that incorporates an equal number of data points from both data sets does achieve the highest accuracy in some cases, demonstrating the benefit of training using both datasets. The **BOTH** algorithm's high accuracy indicates the potential for using a single algorithm for both types of tasks and interaction paradigms. However, it is uncertain how the algorithm will perform in other human-robot interaction paradigms, beyond the evaluated paradigms.

3.9 Peer-Based Task Generalizability Analysis

It is important that a workload assessment algorithm generalizes across tasks within a human-robot teaming paradigm, as it is infeasible to train an adaptive teaming system on every task a human may complete within the teaming paradigm. Only data from the peer-based evaluation was used to assess the algorithm's task generalizability, as the supervisory evaluation's tasks are concurrent and not easily separated. The workload assessment algorithm was cross-validated using a leave-one-task-out approach [12], which creates four trained algorithms. Each algorithm was trained on three peer-based tasks using data from all eighteen peer-based participants. The testing set was the data from all eighteen participants for the peer-based task that was not used for training. Individual differences may not affect the results, as data from each participant was used to train each algorithm. The thresholds for classifying the workload conditions are the same as Table 3.36.

It is expected that classification accuracy will decrease when the algorithm is not trained on a specific peer-based task; thus, hypothesis $\mathbf{H_8^{WL}}$ predicts that the algorithm will correctly classify workload states for each workload component and overall workload at least 70% of the time when trained on three peer-based tasks and tested on the remaining task. Although classification accuracy will decrease, it is expected that each trained algorithm will track workload shifts for an unforeseen task. Hypothesis $\mathbf{H_9^{WL}}$ predicts that each trained algorithm's estimates will significantly and positively correlate with the IMPRINT Pro workload models for each peer-based task.

3.9.1 Task Generalizability Results

Only the results for the algorithms' performance on each testing set are presented. For example, a result for $T1_L$ represents task one as the testing set, where the algorithm was trained on tasks two, three, and four. The algorithms' estimates and model's mean and standard deviation's are presented in Table 3.81. The algorithm underestimated cognitive workload for the high workload tasks and overestimated physical workload for the low workload tasks. The algorithm's overall workload estimates were close to the model's values.

Correctly classifying workload for unforeseen tasks is essential to an adaptive teaming system, as it is infeasible to collect data for all tasks. The algorithms' classification accuracies by task and workload component are provided in Table 3.80. The algorithms correctly classify workload $\geq 70\%$, except for $T4_L$, due to the PEER algorithm having low physical workload classification accuracy.

Table 3.80: Generalized Task Classification Accuracy (%) by Peer Task and Workload Component.

Workload	Peer Evaluation Task							
	$T1_L$	$T1_H$	$T2_L$	$T2_H$	$T3_L$	$T3_H$	$T4_L$	$T4_H$
Cognitive	78.32	74.79	93.22	100.00	88.83	96.04	88.82	89.29
Physical	89.51	73.08	73.45	90.78	85.11	100.00	62.94	70.83
Overall	96.50	93.16	90.40	88.83	81.91	100.00	99.41	100.00

Table 3.81: Generalized Task Estimates and Model Values by Peer Task, Cognitive, and Physical Workload.

Workload	**Training**	**Peer Task**							
		$T1_L$	$T1_H$	$T2_L$	$T2_H$	$T3_L$	$T3_H$	$T4_L$	$T4_H$
Cognitive	Model	3.61 (2.48)	5.79 (1.08)	1.78 (2.29)	5.26 (2.67)	3.85 (3.02)	6.49 (0.96)	3.67 (3.05)	6.49 (0.97)
	Algorithm	3.03 (1.53)	4.36 (.54)	1.77 (2.16)	4.06 (2.12)	3.75 (1.52)	4.95 (0.34)	3.05 (2.09)	4.92 (0.55)
Physical	Model	3.61 (2.48)	5.79 (1.08)	1.78 (2.29)	5.26 (2.67)	3.85 (3.02)	6.49 (0.96)	3.67 (3.05)	6.49 (0.97)
	Algorithm	5.27 (3.59)	6.77 (1.74)	4.07 (2.82)	5.23 (0.87)	3.31 (2.74)	4.08 (1.96)	4.68 (2.4)	6.16 (1.35)
Speech	Model	1.18 (0.70)	1.03 (0.64)	1.58 (0.67)	1.43 (0.77)	0.44 (0.55)	0.78 (0.55)	0.44 (0.55)	0.78 (0.55)
	Algorithm	0.98 (1.51)	1.18 (1.54)	1.20 (1.59)	1.31 (1.61)	0.59 (1.27)	0.98 (1.50)	0.58 (1.25)	0.78 (1.40)
Overall	Model	14.92 (7.12)	20.62 (3.07)	11.78 (2.29)	19.35 (3.47)	12.28 (9.5)	20.53 (2.83)	15.17 (9.61)	24.00 (2.94)
	Algorithm	15.17 (6.11)	19.72 (1.80)	11.82 (5.57)	18.85 (2.07)	13.11 (7.31)	19.06 (1.81)	13.49 (8.41)	21.11 (1.36)

The algorithm's ability to track workload shifts for each task is analyzed using the Pearson's correlation coefficient, which are presented in Table 3.82. Each algorithms' estimates correlated significantly with each corresponding model, which demonstrates that the algorithms' estimates correctly reflects workload shifts for unforeseen tasks.

Table 3.82: Generalized Task Pearson's Correlation Coefficients by Peer Task and Workload Component.

Workload	Peer Evaluation Task							
	$T1_L$	$T1_H$	$T2_L$	$T2_H$	$T3_L$	$T3_H$	$T4_L$	$T4_H$
Cognitive	0.95*	0.70*	0.97*	0.95*	0.82*	0.48*	0.82*	0.48*
Physical	0.98*	0.85*	0.41*	0.63*	0.80*	0.96*	0.64*	0.88*
Speech	-0.11*	0.02	-0.19*	0.01	0.10*	0.09*	0.14*	0.01
Overall	0.98*	0.88*	0.89*	0.95*	0.96*	0.92*	0.96*	0.89*

3.9.2 Task Generalizability Discussion

A workload assessment algorithm's ability to generalize across tasks within the same human-robot teaming paradigm is essential for an adaptive teaming system, as it is difficult to collect training data for each task within a paradigm. Hypothesis $\mathbf{H_8^{WL}}$ focused on evaluating the algorithm's task generalizability and is supported for the cognitive and overall workload classification. The hypothesis is only partially supported for physical workload classification, as the algorithm fails to correctly classify physical workload for $T4_L$ at least 70% of the time. This result is due to task $T4_L$ substantially differing from the other peer-based tasks. Partially supporting Hypothesis $\mathbf{H_8^{WL}}$ for physical workload classification is not detrimental to the algorithm's task generalizability, as the average physical workload accuracy is above 70%. Further, none of the state-of-the-art workload assessment algorithms generalize across tasks; thus, showing that the algorithm generalizes across tasks for cognitive and overall workload, is a significant contribution.

The workload assessment algorithm's estimates must accurately reflect a shift in workload regardless of the task, which represents hypothesis $\mathbf{H_9^{WL}}$. This hypothesis is upheld, as the algorithm's estimates significantly correlate to the workload model values for each task. Having high confidence that a shift in the algorithm's estimates correctly represents a shift in workload allows an adaptive teaming system to better gauge the affect an adaptation has on the human's workload state. For example, if the human is overloaded and the workload estimates show the workload level is decreasing appropriately, then the adaptive system can be confident that there is no need for adaptation.

The task generalizability analysis demonstrates the workload assessment algorithm's ability to generalize across similar tasks for the **PEER** dataset. It is unclear if the algorithm can generalize across tasks for other human-robot teaming domains or what happens if an unforeseen task differs substantially from the tasks on which the algorithm was trained.

3.10 Summary

A workload assessment algorithm capable of estimating overall workload and its contributing components (i.e., cognitive, physical, visual, auditory, and speech) was developed. The algorithm used machine-learning techniques to estimate cognitive, physical, auditory, and speech workload, while IMPRINT Pro workload models were used to estimate the visual workload component. Data from two human-robot teaming evaluations (supervisory-based and peer-based) were used to train and validate the algorithm. Objective and subjective data were analyzed from the supervisory-based evaluation in order to validate the workload conditions experienced by the participants. The developed algorithm's ability to estimate workload for the supervisory-based and peer-based evaluations was analyzed, while its ability to generalize across human-robot teaming paradigms and peer-based tasks was also analyzed.

The results highlight that the algorithm achieves high classification accuracy when trained and tested on data from the same human-robot teaming paradigm. Further, the algorithm generalized across the participant populations within each human-robot teaming evaluation. The algorithm's task generalizability was analyzed using tasks from the peer-based evaluation, which showed that the algorithm generalized across peer-based tasks for cognitive and overall workload estimation. The impact of training the algorithm on multiple data sets was also analyzed by training the algorithm on two human-robot teaming data sets, where the results demonstrated that there is minimal negative impact on the algorithm's performance. Finally, the algorithm performed well in emulated real-world conditions, where workload transitions rapidly between workload levels.

An adaptive teaming system designed to improve human-robot team performance may use the developed algorithm to understand the human's workload state and the associated contributing factors. The system can have high confidence ($\geq$ 80%) in the algorithm's workload state classifications, even with unknown users. This confidence in the workload component classifications will allow the system to understand how an adaptation may affect the human. Other state-of-the-art workload assessment algorithms do not provide workload component information; thus, limiting their viability for an adaptive teaming system. The developed algorithm is the first algorithm that provides information about each workload component and the overall workload state.

The algorithm provided overall workload estimates, rather than just classifications, which can be used by a system to trigger an adaptation that may prevent an underload or overload workload state from occurring. These estimates correlated strongly with the corresponding IMPRINT Pro overall workload models, giving an adaptive system confidence that a trend in the overall workload estimates reflects the human's overall workload trend accurately. There is less confidence in the workload component estimate trends in the underload or overload conditions. This lower confidence is not detrimental to the algorithm's inclusion into an adaptive system, as the system can account for the lower confidence.

The algorithm had some difficulty estimating physical workload in general, relative to the other workload components. This difficulty is due to posture magnitude and respiration-rate being only somewhat sensitive to physical workload [50], while heart-rate is confounded by cognitive workload. The physical workload metrics may not be sensitive enough to the fine-motor and tactile demands, which comprise the majority of the supervisory evaluation's physical workload task components. There is little research related to the incorporated workload metrics and their sensitivity to fine-motor and tactile demands, as most research has focused on gross-motor demands [20, 22]. Additional workload metrics, such as electromyography captured by a wearable device (i.e. the Myo device) may be needed in order to capture the fine-motor and tactile task demands [95].

Another common theme was that the algorithm's speech workload estimates did not significantly correlate with the IMPRINT Pro workload models, which is attributed to the workload model's composition. The supervisory-based evaluation consisted primarily of complex speech (as determined by IMPRINT Pro), as the response to air traffic control messages contained more than three words. Simple speech (1 to 2 words) occurred during the in-situ workload ratings. Better algorithmic performance can be achieved by producing the IMPRINT Pro complex speech workload value (4.0) when the participant is speaking and zero in the absence of speech, but this approach limits the algorithm to scenarios that only contain complex speech. Further, IMPRINT Pro's three speech workload model values may be insufficient for properly capturing the actual workload experienced by the participant. Speech-rate and pitch varied across workload conditions [64], which demonstrated that participants experienced multiple speech workload levels. Thus, a more sophisticated speech workload model, rather than using IMPRINT Pro built-in model, may be necessary to accurately reflect the participant's expected workload. Developing such a model is outside the scope of this dissertation.

The current workload assessment algorithm uses IMPRINT Pro workload models to estimate the visual workload component. Visual workload is difficult to estimate in dynamic domains (e.g., first response), as eye-tracking metrics require known focus regions. Thus, for real-world, dynamic domains, a visual workload model will continue to be used, at least until appropriate technology is available.

Decomposing overall workload into its corresponding components may allow for better adaptations, but the resulting components are not completely separable (e.g., an increase in auditory workload will result in an increase in cognitive workload). Additionally, relying on physiological measures sensitive to multiple workload components may hinder algorithm performance (e.g., is an increase in heart-rate due to cognitive or physical workload). Thus, contextual features are needed to achieve high workload assessment algorithm performance, as the features provide valuable insight into the task's composition. If heart-rate increases and contextual features show that the current task is mainly cognitive, then the heart-rate increase is most likely due to cognitive workload, rather than physical workload.

The reliance on contextual features requires an activity recognition algorithm to be used in conjunction with the developed workload assessment algorithm; thus, limiting the algorithm to domains for which activity recognition is applicable. Additional workload training data from multiple task domains may decrease the algorithm's reliance on activity recognition; however, it is expected that the workload algorithm will always need an activity recognition algorithm, due to the nature of dynamic task domains. Future work as discussed in Chapter VI, will investigate incorporating additional workload training data.

Diagnostically assessing workload is imperative for the adaptive teaming system, as assessing overall workload and each workload component determines what distinct components are contributing to the overall workload state, or simply why the human is in the current workload state. Identifying the distinct contributing components permits targeted adaptions to the components in order to normalize the workload state. An adaptation based solely on the overall workload state cannot target a specific workload component; thus, the adaptation may be ineffective (i.e., reallocating a task that is not a primary contributor to the workload state). A diagnostic workload assessment permits projection of future interactions or task allocations onto the human's current workload state to determine and account for the potential impact.

Chapter 4

Real-Time Workload Assessment

The analysis in Chapter III demonstrated the developed workload assessment algorithm's ability to estimate workload in a post-hoc fashion. An adaptive teaming system requires real-time workload estimates in order to adapt its interactions intelligently. Further, the previous analysis focused on a stationary supervisory-based human-robot team; however, there are supervisory-based environments that require physical movement throughout the task environment (e.g., a nuclear power-plant control room). This chapter details a non-stationary evaluation, where human workload was assessed in real-time.

4.1 Real-Time Workload Assessment

The workload assessment algorithm described in Chapter 3.1 was used in real-time to estimate overall workload and each workload component every five seconds. The IMPRINT Pro workload models were used to estimate the visual and speech components. The real-time speech workload assessment was implemented for the adaptive teaming system only (Chapter 5.2). Estimating workload every five seconds permits balancing the feature extraction computational time and the minimum update rate required to determine how an interaction occurs. Estimating workload too quickly increases computational time, while estimating it too slowly may lead to a system adapting its interactions based on outdated information.

4.1.1 Experimental Design

The within-subjects real-time workload assessment evaluation manipulated workload (i.e., underload, normal load, and overload) as the independent variable. The dependent variables included physiological, performance, and subjective metrics. Participants completed one 52.5 minute trial using an adapted version of the NASA MATB-II, where the trial consisted of seven consecutive 7.5 minute workload conditions. Three workload condition orderings were used:

- UL-NL-OL-UL-OL-NL-UL
- NL-OL-UL-OL-NL-UL-NL
- OL-UL-OL-NL-UL-NL-OL

These orderings were chosen to ensure that each workload state transition (i.e., UL-NL, OL-UL) occurred and the orderings mimic the orderings from the supervisory-based evaluation (Chapter 3.4). IMPRINT Pro was used to model each condition ordering prior to conducting the evaluation.

4.1.2 Environment

The original NASA MATB-II required participants to remain stationary, but there are supervisory-based environments that require movement throughout the environment (e..g, a nuclear power-plant). Thus, the NASA MATB-II was adapted to require movement throughout the task environment by physically separating each NASA MATB-II task. This physical layout is depicted in Figures 4.1 and 4.2. Each NASA MATB-II task had a computer monitor dedicated to a particular task, where the computer monitors were stationed such that the participant was unable to visually see more than two tasks simultaneously. This visual hindrance ensured that participants walked around the task environment, instead of staying in one place. The required equipment (e.g., joystik or a keyboard) to complete each task was placed in front of the respective computer monitor. The table surfaces were approximately 4 ft. from the floor. Each monitor was connected to a single computer, which had a NVIDIA 1080 TI graphics card. Participants were free to tilt the computer monitors up or down in order to accommodate height differences.

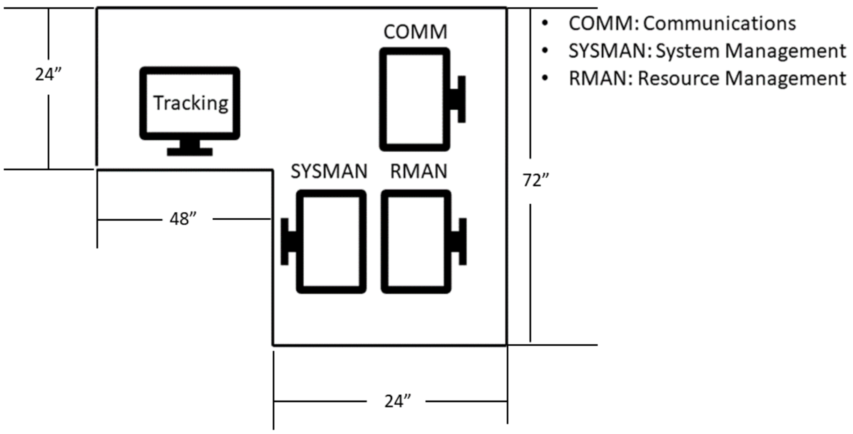


Figure 4.1: Physical Layout of the Adapted NASA MATB-II

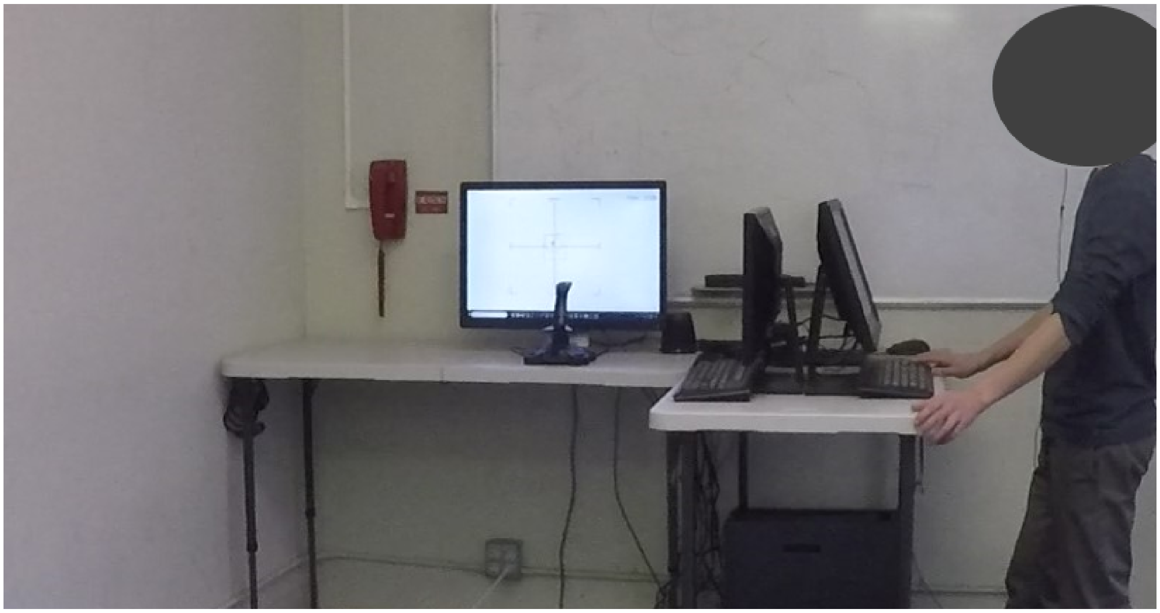
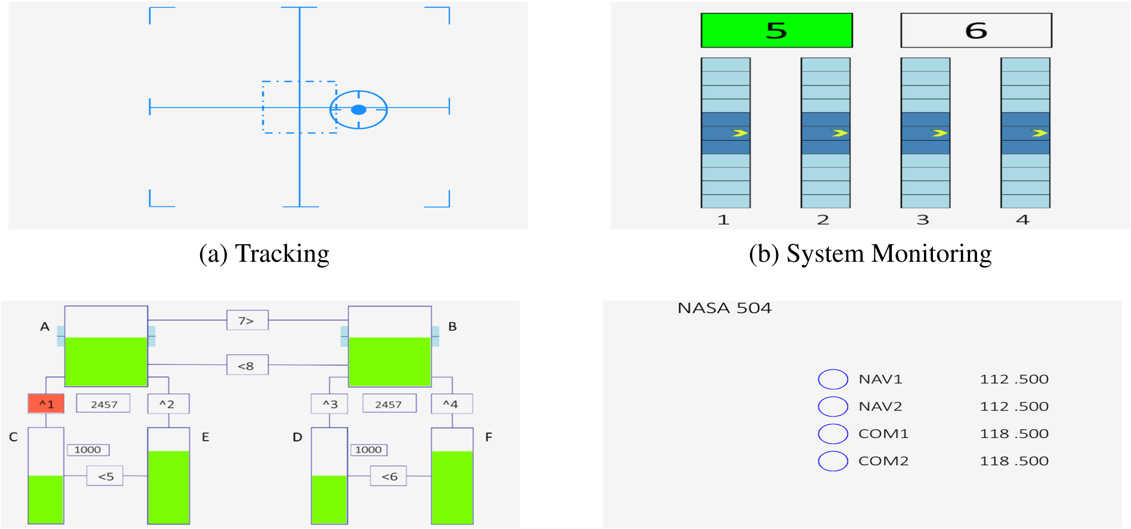


Figure 4.2: Real-Time Evaluation Task Environment.

The physically expanded version of the NASA MATB-II was coded using Python and PyGame in order to have more control over the task environment. The same task parameters (e.g., tank fuel rates) from the original NASA MATB-II were reimplemented. Information regarding the fuel pumps’ rates and task scheduling was omitted in order to reduce the visual screen clutter. Each computer monitor screen is depicted in Figure 4.3. The tasks are explained in Chapter 3.4. The objects in each figure were drawn in PowerPoint in order to improve object quality (e.g., thicker lines), rather than using pictures from the original NASA MATB-II.



(a) Tracking

(b) System Monitoring

(c) Resource Management

(d) Communications

Figure 4.3: The NASA MATB-II Tasks

The evaluation occurred in an empty classroom on Oregon State University's campus.

4.1.3 Workload Models

The workload models were developed in a similar manner to the process described in Chapter 3.4.4. However, a few changes were required. First, there was inherent uncertainty related to which tasks the participant will complete and when, as participants moved around the task environment and can only complete two tasks simultaneously, given the tasks' physical proximities to one another. IMPRINT Pro provides tools for modeling this uncertainty, which were used. Second, the participant's gross motor (walking around the task environment) was added to the IMRPRINT Pro models, where the walking was anchored to a gross motor value of 1.0. Lastly, instances of the communication task (Table

3.3) were removed from the underload condition, as the communication requests may keep participants vigilant, which was undesired.

The number of task instances per minute for each workload condition are provided in Tables 4.1 - 4.3, which were the similar to the first 7.5 minutes of Tables 3.2 - 3.4.

Table 4.1: Number of Tasks Per Minute for the Underload Condition. Note: Vertical Bold Line Designates when an In-Situ Workload Rating was assessed (7:00 minutes) and TRCK = tracking, SYS = system management, COMM = communication, and RESP = response to the communication

Task	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
TRCK	0	0	0	0	0	0	0	0
SYS	0	0	0	1	0	0	0	1
COM	0	0	0	0	0	0	0	0
RESP	0	0	0	0	0	0	0	0
Total	0	0	0	1	0	0	0	1

Table 4.2: Tasks Per Minute for the Normal Load Condition

Task	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
TRCK	1	1	0	1	0	1	1	0
SYSM	4	4	3	1	4	3	3	2
COM	2	2	1	2	0	0	1	1
RESP	1	2	0	1	0	0	1	1
Total	8	9	4	5	4	4	6	4

Table 4.3: Tasks Per Minute for the Overload Condition

Task	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
TRCK	1	1	1	1	1	1	1	1
SYSM	19	18	15	15	21	17	18	20
COM	3	4	5	4	3	4	4	3
RESP	2	2	4	3	2	3	3	2
Total	25	25	25	23	27	25	26	26

The developed workload assessment algorithm required contextual information (Chapter 3.1) regarding the participant's current task focus. This information was derived from the IMPRINT Pro workload models, which were assumed to be correct.

4.1.4 Procedure

The participants completed a consent form and a demographic questionnaire upon arrival, after which participants were fitted with a BioPac Bioharness BT, a Schure Microphone, and two Myo devices. This BioPac Bioharness BT was a newer version than the ones described in Chapters 3.2.1.6 and 3.2.2.6. However, the device was unable to collect skin-temperature data. The Myo devices were fitted on the participant's forearms and collected acceleration and electromyography data. This Myo data is intended to help develop an activity recognition algorithm, as described in Chapter 6.2.

A 15-minute training session occurred before the 52.5-minute trial. Participants completed the NASA-TLX and a post-session questionnaire after finishing the trial. In-situ workload ratings were verbally administered at 7 minutes into the trial and every 7.5 minutes after the initial rating.

4.1.5 Participants

The thirty-one participants (14 females and 17 males) had a mean age of 27.61 (St. Dev. = 9.06), with an age range from 18 to 64 years. Fourteen participants held a high school degree, nine participants held an undergraduate degree, seven participants held a master's degree, and two participants held a doctorate degree. The majority of participants (twenty-one) indicated that they played video games for three or fewer hours per week. Participants also rated their video game skill level on a Likert scale (1-little to 9-expert) on average as 4.38 (St. Dev. = 2.29). Fifteen participants had 0 oz. of caffeine the day of the experiment, while thirteen participants drank at most 16 oz. Participants exercised on average 4.94 (St. Dev. = 4.02) hours a week.

The participants slept an average of 6.95 (Std. Dev. = 1.23) hours the night before the experiment and an average of 7.13 (Std. Dev. = 1.35) hours two nights prior. The participants' average stress and fatigue levels rated on a Likert scale (1-little to 9-extreme) was 2.77 (Std. Dev. = 1.50) and 3.19 (Std. Dev. = 1.82), respectively.

4.1.6 Metrics

The objective and subjective metrics were collected throughout the evaluation, where an overview is provided in Table 4.4. The objective metrics consisted of physiological signals (e.g., heart-rate and respiration-rate) and performance metrics. The physiological metrics were captured by the BioHarness BT, which was different from the BioHarness in the supervisory and peer evaluations. The BioHarness BT does not capture skin-temperature, but it does capture the other metrics (i.e., heart-rate, heart-rate variability, respiration-rate, and posture magnitude). The noise-level and speech-based metrics were captured by the same equipment in supervisory evaluation (Chapter 3.2.1).

The physically separated NASA MATB-II collected performance metrics in a similar manner to the original NASA MATB-II. The tracking task's performance was measured as the error in pixels between the center of the cross-hairs and the center of the object (Figure 4.3 a), which was collected every second. The system monitoring task's performance was determined by response time and failure rate. Response time was the number of seconds a participant took to click on a light or gauge, once the respective light or gauge went out of range. Failure rate represented the number of out of range lights and gauges that were not corrected within fifteen seconds, which is the default threshold for the NASA MATB-II. The resource management task's performance was determined by the time fuel Tanks A and B were out of range (i.e., the fuel levels were not between 2,000 and 3,000 units), where the fuel levels of each tank were collected every second. The number of failed communication requests (i.e., the participant failed to respond or the number of times the radio was tuned to the wrong frequency) determined the communications task performance.

The in-situ workload ratings and NASA-TLX subjective metrics were collected.

Table 4.4: The Objective and Subjective Metrics for the Real-Time Evaluation.

Metric Type	**Metric**
Algorithm	Heart-Rate Heart-Rate Variability Respiration-Rate Posture Noise Level Speech-Rate Pitch Voice Intensity
Other Objective	Body Activity Arm Acceleration Forearm Electromyography Tracking Task: Tracking Error System Monitoring Task: Reaction Time System Monitoring Task: Failure Rate Resource Management Task: Time-in-Range Communications Task: Reaction Time
Subjective	In-Situ Workload Ratings NASA-TLX

4.2 Hypotheses

The remainder of this chapter analyzes the performance and subjective metric data and validates the developed workload assessment algorithm's ability to estimate workload in real-time correctly. The analyses are broken into three sections: Performance Metrics, Subjective Metrics, and Algorithm Analysis. The algorithm analysis section contains three trained algorithm variants: **SUP**, **RT**, and **POST-HOC**. The **SUP** and **RT** variants were used in real-time and correspond to the algorithm being trained solely on the supervisory-based evaluation's data and a combination of the supervisory-based evaluation's and a portion of the real-time evaluation's data, respectively. The **POST-HOC** variant was trained on the real-time evaluation's data, but was not used in real-time.

Several hypotheses were formed for these analyses, where an overview of the hypotheses is presented in Table 4.5. The hypotheses are explained in more detail in the following sections.

Table 4.5: Chapter 4 Hypotheses

Analysis	**Hypothesis**	
Metric Comparison	$\mathbf{H_1^{RT}}$	The objective metrics and the In-Situ workload ratings will trend in a similar manner across conditions for the Real-Time Evaluation and the Supervisory Evaluation's Day 2.
Real-Time	$\mathbf{H_2^{RT}}$	The **SUP** and **RT** algorithm's estimates will be within a standard deviation of the corresponding IMPRINT Pro workload model values.
	$\mathbf{H_3^{RT}}$	The **SUP** and **RT** algorithms will classify each workload state at least 80% of the time.
	$\mathbf{H_4^{RT}}$	The **SUP** and **RT** algorithm's estimates will significantly and positively correlate with the corresponding IMPRINT Pro workload models.
Post-Hoc	$\mathbf{H_5^{RT}}$	The **POST-HOC** algorithm's estimates will be within a standard deviation of the corresponding IMPRINT Pro workload model values.
	$\mathbf{H_6^{RT}}$	The **POST-HOC** algorithms will classify each workload state at least 80% of the time.
	$\mathbf{H_7^{RT}}$	The **POST-HOC** algorithm's estimates will significantly and positively correlate with the corresponding IMPRINT Pro workload models.
	$\mathbf{H_8^{RT}}$	The **POST-HOC** algorithm will be more accurate than the **RT** algorithm.

4.3 Real-Time Evaluation Results

Five analyses were conducted for the real-time evaluation: workload metrics, performance metrics, subjective metrics, real-time analysis and post-hoc analysis. The workload, performance, and subjective metric analyses were performed in a similar manner to the corresponding supervisory-based evaluation analyses (Chapters 3.4 and 3.5). The algorithm analysis focused on the developed workload assessment algorithm's real-time capabilities. The algorithm requires contextual information about the participant's current task, which was assumed to be known correctly.

4.3.1 Workload Metrics Analysis

It was expected that the value representing the highest workload for each metric will occur during the overload condition (e.g, the lowest heart-rate variability value will occur during the overload condition). Additionally, the metrics were expected to trend across the workload conditions in a similar manner to how the metrics trended in the Supervisory Evaluation's Day 2 (Chapter 3.5.1), as predicted by hypothesis $\mathbf{H_1^{RT}}$. This hypothesis tests if the workload conditions for each evaluation (Supervisory and Real-Time) are similar and that the metrics reflect the conditions. Significance testing was not performed for testing the hypothesis, as the metrics will differ significantly between the two evaluations, due to the physical nature of the real-time evaluation.

Heart-Rate

The associated descriptive statistics for heart-rate are provided in Table 4.6. The highest values tended to occur during the overload condition. A two-way MANOVA determined that there was a significant effect on workload (F(2,26) = 179.29, $p < 0.01$) and on order (F(2,26) = 3764.00, $p < 0.01$). There was also a significant interaction between order and workload condition (F(10,48) = 23.16, $p < 0.01$).

Table 4.6: Heart-Rate Descriptive Statistics. The highest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	92.28 (14.34)	**93.13 (14.24)**	**93.95 (14.47)**
2	81.63 (11.89)	82.38 (12.02)	**83.41 (12.57)**
3	85.54 (15.28)	87.12 (16.57)	**90.02 (17.54)**
Overall	87.25 (14.58)	86.61 (14.52)	**88.91 (15.54)**

Heart-Rate Variability

Heart-rate variability decreases as workload increases. The means and standard deviations by workload condition and ordering are provided in Table 4.7. The lowest values occurred during the overload condition and the values differed significantly between the workload conditions (F(2,26) = 67.91, $p < 0.01$). There was a significant effect on the workload orderings (F(2,26) = 2101.40, $p < 0.01$) and a significant interaction between the orderings and conditions F(10,48) = 7.42, $p < 0.01$).

Table 4.7: Heart-Rate Variability Descriptive Statistics. The lowest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	0.67 (0.12)	0.67 (0.14)	**0.66 (0.13)**
2	0.75 (0.14)	0.75 (0.18)	**0.74 (0.18)**
3	0.73 (0.15)	0.73 (0.19)	**0.71 (0.19)**
Overall	0.71 (0.14)	0.72 (0.17)	**0.70 (0.18)**

Respiration Rate

An increase in workload may be reflected by a decrease in respiration rate. The associated descriptive statistics by workload condition and ordering are provided in Table 4.8. The lowest respiration-rates occurred during the overload condition for each order. Respiration rate differed significantly between the workload conditions (F(2,26) = 2155.04, $p < 0.01$) and between the orders (F(2,26) = 128.98, $p < 0.01$). There was a significant interaction between the orders and the workload conditions (F(10, 48) = 100.87, $p < 0.01$).

Table 4.8: Respiration-Rate Descriptive Statistics. The lowest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	18.60 (3.76)	18.45 (4.04)	**15.49 (5.03)**
2	18.90 (4.65)	17.87 (4.43)	**16.02 (4.91)**
3	18.21 (4.68)	16.50 (5.23)	**15.78 (5.87)**
Overall	18.62 (4.28)	17.76 (4.55)	**15.77 (5.28)**

Posture Magnitude

Posture magnitude is associated with an increase in workload. The means and standard deviations by workload condition and ordering are provided in Table 4.9. The largest values occurred either during the underload or normal load conditions, which was unexpected. A two-way MANOVA determined that there was a significant main effect on workload (F(2,26) = 101.01, $p < 0.01$) and on the ordering (F(2,26) = 2691.46, $p < 0.01$). There existed a significant interaction between the workload conditions and orderings (F(10, 48) = 2.90, p = 0.02).

Table 4.9: Posture Magnitude Descriptive Statistics. The highest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	**-13.82 (57.36)**	-15.1 (56.46)	-17.37 (55.98)
2	**7.61 (7.36)**	6.18 (7.63)	2.40 (9.81)
3	4.55 (9.83)	**4.97 (10.07)**	0.37 (12.39)
Overall	-2.70 (40.29)	**-0.49 (33.21)**	-4.62 (34.23)

Noise-Level

An increase in auditory workload may be represented by an increase in noise-level. The associated descriptive statistics are presented in Table 4.10. The highest values occurred during the overload condition and there was a significant effect by workload condition (F(2,26) = 19.22, $p < 0.01$). No significant effect was found for the orderings or for the interaction between the orderings and workload conditions.

Table 4.10: Noise-Level Descriptive Statistics. The highest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	61.33 (202.28)	66.94 (198.62)	**68.79 (155.79)**
2	59.43 (187.02)	65.84 (199.79)	**68.57 (153.86)**
3	53.44 (130.74)	66.61 (197.85)	**70.66 (163.51)**
Overall	58.93 (183.3)	66.32 (199.03)	**69.34 (157.76)**

Speech-Rate

It was expected that participants will speak faster as workload increased. The means and standard deviations by workload condition and ordering are provided in Table 4.11. There was no general trend for speech-rate across the workload conditions, but there may be an order effect. A two-way MANOVA determined that speech-rate differed significantly across the workload conditions (F(2,26) = 14.78, $p < 0.01$) and orderings (F(2,26) = 6620, $p < 0.01$), while a significant interaction occurred between the two (F(10, 48) = 83.70, $p < 0.01$).

Table 4.11: Speech-Rate Descriptive Statistics. The highest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	**3.31 (1.11)**	3.21 (1.16)	3.16 (1.18)
2	2.48 (1.72)	2.55 (1.61)	**2.57 (1.43)**
3	1.00 (0.81)	1.25 (0.98)	**1.57 (1.08)**
Overall	**2.76 (1.52)**	2.60 (1.52)	2.48 (1.40)

Voice Intensity

The participant's voice intensity was expected to increase as workload increased, as seen in 4.12. The highest voice intensity tended to occur during the overload condition. Voice intensity differed significantly across the conditions (F(2,26) = 181.99, p $<$ 0.01) and orders (F(2,26) = 87.15, p $<$ 0.01). There was also a significant interaction between the workload conditions and orders (F(10, 48) = 7.25, p $<$ 0.01).

Table 4.12: Voice Intensity Descriptive Statistics. The highest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	125.68 (114.35)	162.2 (134.9)	**179.31 (148.98)**
2	147.41 (96.92)	177.88 (128.35)	**203.89 (139.68)**
3	138.49 (132.61)	138.55 (132.05)	**170.31 (133.26)**
Overall	142.14 (109.6)	167.81 (130.79)	**187.09 (138.98)**

A human's dominant speaking frequency or pitch is expected to increase with an increase workload. The associated descriptive statistics are provided in Table 4.13. The highest pitches tended to occur in the overload condition for each order. The participants' pitches differed significantly across the workload conditions (F(2,26) = 272.06, p $<$ 0.01) and between the orders (F(2,26) = 1969.46, p $<$ 0.01). There was also a significant interaction between the workload conditions and orders (F(10,48) = 25.97, p $<$ 0.01).

Comparison to the Supervisory Evaluation

An overview of the metric trends for an increase in workload for each evaluation (Real-Time and Supervisory) is provided in Table 4.14. A no trend rating was given if the metric did not increase or decrease across the three workload conditions. The expected column

Table 4.13: Pitch Descriptive Statistics. The highest values for each order are in **Bold**.

Order	Underload	Normal Load	Overload
1	248.63 (113.56)	281.55 (107.47)	**300.81 (100.4)**
2	307.13 (85.04)	320.88 (86.96)	**328.39 (86.26)**
3	175.41 (72.79)	198.3 (84.6)	**232.87 (94.79)**
Overall	259.55 (107.21)	288.34 (103.03)	**291.19 (101.5)**

details the metric's corresponding theoretical trend, as shown in Table 2.1. Four of the eight metrics trend across the workload conditions the same way for the two evaluations. Two of the metrics (i.e., heart-rate and respiration-rate) had no trend for the supervisory evaluation's second day, but had a discernible trend for the real-time evaluation. These trends for the real-time evaluation reflected the metric's predicted response to workload (i.e., heart-rate increases as workload increases). Posture magnitude had no trend for the real-time evaluation, but increased as workload increased for the supervisory evaluation. This result was attributed the participants walking during the real-time evaluation and sitting during the supervisory evaluation. Speech-rate decreased as workload increased for the real-time evaluation, but the metric increased for the supervisory evaluation. This result may be attributed to individual differences that exists between the participants.

Table 4.14: The Workload Metric Trends for the Real-Time Evaluation and the Supervisory Evaluation's Second Day.

Metric	Expected	Real-Time	Supervisory Day 2
Heart-Rate	Increases	Increases	No Trend
Heart-Rate Variability	Decreases	Decreases	Decreases
Respiration-Rate	Decreases	Decreases	No Trend
Posture Magnitude	Increases	No Trend	Increases
Noise-Level	Increases	Increases	Increases
Speech-Rate	Increases	Decreases	Increases
Voice-Intensity	Increases	Increases	Increases
Pitch	Increases	Increases	Increases

Workload Metrics by Workload Transition

The real-time evaluation contained workload transitions in order to emulate real-world conditions. It is necessary for the objective workload metrics to be sensitive to these transitions in order for the workload assessment algorithm to estimate the transitions accurately. The Spearman correlation coefficients between each workload metric and the IMPRINT Pro overall workload model by workload transition are provided in Table 4.15. The same window size (60 seconds) for each workload transition was chosen as in Chapter 3.6.1.

Table 4.15: The Spearman Correlation Coefficients between Each Workload Metric and the IMPRINT Pro Overall Workload Model by Workload Transition. Note: * indicates p < 0.05.

Metric	**UL-NL**	**UL-OL**	**NL-UL**	**NL-OL**	**OL-UL**	**OL-NL**
Heart-Rate	0.12*	0.23*	0.14*	-0.04	0.21*	0.09
Heart-Rate Variability	0.01	-0.16*	-0.08	0.04	-0.16*	-0.05
Noise Level	0.44*	0.74*	0.55*	0.53*	0.76*	0.29*
Posture Magnitude	-0.06	-0.28*	-0.05	-0.22*	-0.24*	-0.15*
Respiration-Rate	-0.33*	-0.36*	-0.55*	0.06	-0.63*	-0.57*
Speech-Rate	0.04	0.21*	0.24*	-0.02	0.35*	0.21*
Pitch	0.13*	-0.18*	-0.39*	0.00	-0.44*	-0.28*
Voice Intensity	0.04	0.11*	0.27*	-0.09	0.32*	0.19*

Overall, there were significant correlations between the workload metrics and overall workload model when workload transitioned between the underload and overload conditions and vice-versa. These correlations also reflected the expected trend for each metric, (i.e, the corresponding correlations for heart-rate were positive). However, the majority of the correlations are considered to be small (r < 0.30), which was to be expected. The overall workload model ranges from 4 - 65, while the workload metrics have much narrower ranges (i.e., heart-rate variability ranges from 0.55 to 1.10).

The smallest correlations occurred when workload transitioned between the normal load and overload conditions and vice-versa. The speech-based metrics and cardiovascular metrics did not significantly correlate with the overall workload models during these transitions. This result is attributed to the intrinsic nature of these specific workload transitions (i.e., normal load to overload and overload to normal load). The participants were already experiencing significant workload in the normal load condition; thus, participants did not have sufficient resources to allocate to the tasks when workload transitioned to the overload condition. Thus, these metrics may have hit a “red line”, where an increase in workload is no longer associated with a response in the workload metrics.

Respiration-rate and noise level typically produced the highest correlations across the workload transitions. Respiration-rate is sensitive to multi-task environments, which may have attributed to its correlations with the IMPRINT Pro overall workload model. The correlations between noise level and overall workload is likely due to noise level being a task demand measure (Table 2.1), not a physiological response. Additionally, noise-level is not sensitive to individual differences between participants.

The correlation analysis focused on each individual metric's response to changes in overall workload, but the correlations were typically not strong ($r > 0.5$). This result is not concerning, as the workload assessment algorithm relies on all of the objective workload metrics. The neural networks within the algorithm can learn the interdependicies between the metrics in order to better estimate workload.

Each workload metric was sensitive to one or more workload components, which permitted analyzing each metric's correlation to the corresponding IMPRINT Pro workload component model for the workload transitions. This correlation analysis was limited to metrics that the developed workload assessment algorithm used. Noise level was the only metric used to estimate auditory workload, where the corresponding Spearman correlation coefficients are provided in Table 4.16. Significant and positive correlations occurred for four workload transitions (i.e, UL-OL, NL-UL, OL-UL, OL-NL), where the largest correlation coefficients happened when workload transitioned vastly (i.e., from underload to overload). Noise level did not significantly correlate with the auditory workload model when workload transitioned from the underload state to the normal load state and from the normal load state to the overload state.

Table 4.16: The Spearman Correlation Coefficients between Each Workload Metric and the IMPRINT Pro Auditory Workload Model by Workload Transition. Note: * indicates $p < 0.05$.

Metric	**UL-NL**	**UL-OL**	**NL-UL**	**NL-OL**	**OL-UL**	**OL-NL**
Noise Level	0.1	0.38*	0.15*	0.05	0.53*	0.25*

The heart-rate, heart-rate variability, and noise level metrics were used to estimate cognitive workload. The Spearman correlation coefficients between these metrics and the IMPRINT Pro cognitive workload model are provided in Table 4.17. Noise level had the largest correlations out of the three workload metrics, where the correlations were positive and significant for each workload transition. Heart-rate and heart-rate variability significantly correlated with cognitive workload for the UL-OL, OL-UL, and OL-NL workload transitions. These significant correlations were the expected sign (positive or negative), which demonstrates that the heart-rate and heart-rate variability metrics where sensitive to the workload transitions. However, these two metrics did not significantly correlate when workload transition from underload to normal load, normal load to underload, or from normal load to overload.

Table 4.17: The Spearman Correlation Coefficients between Each Workload Metric and the IMPRINT Pro Cognitive Workload Model by Workload Transition. Note: * indicates $p < 0.05$.

Metric	**UL-NL**	**UL-OL**	**NL-UL**	**NL-OL**	**OL-UL**	**OL-NL**
Heart-Rate	0.09	0.17*	0.1	-0.02	0.15*	0.11*
Heart-Rate Variability	-0.08	-0.16*	-0.1	0.0	-0.14*	-0.15*
Noise Level	0.4*	0.74*	0.51*	0.4*	0.76*	0.25*

The workload assessment algorithm relied on three workload metrics (heart-rate, posture magnitude, and respiration-rate) for the physical workload estimations. The correlation coefficients between each metric and the IMPRINT Pro physical workload model are provided in 4.18. Respiration-rate was the most sensitive to physical workload, as the metric had the largest coefficients. However, respiration-rate did not significantly correlate with the workload model during the normal load to overload transition. Significant correlations occurred between heart-rate and the workload model for two of the six workload transitions. This result may be attributed to the participants experiencing cognitive workload and physical workload, as heart-rate is sensitive to both components. Posture magnitude produced similar correlation coefficients to heart-rate, which may demonstrate that these two workload metrics were not sensitive to physical workload.

Table 4.18: The Spearman Correlation Coefficients between Each Workload Metric and the IMPRINT Pro Physical Workload Model by Workload Transition. Note: * indicates p < 0.05.

Metric	UL-NL	UL-OL	NL-UL	NL-OL	OL-UL	OL-NL
Heart-Rate	0.04	0.16*	-0.02	-0.02	0.11*	-0.07
Posture Magnitude	0.01	-0.17*	-0.03	-0.13*	-0.09	-0.09
Respiration-Rate	-0.21*	-0.35*	-0.39*	0.1	-0.35*	-0.11*

Speech-rate, pitch, and voice intensity were expected to correlate to speech workload, where the corresponding Spearman correlation coefficients are presented in Table 4.19. The pitch and voice intensity workload metrics significantly correlated with the speech workload model for each workload transition, where pitch had the largest correlation coefficients. Speech-rate did not significantly correlate with the workload model for the normal load to overload workload transitions, but did produce significant correlations for the remaining workload transitions.

Table 4.19: The Spearman Correlation Coefficients between Each Workload Metric and the IMPRINT Pro Speech Workload Model by Workload Transition. Note: * indicates p < 0.05.

Metric	UL-NL	UL-OL	NL-UL	NL-OL	OL-UL	OL-NL
Speech-Rate	0.17*	0.18*	0.14*	0.09	0.3*	0.21*
Pitch	-0.24*	-0.36*	-0.37*	-0.23*	-0.45*	-0.37*
Voice Intensity	0.16*	0.17*	0.17*	0.12*	0.3*	0.16*

Overall, at least one workload metric significantly correlated with the IMPRINT Pro workload component model for each workload component and transition. However, the lowest correlations tended to occur when workload transitions from the normal load to the overload condition. This result may be attributed to the physiological metrics no longer responding to an increase in workload, as the participants did not have any remaining resources to allocate to the workload increase. Additionally, the smallest correlations occurred for the physical and speech workload components. This result indicates that additional metrics or improved workload models are needed for these workload components.

4.3.2 Performance Metrics Analysis

It was expected that the participants will perform better during the normal load condition than the other conditions, as underload and overload can negatively impact task performance. The best performance value was bolded for each table.

Tracking Task Performance

Task performance for the tracking task is determined using the average RMSE between the center of the cross-hairs and center of the object to be tracked. The resulting descriptive statistics are presented in Table 4.20. The underload statistics are not provided, as that condition does not require the participant to track the object. The participants achieved the highest performance during the normal load condition. A two-way MANOVA determined that there was a significant main effect on the workload condition (F(1,24) = 1372.63, $p < 0.01$) and ordering (F(2,23) = 21.39, $p < 0.01$). There was also a significant interaction (F(2,23) = 17.16, $p < 0.01$) between the workload conditions and orderings.

Table 4.20: Tracking Task Performance Descriptive Statistics for Average Root-Mean Squared Error. Note: Lower is Better.

Order	Normal Load	Overload
1	**160.28 (108.35)**	199.49 (112.67)
2	**157.58 (104.48**)	207.86 (118.41)
3	**140.59 (93.83)**	200.28 (111.60)
Overall	154.23 (10.344)	202.52 (114.22)

Resource Management Task Performance

The time in seconds that the fuel tanks were out of range determined task performance for the resource management task, where the higher the value represents poorer performance. The descriptive statistics are presented in Table 4.21. The participants maintained the fuel levels better during the underload condition for Order 1, the normal load condition for Order 2, and the underload and normal load conditions for Order 3. There was a significant main effect on workload condition (F(2,23) = 914.83, $p < 0.01$) and ordering (F(2,23) = 798.01, $p < 0.01$). Additionally, there was a significant interaction (F(2,23) = 29.11, p <

0.01) between conditions and orderings.

Table 4.21: Resource Management Task Performance Descriptive Statistics for Time in Range (%). Note: Higher is Better.

Order	Underload	Normal Load	Overload
1	**0.95 (0.21)**	0.90 (0.29)	0.78 (0.41)
2	0.74 (0.44)	**0.78 (0.41**)	0.61 (0.48)
3	**0.87 (0.33)**	**0.87 (0.33)**	0.75 (0.43)
Overall	**0.86 (0.33)**	**0.85 (0.36)**	0.72 (0.44)

System Monitoring Task Performance

The system monitoring task contained two task performance metrics: mean reaction time and failure rate. The descriptive statistics for mean reaction time are presented in Table 4.22. The lowest reaction times typically occurred during the underload condition. A two-way MANOVA determined that reaction time significantly differed by workload condition (F(2,23) = 18.28, $p < 0.01$), but not by the condition ordering. There was no significant interaction between the workload conditions and orderings.

Table 4.22: System Monitoring Task Performance Descriptive Statistics for Mean Reaction Time in Seconds. Note: Lower is Better.

Order	Underload	Normal Load	Overload
1	**4.22 (3.87)**	5.77 (6.01)	6.01 (4.27)
2	5.97 (4.73)	**5.49 (3.75)**	6.28 (4.24)
3	**4.32 (4.16)**	5.37 (3.52)	6.25 (4.29)
Overall	**4.79 (4.27)**	5.53 (3.78)	6.19 (4.27)

The average success rates for the system monitoring task by workload condition and order are provided in Table 4.23. The highest success rates typically occurred during the underload condition for each order, except for order 3, which had the highest success rates during the normal load condition. There was a significant main effect on workload condition (F(2,23) = 7.09, $p < 0.01$). There was no significant effect for order or the interaction between the orders and workload conditions.

Communications Task Performance

Table 4.23: System Monitoring Task Performance Descriptive Statistics for Success Rate (%). Note: Higher is Better.

Order	Underload	Normal Load	Overload
1	**70.00 (23.45)**	68.33 (22.90)	50.57 (21.80)
2	**68.18 (33.71)**	65.98 (26.49)	41.79 (22.45)
3	68.51 (25.61)	**79.40 (7.32)**	60.02 (12.95)
Overall	68.88 (27.27)	**70.79 (21.27)**	50.18 (20.60)

The descriptive statistics for the communications task's reaction times by condition and ordering are provided in Table 4.24. No results are presented for the underload condition, as no communications request occurred. The lowest reaction times occurred during the overload condition, which may be attributed to participants focusing on the communications task more during the overload condition. A two-way MANOVA determined that there was a significant main effect on workload (F(1,24) = 21.80, $p < 0.01$) condition and condition ordering (F(1,24) = 3.45, $p = 0.03$). There was also a significant interaction (F(1,24) = 4.80, $p = 0.01$) between workload condition and ordering.

Table 4.24: Communication Task Performance Descriptive Statistics for Reaction Time. Note: Lower is Better.

Order	Normal Load	Overload
1	11.56 (2.07)	**9.34 (4.41)**
2	9.97 (2.27)	**9.28 (4.04)**
3	10.41 (1.78)	**9.83 (4.36)**
Overall	10.49 (2.18)	**9.52 (4.27)**

A failure occurred during the communications task when participants did not react within 15 seconds or incorrectly tuned the radio. The descriptive statistics for the participants' success rate by workload condition and ordering are provided in Table 4.25. Overall, there was no significant difference between the two workload conditions. However, there was a significant difference between the orders (F(1,24) = 5.58, $p < 0.01$). This difference was due to Order 3 having higher success rates during the normal load condition, while the other two orders had the highest success rates during the overload condition. There was no significant interaction between the workload conditions and orderings.

Table 4.25: Communication Task Performance Descriptive Statistics for Success Rate (%). Note: Higher is Better.

Order	Normal Load	Overload
1	70.71 (25.96)	**73.57 (13.69)**
2	77.59 (25.50)	**84.12 (13.00)**
3	**96.82 (5.18)**	85.71 (8.58)
Overall	**81.07 (23.67)**	**81.11 (12.89)**

Comparison to the Supervisory Evaluation

An overview of the performance metric trends for an increase in workload for each evaluation (Real-Time and Supervisory) is provided in Table 4.26. The communication task reaction-time performance metric is not included in the table, as the original NASA MATB-II did not collect this metric. Four out of the five performance metrics trended in a similar manner for each evaluation. The system monitoring task's reaction time increased as workload increased for the real-time evaluation, but had no trend for the supervisory evaluation's second day.

Table 4.26: The Performance Metric Trends for the Real-Time Evaluation and the Supervisory Evaluation's Second Day.

Metric	Expected	Real-Time	Supervisory Day 2
TRCK: RMSE	Decreases	Decreases	Decreases
RES: Time in Range	Decreases	Decreases	Decreases
SYS: Reaction Time	Increases	Increases	No Trend
SYS: Success Rate	Decreases	Decreases	Decreases
COMM: Success Rate	Decreases	No Trend	No Trend

4.3.3 Subjective Metric Analysis

The highest value for each table is bolded, as the value represents the highest perceived workload level.

In-Situ Workload Ratings

The in-situ workload ratings subjectively assessed workload across six dimensions: auditory, visual, speech, motor, tactile, and cognitive. Each rating ranges from 1 (little to no demand) to 5 (extreme demand).

The auditory workload rating's descriptive statistics are provided in Table 4.27. The highest auditory ratings occurred during the overload conditions for each workload condition ordering. A two-way MANOVA determined that there were significant differences for workload condition (F(2,23) = 112.11, $p < 0.01$) and for the orderings (F(2,23) = 11.45, $p < 0.01$). There was a significant interaction between the conditions and orderings (F(4,23) = 3.27, $p = 0.01$).

Table 4.27: Descriptive Statistics for In-Situ Auditory Workload Ratings.

Order	**Underload**	**Normal Load**	**Overload**
1	1.23 (0.68)	2.95 (1.39)	**4.0 (1.03)**
2	1.36 (0.73)	2.61 (0.79)	**3.27 (1.08)**
3	1.2 (0.41)	1.9 (0.64)	**2.9 (0.76)**
Overall	1.26 (0.63)	2.51 (1.03)	**3.32 (1.03)**

The means and standard deviations by group and condition for the visual workload ratings are presented in Table 4.28. The participant rated their visual workload the highest during the overload condition. There was a significant effect for workload condition (F(2,23) = 39.85, $p < 0.01$). There was no significant effect for the workload condition ordering or for the interaction between workload condition and ordering.

Table 4.28: Descriptive Statistics for In-Situ Visual Workload Ratings.

Order	**Underload**	**Normal Load**	**Overload**
1	2.23 (1.3)	3.50 (1.05)	**4.00 (1.03)**
2	2.59 (1.22)	3.30 (0.92)	**3.59 (0.73)**
3	1.80 (0.83)	2.85 (1.04)	**3.57 (1.04)**
Overall	2.22 (1.19)	3.23 (1.01)	**3.69 (0.96)**

The speech workload ratings by group and condition are presented in Table 4.29. Similar to the previous ratings, the largest speech workload ratings occurred during the overload condition. A two-way MANOVA determined that there was a significant effect of workload condition (F(2,23) = 74.78, $p < 0.01$) and no significant effect for the ordering or the interaction between workload condition and ordering.

Table 4.29: Descriptive Statistics for In-Situ speech workload ratings.

Order	Underload	Normal Load	Overload
1	1.13 (0.51)	2.45 (1.23)	**3.45 (1.39)**
2	1.36 (0.79)	2.24 (0.87)	**3.05 (1.05)**
3	1.20 (0.41)	2.10 (0.79)	**2.80 (0.92)**
Overall	1.22 (0.59)	2.26 (0.96)	**3.06 (1.12)**

The motor in-situ rating is a subcomponent of physical workload, where the descriptive statistics are provided in Table 4.30. The participants rated their motor demands the highest during the overload condition. There was a significant difference between workload conditions (F(2,23) = 49.43, p <0.01) and between condition orders (F(2,23) = 5.99, p <0.01). There was no significant interaction between the workload conditions and orderings.

Table 4.30: Descriptive Statistics for In-Situ motor workload ratings.

Order	Underload	Normal Load	Overload
1	2.13 (1.22)	3.00 (1.12)	**3.75 (0.97)**
2	1.82 (0.96)	3.06 (0.93)	**3.27 (0.98)**
3	1.45 (0.51)	2.65 (0.88)	**3.10 (0.84)**
Overall	1.85 (1.02)	2.93 (0.98)	**3.33 (0.95)**

Tactile workload is another subcomponent of physical workload, while the means and standard deviations for the tactile ratings are presented in Table 4.31. The ratings were the largest during the overload condition for each workload condition ordering. A two-way MANOVA found a significant difference for workload condition (F(2,23) = 39.92, p < 0.01). There was a significant main effect on ordering (F(2,23) = 19.81, p < 0.01), but no significant interaction between the orderings and workload conditions.

Table 4.31: Descriptive Statistics for In-Situ Tactile workload ratings.

Group	Underload	Normal Load	Overload
1	1.9 (1.16)	2.80 (1.15)	**3.45 (0.94)**
2	2.05 (1.13)	3.15 (0.76)	**3.41 (1.22)**
3	1.30 (0.47)	1.85 (0.59)	**2.63 (0.89)**
Overall	1.78 (1.04)	2.70 (1.0)	**3.10 (1.08)**

The last in-situ rating is cognitive workload, where the descriptive statics by workload condition and order are provided in Table 4.32. A significant main effect for workload condition (F(2,23) = 108.16, $p < 0.01$) and condition ordering (F(2,23) = 3.72, $p = 0.02$) was found, along with a significant interaction between the workload condition and orderings (F(2,23) = 2.69, $p = 0.03$).

Table 4.32: Descriptive Statistics for In-Situ Cognitive workload ratings.

Group	Underload	Normal Load	Overload
1	1.67 (0.99)	3.45 (0.83)	**4.25 (0.72)**
2	2.18 (0.85)	3.15 (0.71)	**3.73 (0.88)**
3	1.65 (0.67)	2.75 (1.02)	**3.70 (0.92)**
Overall	1.82 (0.89)	3.12 (0.87)	**3.86 (0.88)**

The overall in-situ workload rating is the aggregate of the individual ratings and the descriptive statistics are provided in Table 4.33. It was expected that the highest overall ratings occurred during the overload condition for each order. A two-way MANOVA determined that the workload conditions (F(2,23) = 112.91, $p < 0.01$) and orderings (F(2,23) = 10.95, $p < 0.01$) differed significantly, but there was no significant interaction between the two.

Table 4.33: Descriptive Statistics for the Overall In-Situ workload ratings.

Order	Underload	Normal Load	Overload
1	10.30 (5.17)	18.15 (5.93)	**22.90 (4.67)**
2	11.36 (4.54)	17.52 (3.51)	**20.32 (4.00)**
3	8.60 (2.54)	14.10 (3.68)	**18.70 (3.53)**
Overall	10.15 (4.46)	16.75 (4.59)	**20.36 (4.32)**

Comparison to the Supervisory Evaluation

A comparison for the In-Situ workload ratings trends between the evaluations (Real-Time and Supervisory) is presented in Table 4.34. The NASA-TLX ratings are not included, as the ratings cannot be parsed by workload condition. Each in-situ workload rating trended in a similar manner for each evaluation, which demonstrates that the participants' perceived workload was similar between the evaluations.

NASA-TLX Ratings

Table 4.34: The In-Situ Workload Rating Trends for the Real-Time Evaluation and the Supervisory Evaluation's Second Day.

Rating	Real-Time Evaluation	Supervisory Evaluation's Day 2
Auditory	Increases	Increases
Visual	Increases	Increases
Speech	Increases	Increases
Motor	Increases	Increases
Tactile	Increases	Increases
Cognitive	Increases	Increases
Overall	Increases	Increases

The presentation of each NASA-TLX score represents the unweighted score, while the overall NASA-TLX results represent the weighted aggregate. The NASA-TLX was administered after the single trial; thus, the results cannot be broken down by workload condition. The descriptive statistics for each NASA-TLX scale and the associated ANOVA result are provided in Table 4.35. An ANOVA determined that none of the NASA-TLX scales differed significantly between the three orders.

Table 4.35: NASA-TLX Workload Ratings by Workload Condition Ordering.

Order	Effort	Frustration	Mental	Performance	Physical	Temporal	Overall
1	60.00 (25.93)	49.00 (32.73)	71.0 (22.95)	44.50 (23.51)	43.50 (25.72)	63.00 (26.58)	61.07 (21.62)
2	74.55 (14.91)	37.27 (26.11)	68.64 (22.03)	35.91 (21.89)	45.00 (20.00)	71.82 (13.28)	64.15 (11.02)
3	75.50 (14.62)	38.00 (22.51)	66.0 (16.8)	37.00 (22.14)	49.50 (25.22)	61.00 (22.83)	60.00 (12.25)
ANOVA F(2, 28)	2.09, n.s.	0.58, n.s.	0.14, n.s.	0.44, n.s.	0.17, n.s.	0.76, n.s.	0.20, n.s.

4.3.4 Real-Time Analysis

Two trained algorithms were used to analyze the developed workload assessment algorithm's real-time capabilities. The inputs for each algorithm were the workload metrics listed in Table 4.4. The speech and visual workload components results are not presented in the real-time analysis, due to IMPRINT Pro workload models being used to estimate the components. The **SUP** algorithm was trained on all of the data from the supervisory-based evaluation detailed in Chapter 3.4 and estimated workload in real-time for each participant. The **SUP** algorithm was not trained on any data from the real-time evaluation; thus, a second algorithm (**RT**) was trained on all of the supervisory-based evaluation (Chapter 3.4) data and data from the first six participants from the real-time evaluation.

Estimating workload in real-time provides an adaptive teaming system with the necessary insight to adapt interactions and system autonomy in order to maintain or improve the performance of the human-machine team. These workload estimates must be accurate; thus, hypothesis $\mathbf{H_2^{RT}}$ predicted that the **SUP** and **RT** algorithms' estimates will be within a standard deviation of the IMPRINT Pro model values for each workload component. Accurate workload estimates improve the likelihood that the algorithm's workload classifications will also be accurate. Hypothesis $\mathbf{H_3^{RT}}$ predicted that each trained algorithm will accurately classify each workload state at least 80% of the time. The algorithm's workload estimates will correlate positively and significantly with the corresponding IMPRINT Pro models, as predicted by hypothesis $\mathbf{H_4^{RT}}$.

The descriptive statistics are used to determine how close the algorithms' estimates are to the IMPRINT Pro model values. These descriptive statistics are provided in Table 4.36 by algorithm, workload condition, and workload component. The **SUP** algorithm's estimates were within a standard deviation of the IMPRINT Pro model values for each workload component in the normal load condition, but tended to underestimate workload in the overload condition and overestimate workload in the normal load condition. The **RT** algorithm performed similar to, or better than the **SUP** algorithm in all cases, but also tended to underestimate workload in the overload condition. The **RT** workload estimates for the third workload condition ordering are shown in Figure 4.4.

Table 4.36: IMPRINT Pro Modeled and Algorithm Estimated Workload by Workload Condition.

Workload	Algorithm	Underload	Normal Load	Overload
Auditory	Model	0.18 (0.65)	2.06 (1.38)	3.58 (0.88)
	SUP	0.80 (1.17)	1.83 (1.25)	2.20 (1.21)
	RT	0.71 (1.33)	2.23 (1.28)	3.16 (0.97)
Cognitive	Model	2.14 (1.37)	9.46 (1.68)	18.44 (1.86)
	SUP	6.71 (4.65)	10.33 (5.01)	13.31 (5.29)
	RT	3.08 (3.11)	9.86 (3.62)	15.46 (3.78)
Physical	Model	1.31 (0.74)	5.57 (2.78)	10.27 (2.20)
	SUP	3.97 (2.68)	5.84 (3.00)	7.38 (2.97)
	RT	2.45 (2.72)	5.99 (3.29)	7.66 (3.03)
Overall	Model	7.81 (3.33)	29.51 (6.17)	55.79 (5.88)
	SUP	15.52 (7.59)	30.27 (9.14)	46.47 (9.52)
	RT	10.41 (6.3)	30.29 (5.87)	49.70 (6.57)

The workload assessment algorithm classified workload by using thresholds to categorize its estimates into a workload state. The classification accuracies are presented in Table 4.37. The **RT** algorithm classifies workload better than the **SUP** algorithm in almost all cases, the exception being classifying physical workload for the normal load condition. Both algorithms had difficulty classifying physical workload during the real-time evaluation, which may be attributed to inaccuracies in the IMPRINT Pro workload models.

Table 4.37: Classification Accuracy (%) for the Real-Time Evaluation

Workload	Algorithm	Underload	Normal Load	Overload
Auditory	SUP	77	47	48
	RT	**80**	**82**	**82**
Cognitive	SUP	42	61	56
	RT	**88**	**82**	**77**
Physical	SUP	40	**53**	52
	RT	**75**	51	**60**
Overall	SUP	73	76	71
	RT	**83**	**93**	**90**

The ability to track workload trends accurately may allow for more intelligent system

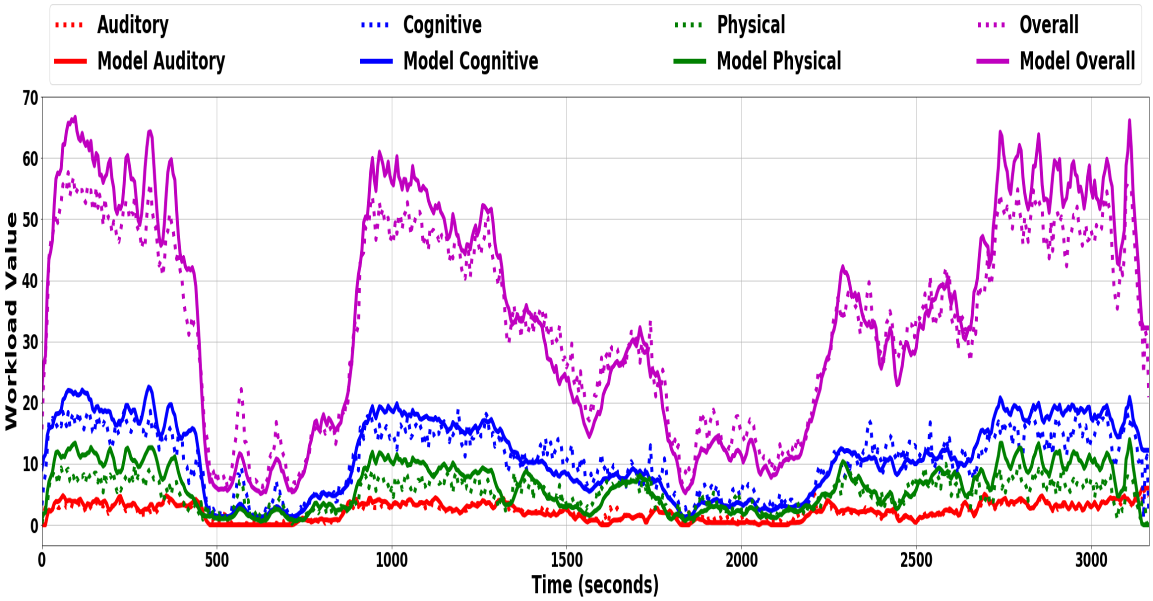


Figure 4.4: The **RT** Workload Estimates for the third condition ordering: OL-UL-OL-NL-UL-NL-OL.

adaptations. The algorithm's ability to track these trends was analyzed using Pearson's correlations between the algorithms' estimates and the corresponding IMPRINT Pro workload models, which are provided in Table 4.38. Almost all correlations are positive and significant, the exception being the **SUP** algorithm's cognitive workload estimates for the normal load condition. The **RT** algorithm's estimates were more positively correlated with the corresponding IMPRINT Pro workload models than the **SUP** algorithm's estimates. The lowest correlations were produced for cognitive workload in the normal load condition.

4.3.5 Post-Hoc Analysis

The **SUP** and **RT** algorithms had limited information about the real-time evaluation's physical aspects, but were able to extract features from the workload metrics in order to produce real-time workload estimates. The incorporation of more task environment specific training data may allow for higher algorithmic performance; thus, the algorithm was trained using data from twenty-seven participants and tested on the remaining participants (n=4) in order to provide **POST-HOC** performance. Additionally, the workload assessment

Table 4.38: Real-Time Evaluation Correlation Coefficients for Within and Across Workload Conditions.

Workload	**Algorithm**	**Within**			**Across**
		Underload	Normal Load	Overload	
Auditory	SUP	0.31*	0.41*	0.28*	0.53*
	RT	**0.52***	**0.80***	**0.69***	**0.82***
Cognitive	SUP	0.12*	0.02	0.14*	0.48*
	RT	**0.78***	**0.19***	**0.29***	**0.84***
Physical	SUP	0.16*	0.27*	0.15*	0.47*
	RT	**0.69***	**0.65***	**0.42***	**0.71***
Overall	SUP	0.27*	0.38*	0.41*	0.85*
	RT	**0.84***	**0.63***	**0.67***	**0.96***

algorithm's speech workload estimation accuracy was analyzed. Similar to the previous hypotheses, four hypotheses focused on algorithmic performance. Hypothesis $\mathbf{H_5^{RT}}$ predicted that the **POST-HOC** workload estimates will be within a standard deviation of the corresponding IMPRINT Pro workload models, while hypothesis $\mathbf{H_6^{RT}}$ predicted that the algorithm will correctly classify overall workload and each workload component at least 80% of the time. The algorithm's estimates will positively and significantly correlate with the corresponding IMPRINT Pro workload models, as predicted by hypothesis $\mathbf{H_7^{RT}}$. The last hypothesis ($\mathbf{H_8^{RT}}$) predicted that the **POST-HOC** algorithm will perform better than the **RT** algorithm.

The **POST-HOC** algorithm estimated workload for four real-time evaluation participants. The algorithm's average estimates and IMPRINT Pro model values are provided in Table 4.39. The algorithm's estimates were within a standard deviation of the corresponding IMPRINT Pro workload values for each workload component and condition, except for speech workload. The algorithm tended to underestimate the speech workload models, which is attributed to the model inaccuracies; for example, the participants may not have spoken when the IMPRINT Pro model indicated that they were speaking and vice versa.

The **POST-HOC** algorithm achieved similar or higher classification accuracies than the **RT** algorithm, as indicated in Table 4.40. The classification accuracies were approximately

Table 4.39: IMPRINT Pro Modeled and POST-HOC Algorithm Estimated Workload by Workload Condition

Workload	Algorithm	UL	NL	OL
Auditory	Model	0.17 (0.62)	2.06 (1.39)	3.53 (0.91)
	POST-HOC	0.65 (1.13)	2.25 (1.24)	3.08 (0.94)
Cognitive	Model	2.13 (1.36)	9.3 (1.56)	18.20 (2.13)
	POST-HOC	3.21 (3.07)	10.15 (3.04)	16.57 (3.25)
Physical	Model	1.31 (0.73)	5.46 (2.74)	10.20 (2.22)
	POST-HOC	2.48 (2.86)	6.91 (3.13)	9.39 (2.31)
Speech	Model	0.05 (0.21)	0.76 (0.65)	2.01 (0.88)
	POST-HOC	0.08 (0.5)	0.17 (0.73)	0.36 (1.03)
Overall	Model	7.78 (3.27)	29.01 (5.93)	55.37 (6.47)
	POST-HOC	10.51 (6.64)	31.50 (6.04)	52.48 (6.64)

80% or higher for most workload components and conditions. The algorithm had difficulty classifying physical workload for the normal load condition.

Table 4.40: Post-Hoc Classification Accuracy (%).

Workload	Underload	Normal Load	Overload
Auditory	79	88	81
Cognitive	92	85	85
Physical	75	57	89
Overall	81	92	93

Tracking workload trends within and across workload conditions is important for adaptive teaming systems. The Pearson's correlation coefficients between the **POST-HOC** algorithm estimates and corresponding IMPRINT Pro workload models are provided in Table 4.41. The algorithm's estimates correlated positively and significantly with the IMPRINT Pro workload models for each workload condition and component. The lowest correlations resulted when estimating speech workload. Each correlation is larger than the corresponding correlations in Table 4.38, which demonstrated that the **POST-HOC** algorithm's estimates represent workload trends more accurately, than the other trained algorithm (**RT** and **SUP**) variations' estimates.

Table 4.41: POST-HOC Correlation Coefficients for Within and Across Workload Conditions.

Workload	**Within**			**Across**
	Underload	Normal Load	Overload	
Auditory	0.65*	0.82*	0.78*	0.86*
Cognitive	0.80*	0.24*	0.45*	0.90*
Physical	0.79*	0.76*	0.57*	0.83*
Speech	0.12*	0.10*	0.09*	0.17*
Overall	0.86*	0.71*	0.79*	0.97*

4.4 Discussion

The real-time evaluation focused on estimating workload in real-time for a non-stationary supervisory-based environment. The objective and subjective workload metrics analysis demonstrated that there were differences between the workload conditions that each participant experienced. Additionally, the majority of the metrics responded to an increase in workload in a similar manner to the corresponding metrics for the supervisory evaluation's second day (Chapters 3.5.1 - 3.5.3), which supports hypothesis $\mathbf{H_1^{RT}}$. Supporting the hypothesis demonstrates that the metrics behaved as expected for the real-time evaluation.

The real-time workload assessment is the foundation of the adaptive human-robot teaming system, as the estimates provide information regarding the human's overall workload state and each workload component state. Hypothesis $\mathbf{H_2^{RT}}$ stated that the **SUP** and **RT** algorithm estimates will be with a standard deviation of the IMPRINT Pro workload models. This hypothesis was upheld when estimating workload during the normal load condition for both algorithms, but was only partially upheld for the underload and overload conditions. The underload and overload outcomes are attributed to the uncertainty in the IMPRINT Pro workload models, as the participants were not always completing the task that the workload model indicated (e.g., completing the communications task when the model indicated that the tracking task was being completed). These outcomes may also be attributed to the multi-tasking nature of the task environment.

Although there were inaccuracies in the IMPRINT Pro models, the models still represented the human's overall workload state accurately. Hypothesis $\mathbf{H_3^{RT}}$ predicted that each workload condition will be classified correctly at least 80% of the time, which was upheld for the **RT** algorithm's auditory and overall workload classifications. The **RT** algorithm's cognitive classifications partially uphold the hypothesis, while the physical workload classifications did not uphold the hypothesis. The low physical workload classifications were interesting, as the **RT** algorithm's physical workload estimates were similar to the IMPRINT Pro model's statistics. The hypothesis is not supported for the **SUP** algorithm.

Tracking the workload trend may allow an adaptive teaming system to trigger an adaptation in order to preclude a performance decrement. Therefore, the developed workload assessment algorithm's estimates need to correlate positively and significantly with the corresponding IMPRINT Pro workload models in order to capture the workload trends accurately, as stated by hypothesis $\mathbf{H_4^{RT}}$. This hypothesis is fully supported for the **RT** trained algorithm and partially supported for the **SUP** algorithm. Fully supporting this hypothesis demonstrates that the algorithm captures workload trends correctly; thus, enabling more intelligent adaptations that are tailored to the particular workload components. Additionally, the correlation coefficients were larger when tracking workload across workload conditions than within conditions, illustrating that the algorithm better captures large workload variations better than small workload variations, as expected.

Overall, the **RT** algorithm's estimates were more accurate than the **SUP** algorithms, which demonstrates that having prior training data on the task environment improves performance, which was expected. Both trained algorithm variants were used in real-time with limited information concerning the task environment; however, a more accurate trained algorithm may be produced with more training data. Using the workload metric data collected in real-time, another trained algorithm variant (**POST-HOC**) was analyzed. Hypotheses $\mathbf{H_5^{RT}}$ through $\mathbf{H_7^{RT}}$ were the same as $\mathbf{H_2^{RT}}$ through $\mathbf{H_4^{RT}}$ respectively, but tailored to the **POST-HOC** trained algorithm. Each hypothesis was fully supported for almost all workload components and conditions, besides speech workload. Additionally, the **POST-HOC** algorithm outperformed the **RT** algorithm, which supports hypothesis $\mathbf{H_8^{RT}}$. Fully supporting these hypotheses demonstrates that the real-time workload assessment algorithm's accuracy is limited to the trained algorithm variant, rather than collecting the required workload metric data.

There are hardware limitations to collecting the workload metric data. For example, the BioPac harness used to collect the physiological metrics has a one second update rate for the real-time heart-rate calculations, which restricts the workload assessment algorithm's workload estimation to at most every second. Limitations also exist in the feature extraction process, since extracting speech-based features tended to be relatively computational expensive, due to using the Fast Fourier Transform. This feature extraction process is sensitive to the window size, as processing larger window sizes requires more computational time. Future work will investigate the window size's impact on the speech-based feature extraction process run-time in order to determine the minimal update rate for the developed workload assessment algorithm. The **POST-HOC** trained algorithm estimated speech workload, while the other trained algorithm variants did not. The speech workload estimates were much lower than the IMPRINT Pro workload model values, which was not anticipated. These lower estimates are again attributed to the IMPRINT Pro workload models having only three possible values (i.e., 0, 2, or 4), as discussed in Chapter 3.4.

4.5 Summary

The developed workload assessment algorithm's real-time capabilities were evaluated. The task environment encompassed an adapted version of the NASA MATB-II, in which each task was physically separated from each other. The algorithm was shown to accurately estimate workload during the evaluation; however, more accurate workload estimates occurred when the algorithm was trained on the majority of the real-time evaluation's data set. Additionally, the algorithm was shown to accurately estimate speech workload. The real-time workload assessment algorithm is a necessary component of an adaptive human-robot teaming system, as the system relies on information about the human's overall workload state and the state of each workload component in order to adapt either interactions between the human and machine or the autonomy levels of the system's tasks.

Chapter 5

Adaptive Human-Robot Teaming System

An adaptive human-robot teaming system and a performance prediction model was developed, where a pilot study (Chapter 5.2) demonstrated the system's ability to adapt interactions and autonomy levels. These adaptations were based on the real-time workload assessment algorithm (Chapter 4.1) and a performance prediction model (Chapter 5.1), which was validated. The adaptive system was shown to improve task performance over having no adaptations, which was attributed to the interaction adaptations.

5.1 Performance Prediction

Estimating workload accurately is essential to an adaptive teaming system, but only provides current and previous performance information. The ability to predict accurately if performance is likely to decrease in the future may allow an adaptive teaming system to trigger an adaptation in order to prevent the performance decrements from occurring. A performance prediction model was developed in order to predict performance accurately for future timesteps.

The developed prediction model relied on a long short-term memory neural network architecture [55]. Long short-term memory networks use the previous time-step information in order to predict future time-steps of a sequential data series. The developed model uses the last three workload estimates (i.e., overall workload and each workload component) as inputs. The performance model consisted of three long short-term memory layers each with 256 neurons. Each neuron in the long short-term memory layers had an 80% chance to dropout during training, which means the neuron may be excluded from training activation and weight updates [55]. There was a 256 neuron fully connected layer with a rectified linear unit activation function after the three long short-term memory layers. The model's output regression layer predicted overall task performance for thirty seconds in the future. The ADAM optimizer [69] with a mean-squared error loss function was used to train the performance prediction model.

5.1.1 NASA MATB-II Overall Task Performance

The performance prediction model predicts overall task performance. Each NASA MATB-II task has its own performance measure, but there is no current method to combine these task performance measures into an overall performance measure. Each NASA MATB-II task performance measure was mapped to a value from 0 to 1 in order to permit combining the measures, where 1 represents optimal performance. The tracking task performance measure (i.e., RMSE between the center of the crosshairs and the object) was normalized based on participant data. Performance for the system monitoring task and communications task were measured using two metrics: reaction time and success rate. Reaction time represents the time a participant took to correct an out-of-range light or gauge, while success rate represents the number of out-of-range instances corrected divided by the total number of instances. Reaction time was normalized, while success rate was already within range (0 - 1). A value of 1 was given if the resource management task's fuel levels were within 2,000 and 3,000 units, while the tank levels were normalized outside of that range. The overall performance measure was the uniform average of all active tasks' performance measures, which assumes that the tasks trade offs will be equivalent in terms of performance. For example, if the resource management and system monitoring tasks were the only active tasks, then the overall task performance was the average of those tasks' performance measures.

Relying on normalizing the performance metrics and using an uniform average calculation may not be the optimal solution to generating an overall performance score. Normalizing performance data does not penalize time dependent performance measures. For example, the fuel tank levels can only rise so quickly; thus, fuel levels much smaller than 2,000 units need to be penalized more than fuel levels close to 2,000 units. However, developing appropriate time penalizations is not trivial and tangential to the developed performance model's ability to predict performance. Additionally, using an uniform average to calculate overall task performance does not account for task priority levels. The participants were not given any task priortiziations; thus, the use of an uniform average.

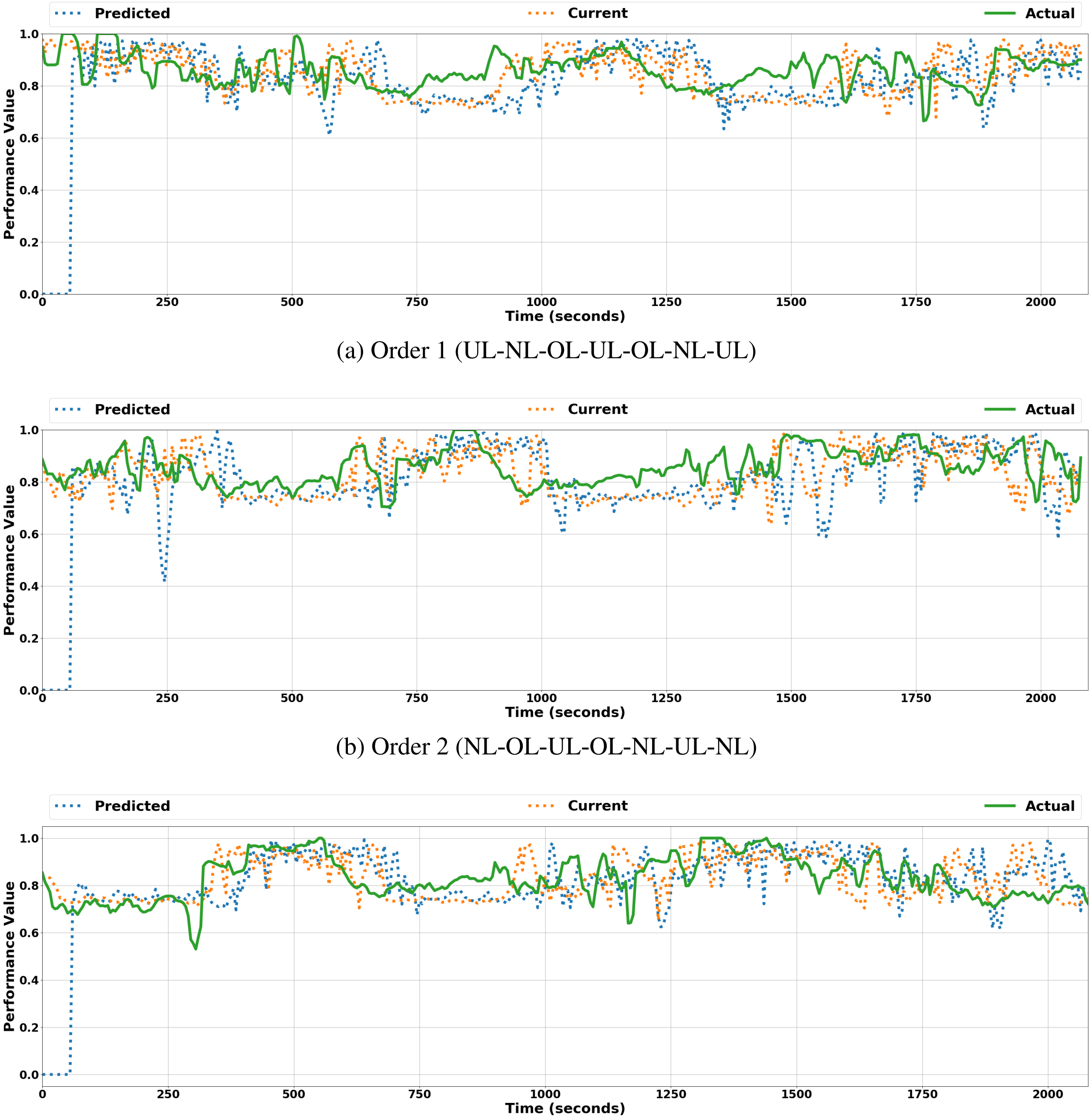
5.1.2 Performance Prediction Results

The performance prediction model was trained using data from the Supervisory Evaluation's (i.e., data from the first day (Chapter 3.4) and tested on the supervisory evaluation's second day data (Chapter 3.5). There were two trained prediction model variants: **Current** and **Predicted**. The **Current** model variant was used to predict a participant's current task performance, while the **Predicted** variant predicted task performance for 1-minute into the future. It was expected that each model variant's predicted performance values will be within a standard deviation of the participants' actual performance, as stated by hypothesis $\mathbf{H_1^P}$. These descriptive statistics are provided in Table 5.1. Each models' performance values for the underload and normal load conditions were within a standard deviation of the participants' actual performance. The predicted performance values for the overload condition were lower than the actual performance values.

Table 5.1: Descriptive Statistics for the Current and Predicted Task Performance and Actual Participant Performance.

Performance Model	**Underload**	**Normal Load**	**Overload**
Actual	0.89 (0.15)	0.84 (0.16)	0.78 (0.13)
Current	0.89 (0.12)	0.84 (0.11)	0.74 (0.03)
Predicted	0.91 (0.13)	0.83 (0.12)	0.74 (0.05)

Using descriptive statistics to compare actual vs. predicted performance is useful, but does not provide any time information (i.e., does the predicted performance trend the same way as the actual performance). Each performance model variants' values and the actual performance values are plotted in Figure 5.1 by workload condition ordering. The **Predicted** performance model's values were set to zero for the first 60 seconds in order to match the predicted values and the actual values. There was approximately a 10 second delay between a large increase in the actual performance and the **Current** model's performance values. A larger delay occurred between the actual performance and **Predicted** performance values.



(a) Order 1 (UL-NL-OL-UL-OL-NL-UL)

(b) Order 2 (NL-OL-UL-OL-NL-UL-NL)

(c) Order 3 (OL-UL-OL-NL-UL-NL-OL)

Figure 5.1: Performance Prediction Results by Workload Condition Ordering

5.1.3 Performance Prediction Discussion

Predicting future task performance accurately is a difficult and complex problem, but the developed performance prediction model is a necessary step to being able to predict task performance. It was expected that the developed model's overall task performance values will be within a standard deviation of the actual performance values, as predicted by hypothesis $\mathbf{H_1^P}$. The hypothesis was supported for the underload and normal load conditions, but not for the overload condition. Not fully supporting the hypothesis is attributed to the training data set, as the data set had clearly separated workload conditions; however, the testing data set incorporated workload transitions. The performance prediction model needs to be trained on these workload transitions in order to achieve better predictive power. However, it is expected that there is a limit to the model's predictive capabilities. For example, predicting task performance for an hour in the future will likely be inaccurate.

The performance prediction model relied on the last three workload estimates to predict task performance for 1-minute into the future. More accurate predicted performance may be achieved by using additional workload estimates or predicting task performance nearer to the current timestep (i.e., fifteen seconds into the future). There may be a trade-off with using additional workload estimates, as incorporating outdated information may actually decrease performance. Future work will vary the number of past workload estimates and the time in the future that task performance is predicted in order to further analyze the developed performance prediction model's capabilities.

Predicting overall task performance for the NASA MATB-II was a necessary step for developing an adaptive system architecture, but the model's predictive capabilities in other task environments is unknown. There are likely limitations to the types of task environments the prediction model can be deployed in. There needs to be an overall task performance measure for the task environment. Developing such a measure may be non-trivial for some task environments or it may be better to predict performance for certain tasks. For example, predicting a human's reaction time to a certain event may be useful information for an adaptive teaming system, as the system may use this predicted reaction time in a task scheduling paradigm. The developed prediction model may be extended to predict individual task performance by incorporating additional neurons in the model's output layer, where each neuron represents an individual task's performance.

5.2 Adaptive Teaming System

The adaptive teaming system combines the developed workload assessment algorithm (Chapter 3.1) and performance prediction model in order to adapt system interactions and automate tasks intelligently. An overview of the architecture, provided in Figure 5.2, is spit into three stages: *Perceive*, *Select*, and *Act*.

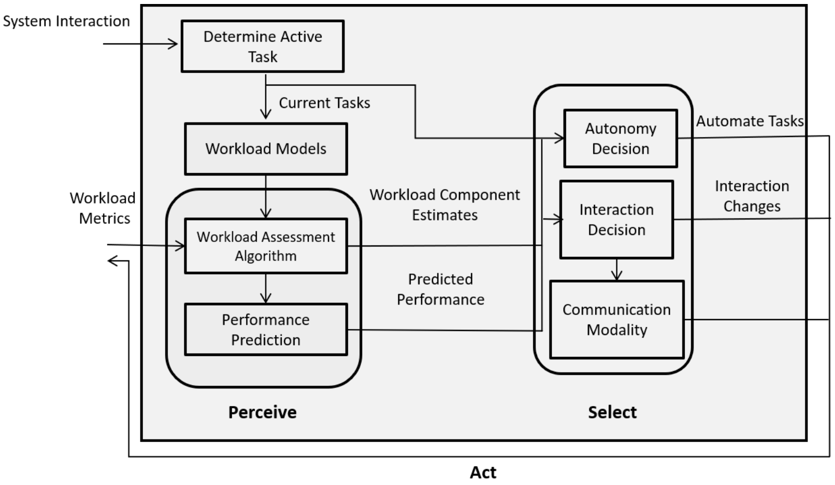


Figure 5.2: Adaptive Teaming System Architecture

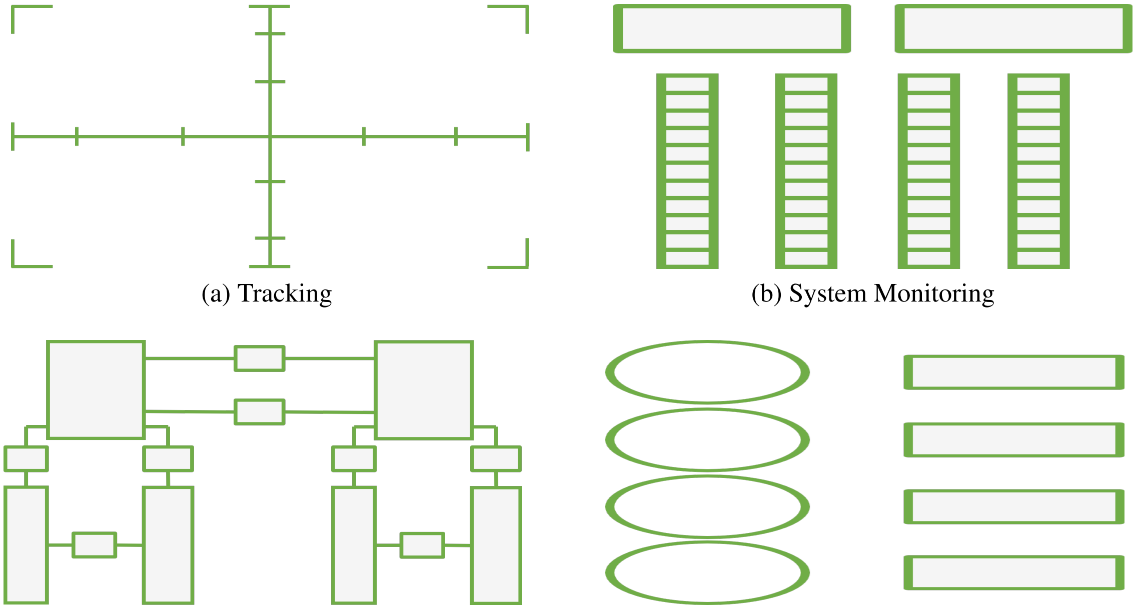
The workload metrics are used as inputs into the *Perceive* stage, such that the developed workload assessment algorithm can estimate overall workload and each workload component. The algorithm's workload estimates are used by the performance prediction model and the *Select* architecture stage. Contextual features calculated from the IMPRINT Pro workload models are required for accurate workload estimates, which requires knowing the participant's current task; however, knowing this task in dynamic domains may not be trivial. This dissertation used a supervisory-based interface (NASA MATB-II); thus, the participant's current active task was always known. The interface tracked the participant's last input (e.g., moving the joystik or a keystroke), which was determined to correspond to the participant's current task. Participants often completed more than one task simultaneously; thus, the closest task to the participant's last input was included in the current task set. For example, if the participant moved the joystik, then the current task set consisted of the tracking and system monitoring tasks. Likewise, if a participant tuned a radio, then the task set consisted of the resource management and communication tasks. The *Select* stage identified if a task needed to be automated or how an interaction occurred using the knowledge of the human's current task set, the workload estimates, and predicted performance. If the human's predicted performance fell below a threshold value (0.70), or when the last three overall workload estimates were in the overload state, then all inactive tasks (as determined by the interface) were transitioned to automation mode. Three was chosen for the number of workload estimates to ensure that the system did not thrash cyclically, where automation is turned off and on each workload estimate. If the last three overall estimates were considered in the underload state, or the human's predicted performance was above a threshold level (0.85), then all tasks transitioned out of automation mode. The threshold levels were chosen based on the overall performance values for the supervisory evaluation (Chapter 3.2.1). The adaptive teaming system architecture determined how a system interaction occurred, once an interaction was expected to occur. The NASA MATB-II interactions occurred when the tracking task switched modes (e.g., manual to automation), when a light or gauge went out of range in the system monitoring task, or when the resource management task's fuel levels went out of range. The adaptive system selected a communication modality (i.e., visual or auditory) based on potential conflicts in the workload channels. A visual modality was used if the participant's visual workload channel was not overloaded, meaning that the participant had sufficient resources to parse the interaction's visual information. An auditory modality was used if the human's speech and auditory workload channels were not loaded, as an auditory stimulus may distract the participant if they were speaking, or if there was substantial environmental noise. The interaction's auditory stimulus was postponed for 5 seconds, if the participant's auditory or speech channels were loaded. If after 5 seconds the participant's workload channels were still loaded, then the interaction used no auditory stimuli.

Interactions pertain to how the system conveyed information to the participant and how the participant interacted with the system (e.g., clicking a mouse). The *Select* stage changed the communication task's interaction modality, depending on the participant's available resources. Participants were able to speak into the microphone in order to change the communications task's radios, instead of using a physical modality (i.e., using the mouse). Participants were told that the system used speech recognition to determine what radio and frequency they were saying, but the system detected that the participant was speaking and assumed that they said the correct radio/frequency.

Icons, provided in Figure 5.3, were used to communicate each task state (e.g., the task was being automated) to the participant. An icon was green if the task was in automation mode, red if the task was out of range (e.g., a light went out of range), or grey if the participant was to manually determine if a task was out of range. The icons appeared on the left side of each computer screen and were greyed out if the participant's visual channel was determined to be overloaded and the corresponding task was not being automated in order to reduce visual workload. There was an interaction icon (Figure 5.4) that appeared on the right side of each computer screen, where the icon represented that the participant was able to interact with the communications task via a speech modality.

5.3 Experimental Design

The adaptive teaming system architecture was implemented in the task environment described in Chapter 4.1.1.1. and was evaluated via a pilot-study using a within-subjects experimental design, with workload and adaptation condition as the independent variables. There were three adaptation conditions that participants completed: *Autonomy*, *Interac-*



(a) Tracking

(b) System Monitoring

(c) Resource Management

(d) Communications

Figure 5.3: The NASA MATB-II Task State Icons.

Figure 5.4: The COMM Interaction Icon.

tion, and *Both*. The adaptive system automated tasks in the *Autonomy* condition, but did not adapt interactions. Likewise, the system adapted interactions in the *Interaction* condition, but did not automate tasks. The *Both* condition automated tasks and adapted interaction modalities. The dependent variables consisted of the workload algorithm's estimates, performance, and subjective metrics. The participants completed a 52.5 minute trial consisting of seven consecutive 7.5 minute workload conditions: OL-UL-OL-NL-UL-NL-OL. The 52.5 minute trial's first half was completed in one adaptation condition, while the second half was completed in a different condition. Each participant completed two adaptation conditions: *Both* and either *Autonomy* or *Interaction*, where the adaptation conditions were counter-balanced across the participants.

Two participants from the prior real-time evaluation (Chapter 4.1) completed the adaptive system evaluation in order to identify potential between subject effects. The main trial's workload condition orderings were dependent on what orderings the participants previously completed. One participant completed the evaluation in the *Both-Autonomy* adaptive condition ordering and the other participant completed the *Both-Interaction* condition ordering.

5.3.1 Environment

The evaluation occurred in an empty conference room. The environmental set-up was the same as described in Chapter 4.1.1.1 and provided in Figure 4.3.

5.3.2 Workload Models

The workload models were the same as the IMPRINT Pro models developed in Chapter 4.1.3.

5.3.3 Procedure

The participants completed a consent form and a demographic questionnaire upon arrival, after which participants were fitted with a BioPac Bioharness, a Schure Microphone, and two Myo devices. The Myos were fitted on the participant's forearm and collected electromyography and acceleration data. A 15-minute training session occurred prior to commencing the 52.5-minute trial. Participants completed the NASA-TLX and a post-session questionnaire upon trial completion. In-situ workload ratings were given at 7 minutes into the trial and every 7.5 minutes after the initial collection.

5.3.4 Participants

The ten participants (5 female and 5 male) including the participants whom completed the real-time evaluation had a mean age of 24.9 (St. Dev. = 1.72), seven of which were graduate students in the Robotics program at Oregon State University. Five participants held an undergraduate degree and five participants held a Master's degree. Participants rated their video game skill level on a Likert Scale (1-little to 9-expert) with an average of 4.90 (St. Dev. = 2.42). Seven participants played video games at most 3 hours a week. Seven participants drank no caffeine the day of the experiment, while three participants drank 16 oz or less.

The participants slept on average 6.75 (St. Dev. = 1.51) hours the night before the experiment and on average 7.80 (St. Dev. = 1.03) hours two nights prior. The participants rated their stress and fatigue levels on a Likert scale (1-little to 9-extreme) with an average stress level of 2.7 (St. Dev. = 1.06) and average fatigue level of 3.3 (St. Dev. = 1.80).

No participant was determined to be an outlier based on their performance and subjective data. Thus, the participant demographics did not impact or marginally impacted the results.

5.3.5 Metrics

The metrics collected for this evaluation are the same as in the real-time evaluation (Chapter 4.1.2).

5.4 Hypotheses

The remainder of this chapter analyzes the workload assessment algorithm's estimates and performance data for the real-time evaluation (Chapter 4.1) and the adaptive system study. The analysis is split into four sections: Between Evaluation, Adaptive Autonomy, Interaction Modality, and Within-Subjects. The between evaluation analysis focused on examining the workload assessment algorithm's estimates and the task performance metrics between the real-time evaluation and the adaptive system study. The adaptive autonomy and interaction modality analyses determined the effectiveness of adapting autonomy and interactions, respectively. The within-subjects analysis focused on data from two participants, who completed the real-time evaluation and the adaptive system study. Several hypothesis were formed for these analyses, where an overview is provided in Table 5.2.

Table 5.2: Chapter 5 Hypotheses

Analysis	**Hypothesis**	
Between Evaluation	$\mathbf{H_1^A}$	The workload algorithm's estimates and the subjective ratings will differ between the real-time evaluation and the adaptive system study.
	$\mathbf{H_2^A}$	Higher task performance will be achieved in the adaptive system study than the real-time evaluation.
Adaptive Autonomy	$\mathbf{H_3^A}$	Lower workload and higher task performance will be achieved in the *Both* condition, when at least one task is automated.
Interaction Modality	$\mathbf{H_4^A}$	The interaction modality adaptations will have a beneficial impact on task performance and workload.
Within-Subjects	$\mathbf{H_5^A}$	The workload assessment algorithm's estimates and the subjective workload ratings will be lower during the overload condition and higher during the underload condition when the participants used the adaptive system.
	$\mathbf{H_6^A}$	The participants will achieve higher task performance during the adaptive system study for all workload conditions.

5.5 Adaptive System Results

The adaptive teaming system was analyzed from four perspectives: between evaluation, adaptive autonomy, interaction selection, and within-subjects. The between evaluation analysis compares workload estimates, performance, and subjective workload ratings between ten participants of the real-time evaluation (*No Adaptation*) and eight participants from the adaptive system evaluation (*Adaptation*). The adaptive autonomy and interaction selection analyses investigate the impact of adapting system autonomy and interactions, respectively. The within-subjects analysis compares the same metrics as the between evaluation analysis for two participants, whom completed both the real-time and adaptive evaluations. Statistical analysis was not conducted on the data due to insufficient power to find meaningful differences across the participants.

5.5.1 Between Evaluation

Three hypotheses were formulated in order to determine if the adaptive system is effective in augmenting task performance. First, it was expected that the adaptive system will have a significant effect on human workload. Specifically, the adaptive system can serve to neutralize workload by lowering workload in the overload condition and increasing workload in the underload condition. Hypothesis $\mathbf{H_1^A}$ predicted that the workload assessment algorithm's estimates and the subjective ratings will differ between the two evaluations, with lower workload experienced in the overload condition and higher workload experienced in the underload condition for the adaptive evaluation. Neutralizing workload may affect task performance; thus, Hypothesis $\mathbf{H_2^A}$ predicted that higher performance will be achieved when using the adaptive system for each NASA MATB-II task. The task performance measures were the same as the previous evaluations described in Chapters 3.2.1.4 and 5.1.

The trained workload assessment algorithm used to estimate workload for the real-time evaluation (Chapter 4.3.5) was also used for the adaptive system evaluation. The algorithm's estimates by workload condition and evaluation type are provided in Table 5.3. The workload estimates for overall workload and each workload component, other than physical workload, were lower during the adaptive evaluation, than the real-time evaluation for the overload condition. Similar physical workload estimates results occurred in the overload condition, which is attributed to the participants primarily remaining stationary during both evaluations. The participants experienced higher overall workload during the underload condition, due to the adaptive system allocating the tracking task to underloaded participants. Auditory workload was also higher using the adaptive system, as an auditory stimulus was used to alert participants of the out of range tasks. Participants experienced similar workload levels when using the adaptive system vs. not using the system during the normal load condition.

Table 5.3: Algorithm Estimated Workload by Workload Condition and Evaluation System Type: Real-Time vs Adaptation.

Workload	**Evaluation**	**UL**	**NL**	**OL**
Auditory	Real-Time	0.71 (1.41)	2.33 (1.37)	3.09 (1.02)
	Adaptation	1.02 (2.07)	2.18 (1.81)	2.41 (1.25)
Cognitive	Real-Time	3.34 (3.25)	10.53 (3.85)	15.56 (3.66)
	Adaptation	3.31 (3.37)	9.55 (3.74)	12.77 (4.47)
Physical	Real-Time	1.50 (1.82)	3.31 (3.87)	1.70 (2.05)
	Adaptation	1.59 (2.05)	3.01 (3.69)	2.12 (2.90)
Speech	Real-Time	0.06 (0.24)	0.79 (0.68)	2.03 (0.88)
	Adaptation	0.51 (1.18)	0.65 (1.3)	0.83 (1.41)
Overall	Real-Time	9.77 (5.17)	28.65 (6.03)	44.04 (5.07)
	Adaptation	10.55 (5.69)	27.05 (6.42)	39.76 (6.65)

The NASA MATB-II tracking task performance was calculated using the RMSE in pixels between the center of the object and the center of the cross hairs (Figure 3.1). The resulting tracking error's descriptive statistics are provided in Table 5.4. The tracking task was in automation mode during the underload condition for the real-time evaluation and the *Interaction* adaptation type; thus, no corresponding results are presented. The Real-Time evaluation produced the lowest performance for the tracking task. The highest performance was achieved using the *Both* adaptation type for the underload and normal load conditions, while the *Interaction* adaptation type achieved the highest performance for the overload condition. This result for the overload condition is attributed to the tracking task being automated in the *Both* and *Autonomy* conditions. The higher tracking errors occurred when the participants were completing the communications task prior to the system identifying an overloaded workload state and automating the tracking task. Participants were able to complete the tracking task and the communications task simulataneously in the *Interaction* condition, due to being able to verbally interact with the communication task.

Table 5.4: Tracking Task: Root-Mean Squared Error Performance Means (Std. Dev.) by Evaluation Type. **Note:** Lower is Better.

Adaptation Type	**Underload**	**Normal Load**	**Overload**
Real-Time Evaluation	-	140.59 (93.83)	200.28 (111.6)
Both	**84.87 (55.83)**	**87.50 (52.44)**	126.14 (86.37)
Autonomy	100.82 (67.09)	119.62 (74.47)	115.84 (75.45)
Interaction	-	89.57 (59.10)	**112.98 (70.83)**

The participants were required to maintain the resource management task's primary fuel tanks' levels within the range of 2000 to 3000 units (Chapter 3.4). The overall percentage of time the tanks were in range by adaptation type are provided in Table 5.5. The participants maintained the fuel levels the best when the system adapted *Interactions* for each workload condition. The real-time evaluation achieved the lowest performance for the underload and normal load conditions, while the *Both* adaptation type performed the worst during the overload condition.

Table 5.5: Resource Management Task: Time in Range (%) by Evaluation Type. **Note:** Higher is Better.

Adaptation Type	**Underload**	**Normal Load**	**Overload**
Real-Time Evaluation	79	79	68
Both	84	76	61
Autonomy	91	85	93
Interaction	**92**	**91**	**99**

The system monitoring task consisted of resetting lights and gauges, if they went out of range. The participants' descriptive statistics for reaction time to the out of range lights or gauges by adaptation type and workload condition are provided in Table 5.6. The participants achieved the best performance in the underload and overload conditions when tasks were automated and the best performance in the normal load condition when interactions were adapted. The lowest performance was achieved when no adaptations occurred.

Table 5.6: System Monitoring Reaction Time Means (Std. Dev.) by Evaluation Type. **Note:** Lower is Better.

Adaptation Type	**Underload**	**Normal Load**	**Overload**
Real-Time Evaluation	4.32 (4.16)	5.38 (3.52)	6.25 (4.30)
Both	3.14 (2.19)	4.53 (2.55)	5.68 (3.67)
Autonomy	**2.86 (1.15)**	5.63 (3.45)	**4.52 (3.18)**
Interaction	5.68 (3.67)	**4.13 (2.60)**	5.83 (3.77)

A failure occurred if participants did not correct a light or gauge within fifteen seconds of when the light or gauge went out of range. The system monitoring success rate by adaptation type is provided in Table 5.7. The participants were less successful when no adaptation occurred. The *Both* and *Autonomy* adaptation types achieved roughly the same performance in the underload and normal load conditions, but the *Autonomy* type achieved better performance in the overload condition.

Table 5.7: System Monitoring Success Rate (%) by Adaptation Type and Workload Condition. **Note:** Higher is Better.

Adaptation Type	**Underload**	**Normal Load**	**Overload**
Real-Time Evaluation	69	79	60
Both	**100**	95	75
Autonomy	**100**	94	**85**
Interaction	92	**99**	76

The participants monitored and responded to simulated air-traffic control messages during the communication task. The average (Std. Dev.) time it took for participants to respond to air-traffic control messages by adaptation condition is provided in Table 5.8. The participants responded to messages quicker when *Both* adaptation types occurred and were the slowest when no adaptation occurred in the real-time evaluation. Similar reaction times were expected when no adaptation occurred and during the *interaction* adaptation type for the normal load condition, as the speech interaction adaptation did not occur during the normal load condition.

Overall performance was calculated using the methodology in Chapter 5.1.1. The resulting calculations are provided in Table 5.9. Directly comparing the calculated overall

Table 5.8: Communications Reaction Time Means (Std. Dev.) by Evaluation Type. **Note:** Lower is Better.

Adaptation Type	Normal Load	Overload
Real-Time Evaluation	10.41 (1.79)	9.83 (4.36)
Both	**8.68 (4.57)**	**3.46 (4.07)**
Autonomy	9.17 (4.9)	5.12 (5.49)
Interaction	10.35 (3.36)	4.36 (5.05)

task performance values between the adaptive conditions is confounded by what tasks were active during the conditions, as the overall task performance value was based on the active task set. For example, the tracking task was inactive during the underload condition for the *Interaction* adaptation type, but was active for the *Both* and *Autonomy* types. Thus, the tracking task performance may deflate the overall performance value artificially. Some general trends can be extrapolated. Adapting interaction modalities (*Interaction*) for each workload condition resulted in higher task performance, than the non-adaptive modalities (real-time evaluation). This comparison is not confounded, as both adaptation types had the same active task set.

Table 5.9: Calculated Overall Performance Descriptive Statistics by Adaptation Type and Workload Condition. **Note:** Higher is Better.

Adaptation Type	Underload	Normal Load	Overload
Real-Time Evaluation	0.85 (0.12)	0.72 (0.20)	0.56 (0.16)
Both	0.81 (0.12)	**0.83 (0.13**)	0.66 (0.27)
Autonomy	0.82 (0.06)	**0.81 (0.14)**	**0.77 (0.16)**
Interaction	**0.98 (0.04)**	**0.83 (0.12)**	0.72 (0.09)

5.5.1.1 Between Evaluations: Subjective Ratings

The in-situ workload ratings were assessed approximately every seven minutes during the adaptive system adaptation. An overview of the ratings' descriptive statistics is provided in Table 5.10. Overall workload was rated lower in the overload condition during the adaptive system evaluation and was similar across the two evaluations for the underload and normal conditions. The lower overall ratings for the overload condition are attributed to the cognitive, tactile, and visual workload components. The NASA-TLX was administered at the end of each evaluation and the average ratings for each NASA-TLX scale are provided in Table 5.11. There was no significant difference in the overall NASA-TLX ratings between the two evaluations. The participants were more frustated using the adaptive system, which may be attributed to the training session. The training session for the real-time evaluation and the pilot study did not use the adaptive system; thus, participants needed to get use to using the adaptive system.

Table 5.10: In-Situ Workload Ratings by Evaluation Condition and Type: No Adaptation vs Adaptation

Condition	**Evaluation**	**Auditory**	**Cognitive**	**Motor**	**Speech**	**Tactile**	**Visual**	**Overall**
Underload	Real-Time	1.2 (0.41)	1.65 (0.67)	1.45 (0.51)	1.2 (0.41)	1.3 (0.47)	1.8 (0.83)	8.6 (2.54)
	Adaptation	1.12 (0.5)	1.5 (0.73)	1.75 (0.93)	1.12 (0.34)	1.62 (0.96)	2.0 (0.82)	9.12 (3.07)
Normal Load	Real-Time	1.9 (0.64)	2.75 (1.02)	2.65 (0.88)	2.1 (0.79)	1.85 (0.59)	2.85 (1.04)	14.1 (3.68)
	Adaptation	2.25 (0.77)	2.69 (0.7)	2.12 (0.81)	2.12 (0.96)	2.0 (1.03)	2.88 (1.09)	14.06 (2.74)
Overload	Real-Time	2.9 (0.76)	3.7 (0.92)	3.1 (0.84)	2.8 (0.92)	2.63 (0.89)	3.57 (1.04)	18.7 (3.53)
	Adaptation	3.04 (1.37)	2.83 (1.17)	1.83 (0.82)	2.92 (1.28)	1.88 (1.19)	2.67 (1.09)	15.17 (4.06)

Table 5.11: NASA-TLX Workload Ratings by Evaluation Type: No Adaptation vs Adaptation

Evaluation	**Effort**	**Frustration**	**Mental**	**Performance**	**Physical**	**Temporal**	**Overall**
Real-Time	75.5 (14.62)	38.0 (22.51)	66.0 (16.8)	37.0 (22.14)	49.5 (25.22)	61.0 (22.83)	60.0 (12.25)
Adaptation	70.62 (16.13)	54.38 (18.41)	59.38 (25.28)	51.25 (24.6)	35.62 (21.62)	55.0 (20.35)	60.17 (13.97)

5.5.1.2 Between Evaluations Discussion

The developed adaptive system targets adaptations to specific workload channels in order to augment task performance. Hypothesis $\mathbf{H_1^A}$ predicted that participants will experience lower workload in the overload condition and higher workload for the underload condition on a between evaluation analysis. The hypothesis was supported for the overload condition, but not for the underload condition. Failing to support the hypothesis for the underload condition was attributed to the adaptive system's sub-optimal approach of transitioning the tracking task out of automatic mode. An adaptive system may allocate other tasks to the participant, but such an approach is not feasible with the NASA MATB-II task environment. The adaptive system cannot recreate communication requests for the participant to respond to or make the system monitoring task's alarms go out of range, as the alarms represented system states.

Targeting adaptations to the overall workload and its contributing components was expected to increase task performance. Hypothesis $\mathbf{H_2^A}$ predicted that higher task performance will be achieved using the adaptive system. This hypothesis was fully supported for the between evaluations analysis, as the highest task performance for each NASA MATB-II task occurred in one of the adaptive conditions (i.e., Both, Autonomy, or Interaction). Fully supporting this hypothesis does not provide insight into which adaptations were most effective for each NASA MATB-II task; thus, additional analysis were conducted to investigate the effectiveness of automating tasks and adapting interactions.

5.5.2 Adaptive Autonomy Results

Analyzing the impact of system's adaptations required parsing task performance data by autonomy and interaction modality. The first set of results parsed data by what tasks were being automated. Hypothesis $\mathbf{H_3^A}$ predicted that lower workload and higher task performance will be achieved in the *Both* condition when at least one task was automated.

Tasks were taken out of automation mode approximately twenty seconds into the underload condition. The resource management task was the only active task during these twenty seconds, as no communications requests or system monitoring alarms was scheduled to occur. Thus, the underload condition was not included in the adaptive autonomy analysis.

The system automated neglected tasks when it determined that a participant was overloaded or that their predicted performance was too low. Overall, at least one task was automated 40% of the time during the normal load condition and 75% of the time during the overload condition for the *Both* and *Autonomy* adaptive conditions. No tasks were automated during the underload condition. The system automatically responded to an out of range light or gauge 20% of the time in the normal load condition and 32% of the time in the overload condition, while the system automatically tuned a communications radio 20% and 32% of the time in the normal load and overload conditions, respectively. The tracking task was automated 16% of the time in the normal load condition and 42% of the time during the overload condition, while the resource management task was automated 48% and 33% of the time during the normal load and overload conditions, respectively. The workload estimates by autonomy and automated tasks are presented in Tables 5.12 and 5.13, respecively. The Task Region column represents where the participants were physically located (TRCK/SYS: Tracking and System Monitoring, COMM/RES: Communications and Resource Management). Automating tasks did not appear to impact the participants' overall workload, as similar overall workload estimates occurred for each adaptive condition and autonomy level (on or off).

Automating inactive tasks was expected to increase task performance. The descriptive statistics for the tracking task performance by adaptive condition and autonomy type are provided in Table 5.14, where the autonomy column indicates which tasks were being automated or if no tasks were being automates (i.e., None). For example, the COMM/RESvalue indicates that the communications and resource management tasks were being automated. The presented results pertain to when the participant was actively completing the tracking

Table 5.12: Algorithm Estimated Workload Means (Std. Dev.) by Autonomy Level for the *Both* Adaptive Condition. Note: TRCK/SYS represents the tracking and system monitoring tasks, while COMM/RES represents the communication and resource management tasks.

Workload	Autonomy	Task Region	Normal Load	Overload
Auditory	Off	TRCK/SYS	1.91 (1.39)	2.29 (1.37)
		COMM/RES	2.00 (1.15)	2.47 (1.30)
	On	TRCK/SYS	1.99 (2.09)	2.35 (1.42)
		COMM/RES	1.85 (1.21)	2.55 (1.19)
Cognitive	Off	TRCK/SYS	9.94 (3.71)	12.22 (4.35)
		COMM/RES	10.39 (3.43)	12.33 (3.64)
	On	TRCK/SYS	9.05 (4.00)	12.55 (4.12)
		COMM/RES	9.03 (3.70)	13.14 (3.63)
Physical	Off	TRCK/SYS	2.91 (3.43)	0.96 (1.21)
		COMM/RES	1.25 (2.26)	2.11 (2.44)
	On	TRCK/SYS	4.59 (4.01)	1.28 (2.55)
		COMM/RES	4.17 (4.42)	2.04 (2.33)
Speech	Off	TRCK/SYS	0.90 (1.46)	0.89 (1.44)
		COMM/RES	0.26 (0.89)	0.36 (0.98)
	On	TRCK/SYS	0.67 (1.32)	1.12 (1.52)
		COMM/RES	0.27 (0.87)	0.41 (1.06)
Overall	Off	TRCK/SYS	27.16 (6.15)	35.93 (7.05)
		COMM/RES	23.91 (5.74)	38.47 (5.46)
	On	TRCK/SYS	28.53 (6.77)	38.27 (6.85)
		COMM/RES	27.05 (7.33)	40.03 (5.31)

task or system monitoring task, as incorporating performance results when the participant was not physically located near these tasks does not provide the necessary insight into how adapting autonomy impacts the corresponding performance. Automating the communications and resource management task in the normal load condition produced similar tracking task performance as not automating the tasks for the *Both* and *Autonomy* conditions. A beneficial impact to performance was seen in the overload condition when the communications and resource management tasks were being automated within the *Both* and *Autonomy* conditions. Automation did not occur during the *Interaction* adaptive condition, but high performance was achieved in the condition.

The participants generally completed the tracking task at the same time as the system monitoring task; thus, a similar trend as the tracking task performance was expected for

Table 5.13: Algorithm Estimated Workload Means (Std. Dev.) by Autonomy and Task Region for the *Autonomy* Adaptive Condition.

Workload	Autonomy	Task Region	Normal Load	Overload
Auditory	Off	TRCK/SYS	2.00 (1.84)	2.77 (1.36)
		COMM/RES	2.03 (1.29)	2.83 (1.12)
	On	TRCK/SYS	1.21 (1.03)	2.14 (1.11)
		COMM/RES	1.71 (1.47)	2.69 (1.20)
Cognitive	Off	TRCK/SYS	9.83 (3.99)	14.38 (4.89)
		COMM/RES	8.72 (3.44)	13.71 (3.58)
	On	TRCK/SYS	9.97 (3.44)	11.13 (4.65)
		COMM/RES	9.81 (3.46)	12.18 (4.30)
Physical	Off	TRCK/SYS	3.78 (3.83)	2.44 (3.44)
		COMM/RES	3.43 (3.81)	1.41 (1.44)
	On	TRCK/SYS	5.74 (3.29)	1.02 (1.63)
		COMM/RES	4.28 (4.19)	2.07 (2.99)
Speech	Off	TRCK/SYS	0.16 (0.70)	0.46 (1.10)
		COMM/RES	0.21 (0.82)	0.05 (0.41)
	On	TRCK/SYS	0.08 (1.40)	0.56 (1.21)
		COMM/RES	0.0 (0.0)	0.39 (1.05)
Overall	Off	TRCK/SYS	27.36 (7.87)	40.40 (8.84)
		COMM/RES	25.29 (7.09)	39.65 (5.39)
	On	TRCK/SYS	31.36 (5.13)	36.27 (6.30)
		COMM/RES	27.56 (7.30)	39.09 (6.76)

Table 5.14: Tracking Error Means (Std. Dev.) by Adaptive Condition and Autonomy Level. **Note**: Lower is Better

Adaptive Condition	Autonomy	Normal Load	Overload
Both	None	82.87 (45.18)	133.05 (82.38)
	COMM/RES	80.61 (50.95)	116.37 (82.02)
Autonomy	None	105.16 (62.55)	145.16 (96.36)
	COMM/RES	111.81 (66.48)	114.53 (73.74)
Interaction	None	87.18 (54.65)	99.33 (58.03)

the system monitoring task performance. The reaction times and success rates for when the participants were actively completing the system monitoring task are provided in Tables 5.15 and 5.16, respectively, by adaptive condition and task autonomy. Automating the communications and resource management tasks did not impact the reaction times for the *Both* and *Autonomy* adaptive conditions, but participants had higher success rates when the tasks were being automated in the overload condition. This trend illustrates that participants were more likely to notice that a light or gauge was out of range when the communications and resource management tasks were automated. Autonomy negatively impacted task performance in the normal load condition for the *Autonomy* adaptive condition, resulting in higher reaction times and lower success rates. The *Interaction* condition resulted in similar or better performance than the other adaptive conditions for both workload conditions.

Table 5.15: System Monitoring Reaction Time Means (Std. Dev.) by Adaptive Condition and Autonomy Level. **Note**: Lower is Better

Adaptive Condition	**Autonomy**	**Normal Load**	**Overload**
Both	None	4.32 (2.03)	5.67 (3.62)
	COMM/RES	4.12 (2.16)	5.53 (3.43)
Autonomy	None	4.92 (2.42)	3.84 (3.26)
	COMM/RES	6.25 (3.26)	4.14 (2.78)
Interaction	None	4.42 (2.73)	4.87 (3.36)

Table 5.16: System Monitoring Success Rate by Adaptive Condition and Autonomy Level. **Note**: Higher is Better

Adaptive Condition	**Autonomy**	**Normal Load**	**Overload**
Both	None	98	20
	COMM/RES	97	64
Autonomy	None	98	55
	COMM/RES	87	75
Interaction	None	93	87

Automating the system monitoring and tracking tasks allowed participants to focus on the resource management and communications tasks. This focused attention may improve the resource management's task performance; the results for which are provided by adaptive condition and autonomy type in Table 5.17. The results for the TRCK/SYS rows correspond to the tracking and system monitoring tasks being automated. Similar to the tracking task's and system monitoring task's performance, autonomy positively impacted the resource management's task performance during the overload condition, but not for the normal load condition. Additionally, better performance was achieved during the *Interaction* adaptive condition, where no tasks were automated.

Table 5.17: Resource Management Time in Range (%) by Adaptive Condition and Autonomy Level. **Note**: Higher is Better

Adaptive Condition	Autonomy	Normal Load	Overload
Both	None	83	23
	TRCK/SYS	99	71
Autonomy	None	89	97
	TRCK/SYS	91	100
Interaction	None	100	100

Being able to focus on the communications task due to automating the tracking and system monitoring tasks was expected to improve participants' reaction times to a communications request. These reaction times are provided in Table 5.18. Automation did not seem to impact the communications task performance, as participants responded to communications requests in approximately the same amount of time to when tasks were not automated. This result is attributed to participants primarily prioritizing the communications tasks over the other tasks.

Table 5.18: Communications Reaction Time by Adaptive Condition and Autonomy Level. **Note**: Lower is Better

Adaptive Condition	Autonomy	Normal Load	Overload
Both	None	8.50 (4.34)	3.67 (4.40)
	TRCK/SYS	10.21 (3.55)	3.86 (4.37)
Autonomy	None	11.15 (0.99)	9.90 (3.45)
	TRCK/SYS	12.23 (2.03)	9.17 (4.59)
Interaction	None	10.35 (3.36)	4.36 (5.05)

Overall task performance was calculated for when tasks were automated or not, where the resulting performance by workload condition is provided in Table 5.19. The autonomy column represents which tasks were being automated. The participants achieved similar or lower overall task performance during the normal load condition when tasks were being automated, but achieved better performance during the overload condition with automated tasks. The *Interaction* condition achieved similar task performance to the *Both* and *Autonomy* adaptive conditions.

Table 5.19: Overall Task Performance by Adaptive Condition and Autonomy Level. **Note**: Higher is Better

Adaptive Condition	Autonomy	Normal Load	Overload
Both	None	0.82 (0.15)	0.44 (0.33)
	TRCK/SYS	0.79 (0.16)	0.75 (0.30)
	COMM/RES	0.82 (0.07)	0.66 (0.18)
Autonomy	None	0.82 (0.13)	0.44 (0.33)
	TRCK/SYS	0.80 (0.18)	0.77 (0.17)
	COMM/RES	0.74 (0.15)	0.74 (0.14)
Interaction	None	0.83 (0.12)	0.72 (0.09)

5.5.2.1 Adaptive Autonomy Discussion

Hypothesis $\mathbf{H_3^A}$ predicted that automating tasks will result in lower workload and higher task performance. The first portion of this hypothesis was not supported, as there was no discernible difference in workload, when tasks were being automated. This result is attributed to the system automating tasks that were not the participant's active task. For example, automating the tracking and system monitoring tasks when participants were completing the communications and resource management tasks did not impact the primary contributors (e.g, the communications and resource management tasks) to the participant's workload state. Thus, the adaptive system needs to automate neglected tasks and choose an appropriate level of autonomy for the active tasks in order to reduce workload appropriately. Choosing a level of autonomy for the active tasks is not trivial, as the unexpected automation may inadvertently increase workload.

Automating inactive tasks did not reduce workload, but may allow participants to perform better on the current tasks and increase overall sustained performance on all tasks, as the participants did not need to monitor the automated tasks and focused solely on the current tasks. Hence, the second portion of hypothesis $\mathbf{H_3^A}$ predicting that higher task performance will occur when tasks were being automated was partially supported, as higher performance was achieved when tasks were automated during the overload workload conditions. It may not appear that higher performance was achieved when tasks were being automated in the system monitoring task, but participants were more likely to respond to and properly repair an out of range light or gauge with a similar reaction time to no autonomy. Participants performed similar or worse when tasks were being automated to when tasks were not automated in the normal load condition. This trend may indicate a need to automate more tasks during the normal load condition, as more automation can decrease the amount of times participants walk between tasks.

The *Interaction* condition tended to have similar or higher task performance metrics to the other two conditions, despite no tasks being automated. This trend may be attributed to participants being able to verbally interact with the communications task, while completing the tracking and system monitoring tasks. The adaptive system prioritized automating the communication's task over allowing participants to verbally interact with the task during the *Both* condition, resulting in participants using the speech interaction modality less frequently during the normal load and overload conditions. Flipping the priority of automating tasks to rely on a speech interaction modality more may result in higher performance in the *Both* condition. Additionally, the resource management task can be neglected for long durations without considerable performance decrements (i.e., the participants can leave pumps on and intermentally check on the fuel levels). Thus; automating the resource management task has minimal impact on overall performance.

5.5.3 Interaction Modality Results

A primary contribution of the adaptive system was that interaction modalities were adapted based on the workload component estimates, not just the overall workload state. System interactions occurred in the form of alarms (e.g., a light or gauge goes out of range), where an auditory modality was used for the alarm, if the participant's speech and auditory workload channels were not loaded. Hypothesis $\mathbf{H_4^A}$ predicted that adapting the interaction modality appropriately will have a beneficial impact on the system monitoring task's performance and on the communications task's workload. Analyzing the communication task's performance, with respect to interaction modality adaptations was not completed, as reaction time corresponded to when the participant physically tunes a radio (i.e., pushes enter after tuning the radio). The radio automatically updated as soon as the participant started speaking, when the participant interacted via a speech modality. Thus, the reaction times between the two modalities do not represent the same performance. The tracking and resource management tasks were removed from the interaction modality analysis, as interactions were rarely adapted for these tasks.

The IMPRINT Pro workload models predicted that a conflict may occur with the participant's speech or auditory channels, if an auditory modality was used 33% (underload), 32% (normal load), and 85% (overload) of the time. The adaptive system identified a potential conflict 15% (underload), 65% (normal load), and 90% (overload) of the time, demonstrating that the system adapted interactions less frequently than expected for the underload condition and more frequently than expected for the normal load condition. A visual modality was selected for the underload condition, primarily due to the participants speaking. This result was expected, as no auditory conflicts (i.e., from the communications task) were to occur in the underload condition. The adaptive system selected a visual modality in the normal load and overload conditions, due to the participants speaking, or an auditory conflict from the communications task.

Changing an alarm's interaction modality intelligently was expected to have a beneficial effect on the system monitoring task performance. It was expected that using an auditory modality, when the participants were underloaded will lower reaction times and improve success rates over a visual modality. Likewise, using a visual modality, when the participants were overloaded will lower reaction times and improve success rates over an auditory modality. The performance metrics (reaction time and success rate) by interaction modality and task region are provided in Tables 5.20 and 5.21, respectively, where the task region represents the participants' physical location when a light or gauge went out of range. Overall, longer reaction times and lower success rates occurred when the participants were located by the communications or resource management tasks for both interaction modalities, which was expected. Using a visual modality typically elicited shorter reaction times, but lower success rates when the participants were located by the communications or resource management task. Similar reaction times and success rates were achieved across the two interaction modalities when the participants were completing the tracking or system monitoring task during the normal load or overload workload conditions. A complete comparison between the two interaction modalities cannot be conducted for the underload condition, as the participants were never in a specific task region when an alarm went off (e.g., participants were never near the communications/resource management tasks when a visual modality was selected during the *Both* condition).

Table 5.20: System Monitoring Reaction Times by Task Region, Adaptive Condition and Selected Interaction Modality.

Condition	Selected Modality	Task Region	UL	NL	OL
Both	Auditory	TRCK/SYS	3.34 (1.19)	4.55 (2.38)	5.31 (3.26)
		COMM/RES	11.76 (0.00)	6.08 (2.86)	10.21 (1.73)
	Visual	TRCK/SYS	2.15 (0.53)	4.25 (2.47)	5.47 (3.58)
		COMM/RES	-	5.44 (3.65)	9.21 (3.33)
Interaction	Auditory	TRCK/SYS	-	4.51 (3.33)	5.11 (3.30)
		COMM/RES	4.01 (1.39)	4.79 (3.59)	13.74 (0.00)
	Visual	TRCK/SYS	2.08 (0.10)	4.02 (2.04)	5.93 (3.83)
		COMM/RES	3.15 (1.83)	4.07 (2.60)	9.08 (2.68)
Autonomy	Visual	TRCK/SYS	2.76 (1.21)	4.49 (2.80)	3.77 (2.67)
		COMM/RES	2.46 (0.29)	4.72 (3.31)	8.34 (3.37)

Adapting interaction modalities pertains to changing how the human may interact with the system, as well as changing how the system interacts with the human. The participants were able to interact with the communications task via a physical or speech modality, based on their available workload resources. The participants interacted with the communications task via a speech modality 10% of the time in the normal load condition and 37% of the time in the overload condition. No communication requests occurred during the underload condition. The corresponding physical and speech workload estimates when an interaction occurred for the communications task are provided in Table 5.22. Speech workload was

Table 5.21: System Monitoring Success Rates by Task Region, Adaptive Condition and Selected Interaction Modality.

Condition	Selected Modality	Task Region	UL	NL	OL
Both	Auditory	TRCK/SYS	100	94	76
		COMM/RES	100	89	57
	Visual	TRCK/SYS	100	97	85
		COMM/RES	-	90	22
Interaction	Auditory	TRCK/SYS	-	100	83
		COMM/RES	100	100	50
	Visual	TRCK/SYS	100	100	82
		COMM/RES	75	92	42
Autonomy	Visual	TRCK/SYS	100	96	70
		COMM/RES	100	44	7

estimated to be higher when participants interacted using a speech modality, than when a physical interaction occurred for both adaptation conditions (*Both* and *Interaction*) during the normal load workload condition. A similar pattern occurred for the *Both* adaptive condition during the overload workload condition, but higher speech workload was estimated during the *Interaction* condition when participant used a physical interaction modality. The participants experienced higher physical workload when physically interacting with the communications task, as demonstrated by the workload estimates.

Table 5.22: Descriptive Statistics for the Speech and Physical Workload Component Estimates for the Communications Task by Adaptive Condition and Selected Modality.

Condition	Selected Modality	Normal Load		Overload	
		Speech	Physical	Speech	Physical
Both	Speech	1.98 (1.71)	3.77 (5.35)	0.85 (1.38)	1.88 (1.98)
	Physical	0.15 (0.70)	6.25 (4.92)	0.18 (0.76)	2.13 (2.02)
Interaction	Speech	2.20 (1.91)	0.88 (1.20)	1.65 (1.67)	1.47 (1.63)
	Physical	1.65 (2.34)	4.25 (5.65)	2.41 (1.50)	3.82 (4.00)

The participants were able to interact with the communications task via a speech modality approximately the same amount of time during the *Both* and *Interaction* conditions. However, the participants were typically not in the same location during the two conditions. 90% of the time participants were completing the tracking and system monitoring tasks, while verbally interacting with the communications task during the *Interaction* condition and 27% of the time they were doing the same during the *Both* condition. This physical location difference was due to the system monitoring and tracking tasks being automated during the *Both* condition.

5.5.3.1 Interaction Modality Discussion

The first portion of hypothesis $\mathbf{H_4^A}$ predicted that higher system monitoring task performance will be achieved when adapting interactions (i.e., an alarm used an auditory modality when appropriate). The hypothesis was not supported, as similar or lower reaction times occurred in the underload condition when a visual modality was selected. Similar reaction times occurred between the auditory and visual modalities for the overload condition as well. Not supporting the hypothesis may be attributed to the adaptive system's experimental design. The similar reaction times and success rates between the two interaction modalities may be attributed to the system selecting the appropriate modality. No data was collected corresponding to when an audible alarm conflicts with the communications task or the participant speaking. Additional results are required in order to determine if adapting an alarm's interaction modality is beneficial.

The second portion of hypothesis $\mathbf{H_4^A}$ was supported, as the expected workload impact of adapting how participants may interact with the communications task occurred. The participants experienced higher speech workload when using a speech modality vs. a physical modality, as expected. Additionally, the speech modality was primarily used when participants were located by the tracking and system monitoring tasks, when the system adapted interactions only. This result allowed the participants to complete three tasks at a time, rather than being limited to only two tasks simultaneously during the *Both* condition. Thus, the adaptive system allowed for more work to be completed in the *Interaction* condition, by selecting a speech modality for the communications task.

5.5.4 Within-Subjects Results

The within-subjects analysis focused specifically on the results from the two participants who completed both evaluations. Each participant completed the real-time evaluation (Chapter 4.1) approximately four months prior to completing the adaptive system evaluation and were acquaintances of the experimenter, who specifically ask these participants to complete the adaptive system evaluation. Participant P_1 completed the adaptive evaluation in the *Both-Autonomy* ordering, while P_2 completed the evaluation in the *Both-Interaction* ordering. The intent was for the *Both* condition to occur in the same order for both participants, which was randomly chosen to be first. Participant P_1 was randomly assigned the *Both-Autonomy* ordering, which meant participant P_2 received the *Both-Interaction* ordering.

It was expected that the two participants will experience lower workload and achieve higher performance during the adaptive evaluation, due to the system adaptations and prior experience. Hypothesis $\mathbf{H_5^A}$ predicted that the participants' objective and subjective workload ratings will be lower using the adaptive system in the overload condition and higher in the underload condition. Task performance was expected to be better using the adaptive system for all workload conditions, as predicted by hypothesis $\mathbf{H_6^A}$. The results were analyzed by evaluation type instead of by adaptive condition (i.e., *Both*) in order to determine the overall impact of the adaptive system.

Overall, adapting system autonomy and interactions was expected to impact the workload assessment algorithm's estimates. These estimates are provided in Table 5.23 by participant, evaluation type, and workload condition. Each participant experienced higher overall workload during the underload and normal load conditions, while lower overall workload was experienced during the overload condition when completing the adaptive evaluation. These differences in overall workload were attributed to the cognitive and auditory workload components, for which similar trends are seen in Table 5.23.

Automating all inactive tasks, based on participant workload, may allow participants

Table 5.23: Within Subjects Algorithm Estimated Workload by Workload Condition and Evaluation Type: Real-Time Evaluation vs Adaptation

Workload	Participant	Evaluation	Underload	Normal Load	Overload
Auditory	P_1	Real-Time	0.54 (1.12)	2.39 (1.39)	3.38 (0.87)
		Adaptation	1.82 (2.91)	2.16 (1.78)	2.42 (1.20)
	P_2	Real-Time	0.53 (1.07)	2.31 (1.39)	3.22 (0.88)
		Adaptation	1.1 (2.12)	2.04 (1.87)	2.13 (1.34)
Cognitive	P_1	Real-Time	3.07 (2.71)	9.58 (3.42)	15.92 (3.05)
		Adaptation	3.34 (3.21)	8.92 (3.51)	11.29 (3.99)
	P_2	Real-Time	2.68 (2.75)	9.70 (3.20)	14.98 (3.56)
		Adaptation	3.46 (3.45)	8.08 (3.18)	9.99 (3.74)
Physical	P_1	Real-Time	1.56 (1.72)	4.12 (4.45)	2.57 (2.62)
		Adaptation	1.97 (2.25)	3.31 (3.85)	2.30 (2.90)
	P_2	Real-Time	2.08 (2.99)	3.62 (4.62)	2.89 (1.72)
		Adaptation	2.70 (2.44)	2.93 (3.38)	2.27 (2.88)
Speech	P_1	Real-Time	0.07 (0.25)	0.79 (0.68)	2.03 (0.88)
		Adaptation	0.29 (0.9)	0.54 (1.21)	1.23 (1.57)
	P_2	Real-Time	0.05 (0.22)	0.75 (0.63)	2.09 (0.84)
		Adaptation	0.29 (0.91)	0.79 (1.35)	0.71 (1.31)
Overall	P_1	Real-Time	9.37 (4.61)	28.57 (5.95)	45.55 (5.1)
		Adaptation	11.55 (6.05)	26.6 (6.37)	38.91 (6.16)
	P_2	Real-Time	9.34 (5.7)	28.06 (6.18)	42.62 (4.89)
		Adaptation	11.56 (5.99)	25.34 (6.01)	36.40 (5.56)

to focus on the current task set and improve performance. The tracking task performance when the task was not automated is provided in Table 5.24 by participant and evaluation type. No results are provided for the Real-Time Evaluation evaluation during the underload condition, as the tracking task was always being automated. Overall, both participants achieved better performance for this task when using the adaptive system. Similar task performance was achieved for the overload condition when no adaptation occurred for both participants, while participant P_1 achieved higher performance than participant P_2 using the adaptive system. This higher performance is attributed to the autonomy adaptations P_1 experienced.

It was expected that participants will maintain the proper level ranges for the resource management's fuel tanks more of the time when using the adaptive system. The percentage

Table 5.24: Within Subjects RMSE Tracking Error Means (Std. Dev.) by Evaluation Type. **Note:** Lower is Better.

Participant	Evaluation	Underload	Normal Load	Overload
P_1	Real-Time	-	128.94 (75.27)	200.19 (104.09)
	Adaptation	89.72 (49.81)	**114.92 (82.29)**	**119.02 (67.85)**
P_2	Real-Time	-	147.22 (102.21)	200.75 (110.57)
	Adaptation	91.52 (70.7)	**108.36 (75.93)**	**132.09 (90.39)**

of time the tanks were in range for each participant and evaluation type are provided in Table 5.25. P_1 better maintained the fuel levels in the underload and overload conditions using the adaptive system, but performed worse during the normal load condition. P_2 achieved similar performance (within 5%) for both evaluations.

Table 5.25: Within Subjects Resource Management Time in Range (%) by Evaluation Type. **Note:** Higher is Better.

Participant	Evaluation	Underload	Normal Load	Overload
P_1	Real-Time	71	**97**	67
	Adaptation	**100**	86	**100**
P_2	Real-Time	**97**	84	**73**
	Adaptation	**100**	**89**	**73**

The participants were expected to have faster reaction times to an out of range light or gauge for the system monitoring task. The descriptive statistics for these reaction times are provided in Table 5.26 by participant and evaluation type. The adaptive system did not appear to impact the reaction times. P_1 had quicker reaction times in the underload condition using the adaptive system.

Table 5.26: Within Subjects System Monitoring Reaction Time Means (Std. Dev.) by Evaluation Type. **Note:** Lower is Better.

Participant	Evaluation	Underload	Normal Load	Overload
P_1	Real-Time	5.35 (4.06)	**4.85 (2.68)**	**6.56 (4.36)**
	Adaptation	**3.02 (1.35)**	**4.46 (1.98)**	**6.35 (4.33)**
P_2	Real-Time	**3.93 (3.19)**	**5.98 (4.27)**	**5.95 (4.42)**
	Adaptation	**3.87 (1.52)**	**5.29 (2.99)**	**5.35 (3.84)**

The similar reaction times across the two evaluation types may not reveal the impact of the adaptations, as participants may fail to respond to an out of range light or gauge within the fifteen second time limit. The participants' success rates by evaluation type are provided in Table 5.27. P_2 had higher success rates using the adaptive system, while P_1 had higher success rates during the normal load condition and the same success rate for the underload and overload conditions.

Table 5.27: Within Subjects System Monitoring Success Rate (%) by Evaluation Type. **Note:** Higher is Better.

Participant	Evaluation	Underload	Normal Load	Overload
P_1	Real-Time	**100**	85	**58**
	Adaptation	**100**	**97**	**58**
P_2	Real-Time	89	66	50
	Adaptation	**100**	**92**	**77**

The participants were to respond to simulated air traffic control communications. The average response times to these to theses communications are provided in Table 5.28. Lower reaction times occurred when the adaptive system was used for all workload conditions. The two participants achieved similar reaction times for the adaptive evaluation during the normal load condition, but P_2 had lower reaction times than P_1 in the overload condition.

An anticipated effect was found for the participants' subjective ratings. The in-situ workload ratings are provided in Table 5.29. The workload ratings demonstrated lower workload during the normal load and overload conditions when using the adaptive system for both participants. P_2 rated their overall workload higher for the underload condition during the adaptive evaluation, while P_1 had similar workload ratings. The respective

Table 5.28: Within Subjects Communications Reaction Time Means (Std. Dev.) by Evaluation Type. **Note:** Lower is Better.

Participant	Evaluation	Underload	Normal Load	Overload
P_1	Real-Time	-	9.24 (1.67)	10.51 (3.34)
	Adaptation	-	**6.92 (4.64)**	**3.32 (5.33)**
P_2	Real-Time	-	13.51 (1.35)	9.00 (4.91)
	Adaptation	-	**6.01 (5.59)**	**1.59 (2.69)**

NASA-TLX results are provided in Table 5.30. P_1 had lower NASA-TLX ratings when using the adaptive system, while participant P_2 had similar ratings for the evaluations.

Table 5.29: Within-Subjects In-Situ Workload Ratings by Evaluation Type: Real-Time Evaluation vs Adaptation

Condition	Participant	Evaluation	Auditory	Cognitive	Motor	Speech	Tactile	Visual	Overall
Underload	P_1	Real-Time	1.0 (0.0)	1.5 (0.71)	1.5 (0.71)	1.0 (0.0)	1.5 (0.71)	1.5 (0.71)	8.0 (2.83)
		Adaptation	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	2.5 (2.12)	1.0 (0.0)	7.5 (2.12)
	P_2	Real-Time	1.0 (0.0)	1.0 (0.0)	2.67 (1.15)	1.0 (0.0)	2.33 (1.53)	2.0 (1.0)	10.0 (3.61)
		Adaptation	3.5 (2.12)	3.5 (2.12)	2.5 (0.71)	3.0 (2.83)	3.0 (1.41)	4.0 (0.0)	19.5 (9.19)
Normal Load	P_1	Real-Time	2.5 (0.71)	3.0 (0.0)	2.5 (0.71)	2.5 (0.71)	2.0 (0.0)	3.0 (0.0)	15.5 (2.12)
		Adaptation	1.5 (0.71)	1.5 (0.71)	2.0 (1.41)	1.5 (0.71)	2.5 (0.71)	2.5 (0.71)	11.5 (4.95)
	P_2	Real-Time	5.0 (0.0)	4.0 (0.0)	4.0 (0.0)	4.0 (1.41)	4.0 (0.0)	3.0 (0.0)	24.0 (1.41)
		Adaptation	2.5 (2.12)	2.5 (0.71)	2.5 (0.71)	2.5 (2.12)	3.5 (0.71)	4.5 (0.71)	18.0 (5.66)
Overload	P_1	Real-Time	4.0 (1.0)	4.0 (1.0)	3.33 (1.15)	4.0 (1.0)	3.33 (1.15)	4.0 (1.0)	22.67 (5.86)
		Adaptation	3.0 (1.0)	1.0 (0.0)	1.0 (0.0)	3.0 (1.0)	1.0 (0.0)	1.0 (0.0)	10.0 (2.0)
	P_2	Real-Time	5.0 (0.0)	3.0 (0.0)	4.0 (0.0)	5.0 (0.0)	4.0 (0.0)	3.5 (0.71)	24.5 (0.71)
		Adaptation	2.0 (1.73)	1.67 (1.15)	1.0 (0.0)	2.0 (1.73)	1.33 (0.58)	1.67 (1.15)	9.67 (5.51)

Table 5.30: Within Subjects NASA-TLX Workload Ratings by Evaluation Type: Real-Time Evaluation vs Adaptation

Participant	**Evaluation**	**Effort**	**Frustration**	**Mental**	**Performance**	**Physical**	**Temporal**	**Overall**
P_1	Real-Time	50.0	30.0	55.0	35.0	20.0	55.0	45.67
	Adaptation	10.0	10.0	20.0	50.0	10.0	10.0	22.67
P_2	Real-Time	15.0	5.0	25.0	45.0	5.0	15.0	22.33
	Adaptation	15.0	5.0	20.0	25.0	20.0	45.0	23.67

5.5.4.1 Within-Subjects Discussion

Two participants completed both the real-time and the adaptive system evaluation, which facilitates analyzing potential effects from a within-subjects perspective, at least on a very limited basis. Hypothesis $\mathbf{H_5^A}$ predicted that both participants will experience lower workload in the overload condition using the adaptive system and higher workload in the underload condition. The overall workload estimates fully support this hypothesis, but the in-situ subjective ratings only partially support the hypothesis. The subjective ratings assess perceived workload and not the actual workload experienced; thus, this result was not surprising.

The difference in task performance for the two evaluations was evaluated for hypothesis $\mathbf{H_6^A}$, which predicted that using the adaptive system will lead to higher task performance. The participants achieved similar or better task performance when using the adaptive system, which supports hypothesis $\mathbf{H_6^A}$. Similar system monitoring reaction times were achieved during both evaluations; however, higher success rates occurred with the adaptive system. This trend highlights that participants addressed correctly more out of range lights and gauges using the adaptive system; thus, more work was completed.

5.6 Summary

The adaptive human-robot teaming system relied on workload component estimates and a performance prediction model, which predicted task performance for 1-minute into the future. These future performance predictions provided valuable insight into if performance decrements were going to occur; thus, allowing the system to identify adaptations. However, the performance prediction model was only analyzed for the NASA MATB-II and for its ability to predict current and future (1-minute) task performance. Further investigation is required to determine the model's predictive capabilities for differing time-steps (e.g., 30-seconds) and in different task environments.

The state-of-the-art adaptive systems adapt the system's autonomy based on cognitive workload or the overall workload state in order to augment performance. However, targeting adaptations to specific workload channels can permit more intelligent adaptations and further task performance. An adaptive system was developed in order to investigate how targeting adaptations impacts task performance, where the system's desired functionality and effectiveness was demonstrated in a pilot study. The participants who used the adaptive teaming system achieved higher task performance, than participants who did not use the adaptive system, which based on the pilot study, illustrated the system's effectiveness. The developed adaptive teaming system is the first system capable of adapting autonomy levels and interaction modalities.

The system adaptions (autonomy levels and interactions) were analyzed by examining their impact on workload and task performance. Automating tasks did not have the expected impact on workload, but did allow the participants to focus on the non-automated tasks and generally achieve higher performance. However, similar performance was achieved when adapting either the interactions or the autonomy. This result was attributed to participants being able to complete the tracking, system monitoring, and communications task simultaneously when a speech interaction modality was active for the communications task. This result demonstrates that the adaptive system was able to balance workload across the workload channels. Specifically, the physical workload requirement of walking between the tasks was allocated to the participant's speech workload channel.

Workload was used to determine the adaptive system's adaptations in order to augment task performance, as workload is an indirect measure of task performance. Directly measuring task performance may appear as a more accurate and informative measure on which to base system adaptations, but there are several limitations to such an approach. First, task performance may be difficult to measure directly in dynamic task environments (i.e., first response domains); thus, relying solely on task performance limits the range of environments in which the adaptive system can be deployed. Second, an overall task performance measure does not provide meaningful information about how a specific interaction will affect the human (e.g., will an auditory modality decrease task performance due to resource conflicts). These limitations demonstrate that relying on workload as a surrogate for task performance provides for more robust system adaptations and allows for the adaptive system to be deployed in multiple task environments.

The current adaptive system considers the workload component information as independent estimates, but these estimates are likely correlated. A task analysis may be used to identify the common means to complete the subtasks, which will identify what specific workload channels will be used. Additionally, a correlation or principle component analysis between the component estimates may help determine what the interdependicies are between the components. This information may allow for a more representative overall workload estimate that weights each workload component accordingly. Having a more representative overall estimate will allow the adaptive system to better reason how an adaptation will impact the human. For example, using a visual modality (i.e., text) for an alarm will have a cognitive component associated with the alarm. If the human is nearing an overloaded state, the system may reason that the visual modality will overload the human's cognitive workload and choose not to use that specific modality, as an auditory modality may be more appropriate.

The adaptive system was reliant on continuous workload estimates, which permits treating the system as a multi-variate control system. The workload estimates act as sensors with the same update rates (5-seconds), while the system adaptations can be considered to be system corrections. The update rate of the system corrections can impact the stability and controllability of the control system. If these corrections are updated too frequently, then the system will oscillate, resulting in unstable system states. For example, continuously invoking and revoking automation may increase the human's workload level, as the human has to reassess the task states constantly in order to have appropriate system awareness. Adapting too slowly will result in the system never reaching the desired steady-state (i.e., the human is performing optimally). For example, if autonomy decisions are considered once every 5-minutes, then the human may be in an overloaded state for at most 5-minutes and be making multiple errors during that time frame. Additionally, relying on data from 5-minutes prior is insufficient for determining system interactions, such as choosing an auditory alarm for the NASA MATB-II's system monitoring task. Choosing an appropriate adaptation update rate is non-trivial and likely domain and task specific. Adaptive system designers likely need real-world data from the specific domain and tasks in order to determine proper adaptation update rates and their corresponding impact on the adaptive system's stability.

Other aspects of control theory may be applicable to the adaptive system as well. A proportional integral derivative controller may be applied to the system in order to maintain the human's performance at a desired level. The controller can provide information regarding how much to adapt. If the human is overloaded, then the controller can determine how much autonomy is needed in order to normalize the human's workload level. However, the adaptive system needs to have the necessary level of controllability in order to invoke the needed amount of automation. For example, the current adaptive system has two autonomy levels (on/off) for each NASA MATB-II task and automating neglected tasks did not impact the human's workload level effectively. The controller may automate each task, which may reduce vigiliance. Thus, the current system's controllability may be considered to be low. Implementing more than two levels of autonomy for each task will increase the system's controllability and will allow for a proportional integral derivative controller to be more effective.

The evaluation, as a pilot study, had a limited number of participants, but there was a general increase in task performance when compared to the real-time evaluation participants' performance. Task performance was sampled frequently throughout the evaluation, which decreases the number of participants required to indentify an effect. However, this frequent sampling will not factor out individual differences between participants. A full human-subjects evaluation is necessary in order to better understand the impacts of individual differences on the system's ability to perceive, predict and adapt the human's and system's performance. This future evaluation needs to manipulate the adaptive conditions as the within-subjects variables, while also including a no adaptation condition. Thus, the individual participant performance between the adaptation and no adaptation conditions can be analyzed. Additionally, the no adaptation trial needs to use an auditory modality for all alarms in order to determine if the auditory alarm conflicts with the communication task and reduces performance.

The adaptive system was tailored to the NASA MATB-II, but the system architecture was designed to be generalizable. The *Perceive* state incorporated activity recognition, the workload assessment algorithm, and performance prediction, which can be used in other task domains, assuming that the algorithms are trained sufficiently. The *Select* stage needs to be tailored to the task domain, which is to be expected. Adaptive system designers need to provide information regarding what tasks can be automated and what interactions can occur. For example, applying the adaptive system architecture to a peer-based human-robot team requires the system designers to provide all potential robot interactions.

The adaptive human-robot teaming architecture may be applied to the peer-based evaluation, similar to that described in Chapter 3.3, where the participants completed a search-and-rescue scenario with a robot teammate. The participants were required to speak throughout the scenario, while the robot communicated messages from incident command or provided containment sampling instructions. Workload channel conflicts may occur if the robot communicated a message audibly, when the participant was speaking. Thus, the robot must use a different modality or wait until the participant's speech workload channel is no longer loaded. The adaptive system can identify these potential conflicts, using the interaction decision node in Figure 5.2, postpone the communication until the participant is no longer speaking or use another interaction modality. For example, the robot may send a text message (tactile and visual modality), instead of audibly communicating the message.

Automating tasks for the peer-based evaluation when the participant is overloaded may be difficult, as the robot may be either completing its own task or communicating instructions to the participant. However, the participant is required to communicate current task information to incident command. Thus, the robot may communicate the task information to incident command, if the participant is overloaded, so that the participant can better focus on the current task. Essentially, the robot re-allocates a task from the participant to itself, which can be considered a form of autonomy.

Task domains other than the NASA MATB-II contain aspects that may impact the adaptive system's scalability. Representive task domains may team multiple humans and robots to work together. These domains will require ntegrating methods, such as collation formation, into the adaptive system architecture, where tasks can be allocated to different humans and robots based on task schedules and the humans' workload levels. Collation formation may also be useful for when there is significant distance or communication latency between tasks or agents, as collation formation may incorporate temporal and task scheduling information. However, the adaptive system will have difficulty scaling to incorporate heterogenous, uncoupled, or unstructured tasks. For example, the system may have to incorporate task priorities and be able to probabilistically predict a task's workload composition in order to scale to heterogenous and unstructured tasks. These adaptive system extensions are outside the scope of this dissertation.

Overall, the adaptive system was demonstrated to improve task performance over a version of the system without adaptations. This task performance increase was primarily attributed to the system being able to select the appropriate interaction modality for the system monitoring and communications task. Automating inactive tasks was also beneficial, as the automation allowed participants to focus their attention on the active tasks and increase overall task performance. Further analyses and human subject evaluations are needed to better understand how to adapt system autonomy levels and interactions appropriately, but the developed system is a necessary step towards effective human-robot teaming architectures.

Chapter 6

Conclusion

The ability to augment task performance in high intensity domains (i.e., a NASA Control Room) by adapting to human workload states has gained considerable research interest, as multiple adaptive systems have been developed or theorized. These systems have primarily focused on adaptive automation, where tasks are automated based on the human's overall workload state. However, the overall workload state does not promote understanding why the human is in the current state, which is needed in order to enable more intelligent adaptations. Additionally, the overall workload state does not provide any information about workload channel conflicts, which is needed to adapt interaction modalities. This dissertation developed a diagnostic workload assessment algorithm to provide an adaptive system with information about the overall workload state and its contributing components: auditory, cognitive, physical, speech, and visual. The algorithm was validated using data from two supervisory-based evaluations and a peer-based evaluation, where the peer-based evaluation was conducted by a prior PhD student. A performance prediction model used the workload assessment algorithm's estimates to predict current and future task performance in order to provide additional information to the adaptive teaming system. The developed system was designed to be applied across task domains and teaming roles. A pilot study was conducted to demonstrate the adaptive system's effectiveness in augmenting task performance.

6.1 Contributions

The dissertation resulted in a number of contributions to the field, each of which are summarized below:

1. The primary contribution was a workload assessment algorithm capable of providing a complete assessment of overall workload, by assessing each individual workload component. The workload component estimates ensure that the algorithm is generalizable across tasks and human-robot teaming paradigms, as different tasks and paradigms may encompass different workload components. Knowledge of the underloaded and overloaded workload components provided by the algorithm enables the adaptive system to understand the complete workload state of an individual and intelligently target adaptations based on this knowledge.

   The developed algorithm is the first algorithm capable of estimating each workload component and overall workload in real-time. These real-time workload estimates are needed in order for an adaptive system to understand how an adaptation may impact the human's workload level. Additionally, a system may reason over multiple of these adaptations and their corresponding impact to human workload in order to determine the most effective adaptation to implement.

   Multiple task environments were used to validate the developed workload assessment algorithm's ability to estimate workload. These environments had varying workload contributions, demonstrating that the algorithm is not constrained to a single task environment with specific workload contributions. No other workload assessment algorithm has been validated to develop similar estimates for such a range of task environments.

2. A second contribution was an adaptive human-robot teaming system that improved task performance by automating tasks or selecting appropriate interaction modalities. The system targeted these adaptations based on specific workload components and the overall workload state, where current adaptive systems only consider the overall or cognitive workload state. A pilot study demonstrated that targeting adaptations intelligently can improve task performance in a supervisory-based human-robot team-

ing paradigm.

The existing adaptive system literature results only focus on adapting the level of autonomy with a perfect system (i.e., optimal task performance is achieved with full autonomy). The existing results ignore imperfect systems and other types of adaptation, such as changing interaction modalities. The developed adaptive system is the first system capable of adapting task autonomy and interactions in order to improve performance.

3. The third contribution was a performance model that relies on the workload assessment algorithm. The performance model predicted if performance will decrease, increase, or remain stagnant for future time-steps. Current systems that predict performance use small segments of data (i.e., 5-seconds) to determine if performance will decrease for the current or immediately following time-step. However, the state-of-the-art systems fail to predict when performance may decrease, increase, or remain the same for future time-steps, which is needed to determine what adaptations occurs when and how.

   Elaborate

6.2 Future Work

There are multiple future directions, which are summarized in Table 6.1 and discussed in the following paragraphs.

Workload Component Interdependicies

The developed workload assessment algorithm considers the workload components as independent measures, as the components are uniformly aggregated into an overall workload estimate. However, correlations exists between the workload components. Future work will quantify the interdependicies between the workload components using task, cor-

Table 6.1: An Overview of the Future Research Directions.

Future Research Questions
Workload Component Interdependicies
Performance prediction model
Better emulate real-world workload conditions
Window size impact on the speech-based feature extraction process
Control system analysis System adaptation impact on situational awareness
Coalition formation extension
Activity recognition
Fine motor, gross motor, and tactile workload component estimation
Visual workload estimation
Finer-grained speech workload model development

relations, and principle component analyses. The analyses will determine what weights are needed for each workload component for the overall workload aggregation. These weights are likely task environment specific; thus, data from the peer-based, supervisory-based, and real-time evaluations will be used to determine the workload component interdependicies.

Performance Prediction Model

The developed adaptive system was demonstrated to improve task performance by targeting adaptations to specific workload channels; however, a full human-subjects evaluation is required in order to fully comprehend the system's impact. First, the system needs incorporate additional, rather than two, levels of autonomy, which will allow the adaptive system to better assist overloaded or underloaded humans. Second, the adaptations' and interactions' expected impact on workload needs to be used by the performance prediction model in order to determine how the adaptation or interaction needs to occur. If the adaptation's expected impact on workload shows a decrease in the performance prediction value, then another adaptation needs to occur,

Better Emulate Real-World Workload Conditions

The algorithm was shown to estimate workload accurately in emulated workload conditions, where each workload condition's duration was uniform (5-minutes). Real-world environments may not contain such task and workload level uniformity; thus, future work will randomly order and choose each workload condition and time-frame in order to determine the corresponding impact on accuracy. The minimum workload condition duration needs to be 30-seconds for the workload metrics to be able to sense the workload change.

Window Size Impact on the Speech-Based Feature Extraction Process

The real-time workload assessment algorithm was the foundation of the adaptive system, where the algorithm estimated speech-workload. However, extracting speech-based features for the speech workload estimation was relatively computationally expensive, due to using the Fast Fourier Transform. This feature extraction process will impact the workload assessment algorithm's real-time capabilities, if other window sizes are used for the speech-based data. For example, using a 1-minute window may not allow the algorithm to estimate workload every five seconds, due to the feature extraction computation time. It is expected that at least a 2-second window size is needed in order to have accurate speech workload estimates. Future work will investigate methods for reducing the feature extraction computation time and determine how quickly the algorithm can estimate workload when larger window sizes are used.

Control System Analysis

The adaptive system may be treated as a multi-variate control system, as described in Chapter 5.6. Future work will investigate control theory and how it may be applied to the adaptive system. This investigation will determine what optimal adaptation rate is necassary using the data from the real-time evaluation and the adaptive system pilot study. Additionally, a methodology for applying various controllers (i.e., a proportional integral derivative controller) to determine how much adaptation is needed (i.e., how much autonomy is needed for each task) will be developed and validated in a future evaluation.

System Adaptation Impact on Situational Awareness

An important aspect of human-robot teaming is situational awareness, which was not assessed during the pilot study. The full human-subjects evaluation needs to incorporate the situational awareness probes [44] in order to properly measure the participant's situational awareness. This information will permit analyzing how the autonomy impacts a participant's ability to understand the current system's global state and accurately predict future states.

Coalition Formation Extension

The current adaptive system was evaluated in a one-to-one supervisory-based task environment, but other task domains may incoporate multiple human and/or system team members. Allocating tasks amongst the multiple team members may allow for workload balancing. Currently, the adaptive system has no mechanism to balance workload for multiple humans, but can be extended to do so using coalition formation [104]. Coalition formation algorithms seek to form teams of agents or individuals to accomplish a task set, while optimizing an objective function. Such algorithms are viable for workload balancing, but choosing the appropriate algorithm for a task environment is difficult. The intelligent Coalition Formation for Humans and Robots decision support system [108] can be incorporated into the *Determine Adaptations* stage in order to select the appropriate coalition formation algorithm and apply that algorithm to (re-)allocating tasks. The decision support system will be extended to reason over various mission features affecting task performance (i.e., human workload, task preemption, number of task switches, and task priority levels) in order to allocate task effectively.

Activity Recognition

Accurate workload estimates were required to intelligently target adaptations, but the developed workload assessment algorithm requires contextual features in order to accurately determine the human's current tasks and provide more accurate adaptations. The pilot study used system information to determine the participant's current task focus in order to calculate these contextual features, but knowing the task focus is not trivial in dynamic domains. Thus, a more sophisticated activity recognition approach needs to be integrated into the adaptive human-robot system architecture. Related work has investigated activity recognition in a clinical domain using data from the Myo wearable sensor device [52]. The activity recognition algorithm decomposed tasks using hierarchical task decomposition and used video data to generate contextual information, which was shown to improve recognition accuracy. This work demonstrated the potential of recognizing the human's current task in a dynamic and complex domain. The real-time evaluation and the adaptive system study collected Myo data; thus, future work will investigate using the collected Myo data in an activity recognition algorithm for the NASA MATB-II. This activity recognition algorithm will be incorporated into the developed adaptive system.

Fine Motor, Gross Motor, and Tactile Workload Component Estimation

The workload assessment algorithm had the most difficulty estimating physical workload, which may be attributed to combining the gross motor, fine motor, and tactile workload components into a physical workload component. This combination was due to the supervisory-based and peer-based evaluations not collecting any workload metric data sensitive to the fine motor and tactile workload components. The real-time evaluation and adaptive system study collected electromyography and arm movement data via the Myo armband, where the data may be sensitive to the fine motor and tactile workload components. The developed workload assessment algorithm will be extended to estimate these components using the collected data from the Myo armband.

Visual Workload Estimation

Dynamic domains will likely require more accurate visual workload information. The current workload assessment algorithm uses the IMPRINT Pro visual workload model to assess visual workload, but these models may become inaccurate in dynamic domains. The workload assessment algorithm may estimate visual workload by relying on visual workload metrics (e.g., pupil dilation or blink rate). These metrics are typically captured via eye-trackers, but these devices are currently not viable for non-stationary task environments or outdoor domains.

Finer-Grained Speech Workload Model Development

Speech workload was difficult to estimate, due to the trinary nature of the speech IMPRINT Pro workload model. A finer-grained workload model is needed in order to for the workload assessment algorithm to estimate speech workload accurately. Future work will extend the IMPRINT Pro speech workload model to be more sensitive to workload by developing finer-grained task anchors, on which IMPRINT Pro relies. The speech in-situ workload ratings and speech-based workload metrics collected during the supervisory and real-time evaluations may be used to develop more task anchors for IMPRINT Pro.

BIBLIOGRAPHY

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

[2] Abbass, H. A., Tang, J., Amin, R., Ellejmi, M., and Kirby, S. (2014). Augmented cognition using real-time EEG-based adaptive strategies for air traffic control. In *Human Factors and Ergonomics Society Annual Meeting*, volume 58, pages 230–234. SAGE Publications.

[3] Ahlstrom, U. and Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, 36(7):623–636.

[4] Archer, S., Gosakan, M., Shorter, P., and Lockett, J. (2005). New capabilities of the Armys maintenance manpower modeling tool. *Journal of the International Test and Evaluation Association*, 26(1):19 – 26.

[5] Assefi, M., Wittie, M., and Knight, A. (2015). Impact of network performance on cloud speech recognition. In *International Conference on Computer Communication and Networks*, pages 1–6.

[6] Backs, R. W. and Walrath, L. C. (1992). Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Applied Ergonomics*, 23(4):243–254.

[7] Baldwin, C. L. (2012). *Auditory Cognition and Human Performance*. CRC Press New York.

[8] Berthold, A. and Jameson, A. (1999). Interpreting symptoms of cognitive load in speech input. *Conference on User Modeling*, pages 235–244.

[9] Besson, P., Dousset, E., Bourdin, C., Bringoux, L., Marqueste, T., Mestre, D. R., and Vercher, J. L. (2012a). Bayesian network classifiers inferring workload from physiological features: Compared performance. In *IEEE Intelligent Vehicles Symposium*, pages 282–287.

[10] Besson, P., Maiano, C., Bringoux, L., Marqueste, T., Mestre, D. R., Bourdin, C., Dousset, E., Durand, M., and Vercher, J.-L. (2012b). Cognitive workload and affective state: A computational study using bayesian networks. In *IEEE International Conference on Intelligent Systems*, pages 140–145.

[11] Bian, D., Wade, J., Swanson, A., Weitlauf, A., Warren, Z., and Sarkar, N. (2019). Design of a physiology-based adaptive virtual reality driving platform for individuals with asd. *ACM Transactions on Accessible Computing (TACCESS)*, 12(1):2.

[12] Bishop, C. M. (2006). *Pattern recognition*, volume 128. Springer-Verlog New York.

[13] Bishop, C. M. (2012). Model-based machine learning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):2012– 2222.

[14] Boles, D. B. and Adair, L. P. (2001). The multiple resources questionnaire (MRQ). *Human Factors and Ergonomics Society Annual Meeting*, 45(25):1790–1794.

[15] Boles, D. B., Bursk, J. H., Phillips, J. B., and Perdelwitz, J. R. (2007). Predicting dual-task performance with the multiple resources questionnaire (MRQ). *Human Factors*, 49(1):32–45.

[16] Borghetti, B. J., Giametta, J. J., and Rusnock, C. F. (2017). Assessing continuous operator workload with a hybrid scaffolded neuroergonomic modeling approach. *Human factors*, 59(1):134–146.

[17] Brenner, M., Doherty, E., and Shipp, T. (1994). Speech measures indicating workload demand. *Aviation, Space, and Environmental Medicine*, 65(1):21–26.

[18] Brouwer, A.-M., Hogervorst, M. A., van Erp, J., Heffelaar, T., Zimmerman, P. H., and Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. *Journal of Neural Engineering*, 9(4):45–48.

[19] Byrne, E. A. and Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological psychology*, 42(3):249–268.

[20] Cain, B. (2007). A review of mental workoad literature. techreport RTO-TR-HFM-121-Part-II, Defence Research and Development Toronto.

[21] Casali, J. and Wierwille, W. (1983). A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load. *Human Factors*, pages 623–642.

[22] Castor, M. (2003). *GARTEUR Handbook of Mental Workload Measurement*. GARTEUR technical publications. Group for Aeronautical Research and Technology in Europe.

[23] Chavaillaz, A., Wastell, D., and Sauer, J. (2016). System reliability, performance and trust in adaptable automation. *Applied Ergonomics*, 52:333–342.

[24] Chen, F. (2013). Effects of cognitive load on trust. Technical Report AOARD-124076, National ICT Australia Limited.

[25] Christensen, J. C., Estepp, J. R., Wilson, G. F., and Russell, C. A. (2012). The effects of day-to-day variability of physiological data on operator functional state classification. *NeuroImage*, 59(1):57–63.

[26] Clark, J. B. and Allen, C. S. (2008). Acoustics issues. In *Principles of Clinical Medicine for Space Flight*, pages 521–533. Springer Nature.

[27] Comstock, J. R. and Arnegard, R. J. (1992). The multi-attribute task battery for operator workload and strategic behavior research. Technical Report NASA Tech. Memorandum 104174, NASA Langley Research Center.

[28] Cooper, G. and Harper, R. (1969). The use of pilot rating in the evaluation of aircraft handling qualities. Technical report, AGARD Report 567.

[29] Dietterich, T. G. (2002). *Ensemble learning*, volume 2. MIT Press: Cambridge, MA.

[30] Dorneich, M. C., Passinger, B., Hamblin, C., Keinrath, C., Vašek, J., Whitlow, S. D., and Beekhuyzen, M. (2017). Evaluation of the display of cognitive state feedback to drive adaptive task sharing. *Frontiers in neuroscience*, 11:144.

[31] Durkee, K., Geyer, A., Pappada, S., Ortiz, A., and Galster, S. (2013). Real-time workload assessment as a foundation for human performance augmentation. In Schmorrow, D. D. and Fidopiastis, C. M., editors, *Foundations of Augmented Cognition*, pages 279–288. Springer-Verlog Switzerland.

[32] Durkee, K., Pappada, S., Ortiz, A., Feeney, J., and Galster, S. (2015a). Using context to optimize a functional state estimation engine in unmanned aircraft system operations. In Schmorrow, D. D. and Fidopiastis, C. M., editors, *Foundations of Augmented Cognition*, pages 24–35. Springer-Verlog Switzerland.

[33] Durkee, K. T., Pappada, S. M., Ortiz, A. E., Feeney, J. J., and Galster, S. M. (2015b). System decision framework for augmenting human performance using real-time workload classifiers. In *IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision*, pages 8–13.

[34] Fan, J., Wade, J. W., Key, A. P., Warren, Z. E., and Sarkar, N. (2018). Eeg-based affect and workload recognition in a virtual driving environment for asd intervention. *IEEE Transactions on Biomedical Engineering*, 65(1):43–51.

[35] Feigh, K. M., Dorneich, M. C., and Hayes, C. C. (2012). Toward a characterization of adaptive systems: A framework for researchers and system designers. *Human Factors*, 54(6):1008–1024.

[36] Fuchs, S. and Schwarz, J. (2017). Towards a dynamic selection and configuration of adaptation strategies in augmented cognition. In *International Conference on Augmented Cognition*, pages 101–115. Springer.

[37] Gawron, V. J. (2008). *Human Performance, Workload, and Situational Awareness Measures Handbook*. CRC Press New York.

[38] Ghosh, A., Danieli, M., and Riccardi, G. (2015). Annotation and prediction of stress and workload from physiological and inertial signals. *IEEE Conference on Engineering in Medicine and Biology Society*, pages 1621–1624.

[39] Grier, R., Wickens, C., Kaber, D., Strayer, D., Boehm-Davis, D., Trafton, J. G., and St. John, M. (2008). The red-line of workload: Theory, research, and design. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 52, pages 1204–1208. Sage Publications Sage CA: Los Angeles, CA.

[40] Gwin, J. T., Gramann, K., Makeig, S., and Ferris, D. P. (2010). Removal of movement artifact from high-density EEG recorded during walking and running. *Journal of Neurophysiology*, 103(6):3526–3534.

[41] Hancock, P. and Parasuraman, R. (1992). Human factors and safety in the design of intelligent vehicle-highway systems (IVHS). *Journal of Safety Research*, 23(4):181–198.

[42] Hancock, P. and Verwey, W. (1997). Fatigue, workload and adaptive driver systems. *Accident Analysis & Prevention*, 29(4):495–506.

[43] Hankins, T. C. and Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation Space and Environmental Medicine*, 69(4):360–367.

[44] Harriott, C. E. (2015). *Workload and task performance in human-robot peer-based teams*. PhD thesis, Vanderbilt University.

[45] Harriott, C. E., Zhang, T., and Adams, J. A. (2013). Assessing physical workload for human–robot peer-based teams. *International Journal of Human-Computer Studies*, 71(7-8):821–837.

[46] Harris, D. (2011). *Human Performance on the Flight Deck*. Ashgate Publishing Limited Surrey, U.K.

[47] Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*, pages 139–183.

[48] Heard, J. and Adams, J. A. (2019). A multi-dimensional human workload assessment algorithm for supervisory human-machine interaction. *Journal of Cognitive Engineering and Decision Making*, (In Press).

[49] Heard, J., Harriott, C. E., and Adams, J. A. (2017). A human workload assessment algorithm for collaborative human-machine teams. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 366–371.

[50] Heard, J., Harriott, C. E., and Adams, J. A. (2018). A survey of workload assessment algorithms. *IEEE Transactions on Human-Machine Systems*, PP(99):1–18.

[51] Heard, J., Heald, R., Harriott, C. E., and Adams, J. A. (2019a). A diagnostic human workload assessment algorithm for supervisory and collaborative human-robot teams. *ACM Transactions on Human-Robotic Interaction*, (In Press).

[52] Heard, J., Paris, R. A., Sullivan, P., Scully, D., McNaughton, C., Ehrenfeld, J. M., Coco, J., Fabbri, D., Bodenheimer, B., and Adams, J. A. (2019b). Automatic clinical procedure detection for emergency services. In *IEEE Conference on Engineering in Medicine and Biology)*.

[53] Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2013). Mental workload during n-back task-quantified in the prefrontal cortex using fnirs. *Frontiers in Human Neuroscience*, 7:1–9.

[54] Hirshfield, L. M., Solovey, E., Girouard, A., Kebinger, J., Jacob, R., Sassaroli, A., and Fantini, S. (2009). Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 2185–2194.

[55] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[56] Hogervorst, M. A., Brouwer, A.-M., and van Erp, J. B. F. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in Neuroscience*, 8:213–225.

[57] Hoogendoorn, R. and van Arem, B. (2013). Driver workload classification through neural network modeling using physiological indicators. In *IEEE Conference on Intelligent Transportation Systems*, pages 2268–2273.

[58] Humphrey, C. and Adams, J. A. (2011). Analysis of complex team-based systems:augmentations to goaldirected task analysis and cognitive work analysis. *Theoretical Issues in Ergonomics Science*, 12(2):149–175.

[59] Hussein, A. and Abbass, H. (2018). Mixed initiative systems for human-swarm interaction: Opportunities and challenges. In *2018 2nd Annual Systems Modelling Conference (SMC)*, pages 1–8. IEEE.

[60] Imran, A., Pandharipande, M., and Kopparapu, S. K. (2013). Speakrite: Monitoring speaking rate in real time on a mobile phone. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 5(1):62–69.

[61] International, A. (2010). Standard guide for operational guidelines for initial response to a suspected biothreat agent. Technical Report ASTM E2770-10, American Society for Testing and Materials.

[62] international, S. (2016). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. *SAE International,(J3016)*.

[63] Jameson, A., Kiefer, J., Müller, C., Großmann-Hutter, B., Wittig, F., and Rummer, R. (2010). Assessment of a users time pressure and cognitive load on the basis of features of speech. *Resource-adaptive cognitive processes*, pages 171–204.

[64] Jamison Heard, J. F. and Adams, J. A. (2019). Speech workload estimation for supervisory human-machine teams. In *Annual Proceedings of the Human Factors and Ergonomics Society*, page (In Press).

[65] Jorna, P. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological Psychology*, 34(2-3):237–257.

[66] Jorna, P. (1993). Heart rate and workload variations in actual and simulated flight. *Ergonomics*, 36(9):1043–1054.

[67] Kaber, D. B. and Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, 5(2):113–153.

[68] Keller, J., Bless, H., Blomann, F., and Kleinbohl, D. (2011). Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *Journal of Experimental Social Psychology*, 47(4):849–852.

[69] Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pages 8–13.

[70] Kitaoka, N., Kakutani, N., and Nakagawa, S. (2005). Detection and recognition of correction utterances on misrecognition of spoken dialog system. *Systems and Computers in Japan*, 36(11):24–33.

[71] Klingner, J., Kumar, R., and Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. In *Symposium on Eye Tracking Research and Applications*, pages 69–72.

[72] Lasley, D. J., Hamer, R. D., Dister, R., and Cohn, T. E. (1991). Postural stability and stereo-ambiguity in man-designed visual environments. *IEEE Transactions on Biomedical Engineering*, 38(8):808–813.

[73] Lin, J., Matthews, G., Reinerman-Jones, L., and Wohleber, R. (2018). Assessing operator psychological states and performance in uas operations. In *International Conference on Augmented Cognition*, pages 131–147.

[74] Lively, S. E. (1993). Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *The Journal of the Acoustical Society of America*, 93(5):2962.

[75] Lorenz, B., Nocera, F. D., Rottger, S., and Parasuraman, R. (2003). Automated fault-management in a simulated spaceflight micro-world. *Aviation, Space, and Environmental Medicine*, 73(9):886 – 897.

[76] Manawadu, U. E., Kawano, T., Murata, S., Kamezaki, M., Muramatsu, J., and Sugano, S. (2018). Multiclass classification of driver perceived workload using long short-term memory based recurrent neural network. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–6. IEEE.

[77] Mark, G., Gudith, D., and Klocke, U. (2008). The cost of interrupted work: more speed and stress. In *IGCHI conference on Human Factors in Computing Systems*, pages 107–110.

[78] Marquart, G., Cabrall, C., and de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, 3:2854–2861.

[79] Massari, D., Pacheco, D., Malekshahi, R., Betella, A., Verschure, P., Birbaumer, N., and Caria, A. (2014). Fast mental states decoding in mixed reality. *Frontiers in Behavioral Neuroscience*, 8.

[80] McCraken, J. and Aldrich, T. (1984). Implications of operator workload and system automation goals. Technical Report ASI-479-024-84B, U.S. Army Research Institution.

[81] Meehan, M., Insko, B., Whitton, M., and Brooks Jr, F. P. (2002). Physiological measures of presence in stressful virtual environments. *ACM Transactions on Graphics*, 21(3):645–652.

[82] Mitchell, D. (2000). Mental workload and arl workload modeling tools. Technical Report ARL-TNL-161, Army Research Lab Aberden Proving Ground MD.

[83] Miyake, S., Yamada, S., Shoji, T., Takae, Y., Kuge, N., and Yamamura, T. (2009). Physiological responses to workload change. a test/retest examination. *Applied Ergonomics*, 40(6):987–996.

[84] Mizuno, T., Sakai, T., Kawazura, S., Asano, H., Akehi, K., Matsuno, S., Mito, K., Kume, Y., and Itakura, N. (2016). Measuring facial skin temperature changes caused by mental work-load with infrared thermography. *IEEE Transactions on Electronics, Information and Systems*, 136(11):1581–1585.

[85] Nagasawa, T. and Hagiwara, H. (2016). Workload induces changes in hemodynamics,

respiratory rate and heart rate variability. In *IEEE 16th International Conference on Bioinformatics and Bioengineering*, pages 176–181.

[86] Nassiri, P., Monazam, M., Dehaghi, B. F., Abadi, L., Zakerian, S., and Azam, K. (2013). The effect of noise on human performance: A clinical trial. *Int J Occup Environ Med.*, 4(2):87 – 95.

[87] Niculescu, A., Cao, Y., and Nijholt, A. (2010). Manipulating stress and cognitive load in conversational interactions with a multimodal system for crisis management support. In *Development of Multimodal Interfaces: Active Listening and Synchrony*, pages 134–147. Springer Nature.

[88] Nourbakhsh, N., Wang, Y., Chen, F., and Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *the Australian Computer-Human Interaction Conference*, pages 420–423.

[89] O'Donnell, R. and Eggemeier, F. (1986). In Boff, K., Kaufman, L., and Thomas, J., editors, *Handbook of perception and human performance*, chapter 42. Wiley and Sons New York.

[90] Oh, H., Hatfield, B. D., Jaquess, K. J., Lo, L.-C., Tan, Y. Y., Prevost, M. C., Mohler, J. M., Postlethwaite, H., Rietschel, J. C., Miller, M. W., Blanco, J. A., Chen, S., and Gentili, R. J. (2015). A composite cognitive workload assessment system in pilots under various task demands using ensemble learning. In Schmorrow, D. D. and Fidopiastis, C. M., editors, *Foundations of Augmented Cognition*, pages 91–100. Springer-Verlog Switzerland.

[91] Or, C. K. and Duffy, V. G. (2007). Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. *Occupational Ergonomics*, 7(2):83–94.

[92] Parasuraman, R. (2003). Neuroergonomics: Research and practice. *Theoretical Issues in Ergonomics Science*, 4(1-2):5–20.

[93] Park, H.-W., Khil, A.-R., and Bae, M.-J. (2013). The improvement of mobile phone voice quality by bone-conduction device. In *IEEE International Conference on Consumer Electronics*, pages 397–398.

[94] Paul, P., Kuijer, F. M., Visser, B., and Kemper, H. (1999). Job rotation as a factor in reducing physical workload at a refuse collecting department. *Ergonomics*, 42(9):1167–1178.

[95] Popovic, D., Stikic, M., Rosenthal, T., Klyde, D., and Schnell, T. (2015). Sensitive, diagnostic and multifaceted mental workload classifier (PHYSIOPRINT). In Schmorrow, D. and Fidopiastis, C. M., editors, *Foundations of Augmented Cognition*, pages 101–111. Springer-Verlog Switzerland.

[96] Putze, F., Jarvis, J., and Schultz, T. (2010). Multimodal recognition of cognitive workload for multitasking in the car. In *IEEE Conference on Pattern Recognition*, pages 3748–3751.

[97] Reid, G. B. and Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In Hancock, P. A. and Meshkati, N., editors, *Advances in Psychology*, volume 52, pages 185–218. Elsevier BV.

[98] Reimer, B., Mehler, B., Coughlin, J., Godfrey, K., and Tan, C. (2009). An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 115–118.

[99] Roscoe, A. H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological Psychology*, 34(2-3):259–287.

[100] Ruff, H., Calhoun, G., Frost, E., Behymer, K., and Bartik, J. (2018). Comparison of adaptive, adaptable, and hybrid automation for surveillance task completion in a multi-task environment. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 155–159. SAGE Publications Sage CA: Los Angeles, CA.

[101] Rusnock, C., Borghetti, B., and McQuaid, I. (2015). Objective-analytical measures of workload – The third pillar of workload triangulation? In Schmorrow, D. D. and Fidopiastis, C. M., editors, *Foundations of Augmented Cognition*, pages 124–135. Springer Science Switzerland.

[102] Rusnock, C. F. and Geiger, C. D. (2017). Simulation-based evaluation of adaptive automation revoking strategies on cognitive workload and situation awareness. *IEEE Transactions on Human-Machine Systems*, 47(6):927–938.

[103] Salvucci, D. D., Taatgen, N. A., and Borst, J. P. (2009). Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption. In *SIGCHI conference on human factors in computing systems*, pages 1819–1828. ACM.

[104] Sandholm, T., Larson, K., Andersson, M., Shehory, O., and Tohmé, F. (1999). Coalition structure generation with worst case guarantees. *Artificial Intelligence*, 111(1-2):209–238.

[105] Sassaroli, A., Zheng, F., Hirshfield, L. M., Girouard, A., Solovey, E. T., Jacob, R. J., and Fantini, S. (2008). Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy. *Innovative Optical Health Sciences*, 1(02):227–237.

[106] Schultze-Kraft, M., Dahne, S., Gugler, M., Curio, G., and Blankertz, B. (2016). Unsupervised classification of operator workload from brain signals. *Journal of Neural Engineering*, 13(3):036008.

[107] Schwarz, J. and Fuchs, S. (2018). Validating a" real-time assessment of multidimensional user state"(rasmus) for adaptive human-computer interaction. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 704–709. IEEE.

[108] Sen, S. D. and Adams, J. A. (2013). A decision network based framework for multiagent coalition formation. In *International Conference on Autonomous Agents and Multi-Agent Systems*, pages 55–62. International Foundation for Autonomous Agents and Multiagent Systems.

[109] Sheridan, T. B. (2011). Adaptive automation, level of automation, allocation authority, supervisory control, and adaptive control: distinctions and modes of adaptation. *IEEE Transactions on Systems, Man, and Cybernetics*, 41(4):662–667.

[110] Shi, Y., Ruiz, N., Taib, R., Choi, E., and Chen, F. (2007). Galvanic skin response (GSR) as an index of cognitive load. In *Extended Abstracts on Human Factors in Computing Systems*, pages 2651–2656.

[111] Sridhar, B., Sheth, K., and Grabbe, S. (1998). Airspace complexity and its application in air traffic management. In *2nd USA/Europe Air Traffic Management R&D Seminar*, pages 1–6.

[112] Steele, B. G., Holt, L., Belza, B., Ferris, S., Lakshminaryan, S., and Buchner, D. M. (2000). Quantitating physical activity in COPD using a triaxial accelerometer. *Chest*, 117(5):1359–1367.

[113] Sterman, M. and Mann, C. (1995). Concepts and applications of EEG analysis in aviation performance evaluation. *Biological Psychology*, 40(1-2):115–130.

[114] Szalma, J. L. and Hancock, P. A. (2011). Noise effects on human performance: A meta-analytic synthesis. *Psychological Bulletin*, 137(4):682–707.

[115] Teo, G., Reinerman-Jones, L., Matthews, G., Barber, D., Harris, J., and Hudson, I. (2016). Augmenting robot behaviors using physiological measures of workload state. In *International Conference on Augmented Cognition*, pages 404–415. Springer.

[116] Teo, G., Reinerman-Jones, L., Matthews, G., Szalma, J., Jentsch, F., and Hancock, P. (2018). Enhancing the effectiveness of human-robot teaming with a closed-loop system. *Applied ergonomics*, 67:91–103.

[117] Thropp, J. E., Oron-Gilad, T., Szalma, J. L., and Hancock, P. A. (2018). Calibrating adaptable automation to individuals. *IEEE Transactions on Human-Machine Systems*, (99):1–11.

[118] Ting, C., Mahfouf, M., Nassef, A., Linkens, D., Panoutsos, G., Nickel, P., Roberts, A., and Hockey, G. (2010). Real-time adaptive automation system based on identification of operator functional state in simulated process control operations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(2):251–262.

[119] Ting, C.-H., Mahfouf, M., Linkens, D. A., Nassef, A., Nickel, P., Roberts, A. C., Roberts, M. H., and Hockey, G. J. (2008). Real-time adaptive automation for performance enhancement of operators in a human-machine system. In *Mediterranean Conference on Control and Automation*, pages 552–557.

[120] Veltman, J. A. and Gaillard, A. W. K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, 41(5):656–669.

[121] Vicente, K., Thornton, D., and Moray, N. (1987). Spectral analysis of sinus arrhythmia: A measure of mental effort. *Human Factors*, 29(2):171–182.

[122] Walter, C., Rosenstiel, W., Bogdan, M., Gerjets, P., and Spüler, M. (2017). Online eeg-based workload adaptation of an arithmetic learning environment. *Frontiers in human neuroscience*, 11:286.

[123] Wang, R., Zhang, Y., and Zhang, L. (2016). An adaptive neural network approach for operator functional state prediction using psychophysiological data. *Integrated Computer-Aided Engineering*, 23(1):81–97.

[124] Weinger, M. B., Herndon, O. W., Zornow, M. H., Paulus, M. P., Gaba, D. M., and Dallen, L. T. (1994). An objective methodology for task analysis and workload assessment in anesthesia providers. *Anesthesiology*, 80(1):77–92.

[125] Wickens, C. D. (1992). *Engineering Psychology and Human Performance*. HarperCollins.

[126] Wickens, C. D., Lee, J. D., Liu, Y., and Becker, S. E. G. (2004). *An Introduction to Human Factors Engineering*. Pearson Education, Inc., 2nd edition.

[127] Wierwille, W. W. and Casali, J. G. (1983). A validated rating scale for global mental workload measurement applications. *Human Factors and Ergonomics Society Annual Meeting*, 27(2):129–133.

[128] Wilson, G. and Fisher, F. (1990). The use of multiple physiological measures to determine flight segment in F4 pilots. In *IEEE Conference on Aerospace and Electronics*, pages 859–861.

[129] Wilson, G. F. and Russell, C. A. (2003). Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human Factors*, 45(4):635–643.

[130] Xu, J., Slagle, J. M., Banerjee, A., Bracken, B., and Weinger, M. B. (2019). Use of a portable functional near-infrared spectroscopy (fnirs) system to examine team experience during crisis event management in clinical simulations. *Frontiers in human neuroscience*, 13:85.

[131] Yager, R. R. and Zadeh, L. A. (2012). *An introduction to fuzzy logic applications in intelligent systems*, volume 165. Springer-Verlog New York.

[132] Ye, J., Janardan, R., and Li, Q. (2004). Two-dimensional linear discriminant analysis. In *Advances in neural information processing systems*, pages 1569–1576.

[133] Yen, Y. and Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, 30:111–120.

[134] Yen, Y., Wickens, C. D., and Hart, S. G. (1985). The effect of varying task difficulty on subjective workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 29(8):765–769.

[135] Yin, Z. and Zhang, J. (2014). Recognition of mental workload levels by combining adaptive exponential feature smoothing and locality preservation projection techniques. In *IEEE Chinese Control Conference*, pages 4700–4705.

[136] Young, M. S. and Stanton, N. A. (2002). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human Factors*, 44(3):365–375.

[137] Zhang, H., Zhu, Y., Maniyeri, J., and Guan, C. (2014). Detection of variations in cognitive workload using multi-modality physiological sensors and a large margin unbiased regression machine. In *IEEE Conference on Medicine and Biology*.

[138] Zhang, J. and Wang, R. (2016). Adaptive fuzzy modeling based assessment of operator functional state in complex human–machine systems. In Dimirovski, G. M., editor, *Complex Systems*, pages 189–210. Springer-Verlog Switzerland.

[139] Zhang, J., Yin, Z., and Wang, R. (2015). Recognition of mental workload levels under complex human-machine collaboration by using physiological features and adaptive

support vector machines. *IEEE Transactions on Human-Machine Systems*, 45(2):200–214.

[140] Zhang, J.-H., Peng, X.-D., Liu, H., Raisch, J., and Wang, R.-B. (2013). Classifying human operator functional state based on electrophysiological and performance measures and fuzzy clustering method. *Cognitive Neurodynamics*, 7(6):477–494.

[141] Zhang, L., Wade, J., Bian, D., Fan, J., Swanson, A., Weitlauf, A., Warren, Z., and Sarkar, N. (2017). Cognitive load measurement in a virtual reality-based driving system for autism intervention. *IEEE transactions on affective computing*, 8(2):176–189.

[142] Zhang, P., Wang, X., Zhang, W., and Chen, J. (2019). Learning spatial–spectral–temporal eeg features with recurrent 3d convolutional neural networks for cross-task mental workload assessment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(1):31–42.

[143] Zhao, G., Liu, Y.-J., and Shi, Y. (2018). Real-time assessment of the cross-task mental workload using physiological measures during anomaly detection. *IEEE Transactions on Human-Machine Systems*, 48(2):149–160.

[144] Zhou, J., Luo, H., Luo, Q., and Shen, L. (2009). Attentiveness detection using continuous restricted boltzmann machine in e-learning environment. In *International Conference on Hybrid Learning and Education*, pages 24–34. Springer.

Appendix A

Cross-Interaction Paradigm Analysis

The peer-based evaluation included human-robot and human-human teaming scenarios, which allows for analyzing the impact that training with additional workload data from the human-human teaming interaction paradigm has. The IMPRINT Pro model results indicate that the human-human teaming tended to have lower workload then the human-robot scenario, but a Kruskal-Wallis test found no significant differences. This analysis focuses solely on the peer-based relationship, as the supervisory-based evaluation did not include human-human teams. Three algorithms were trained: only human-human data (HH), both the human-human and human-robot data (HH-HR), and only the human-robot data (PEER), the same dataset as in Section 3.8. Each algorithm was trained using data from twelve peer-evaluation participants, while data from six peer-evaluation H-R participants were used for testing. The workload classification thresholds are provided in Table 3.36.

It is expected the the human-human teaming data will minimally impact the algorithm's classification accuracy for human-robot teams; thus, hypothesis $\mathbf{H_6}$ predicts that the HH-HR algorithm's classification accuracy will be within 5% of the PEER algorithm's accuracy. The hypothesis also predicts that the HH algorithm's classification accuracy will be within 5% of the PEER algorithm's accuracy, as there is no significant difference between the IMPRINT Pro models for each teaming scenario. Further, it is expected that the HH-HR and HH algorithms will be able to track workload shifts. Hypothesis $\mathbf{H_7}$ predicts that both algorithm's workload estimates will significantly and positively correlate with the workload models for each peer-based task.

A.1 Cross-Interaction Paradigm Results

The workload assessment algorithm's cross-interaction paradigm performance is examined by comparing the algorithms' estimates to the IMPRINT Pro modeled workload values for the peer-based human-robot teaming tasks. The algorithms' estimates are provided in Table A.1. The HH-HR algorithm's estimates are within a standard deviation of the IMPRINT Pro workload model values for each workload component. The HH algorithm's physical and overall workload estimates are within a standard deviation of the IMPRINT Pro values, but the algorithm's workload estimates for high cognitive workload are not. The HH and HH-HR algorithms tend to overestimate high physical workload and underestimate low workload, but do so minimally. The Kruskal-Wallis test found that each algorithms' estimates significantly differed between workload conditions. All three algorithm's estimates are within 10% of each other, illustrating that the respective trained algorithm produces accurate workload estimates.

Table A.1: Descriptive and Kruskall-Wallis Statistics for the IMPRINT Pro Workload Model Values and the HH, HH-HR, and PEER Algorithms' Workload Estimates for the Peer-Based Evaluation.

Workload	**Training**	**Workload Condition**		
		Low	High	χ^2
Cognitive	Model	3.58 (2.84)	6.14 (1.59)	61.75*
	HH	3.49 (2.48)	8.68 (1.22)	44.02*
	HH-HR	3.45 (2.42)	5.50 (1.14)	59.15*
	PEER	3.77 (2.82)	5.95 (1.40)	35.92*
Physical	Model	4.17 (2.63)	5.56 (2.14)	17.85*
	HH	4.00 (2.42)	6.02 (1.88)	17.51*
	HH-HR	4.07 (6.02)	6.02 (1.82)	15.00*
	PEER	4.28 (2.50)	5.35 (1.89)	10.94*
Overall	Model	14.33 (7.83)	21.31 (3.36)	56.26*
	HH	13.63 (7.51)	20.97 (2.41)	72.25*
	HH-HR	13.65 (7.51)	20.78 (2.51)	62.34*
	PEER	14.64 (7.62)	20.91 (2.25)	45.60*

Each algorithm must accurately classify workload for the peer-based evaluation's H-R teaming scenario. The algorithms' classification accuracies are presented in Table A.2. The

HH-HR algorithm achieves high classification accuracy for overall workload and its components. However, the HH-HR and the HH algorithms' physical workload classification accuracy is approximately four percent lower than the PEER algorithm's accuracy, which demonstrates that the H-H teaming data negatively impacts the HH-HR algorithm's classification accuracy. The HH-HR algorithm's cognitive and overall classification accuracy are within two percent of the PEER algorithm's accuracy.

Table A.2: HH, HH-HR, and PEER Trained Algorithms' Classification Accuracy.

Workload	**Training**	**Workload Condition**	
		Low	High
Cognitive	HH	94.92	94.63
	HH-HR	**97.14**	**96.23**
	PEER	97.05	94.52
Physical	HH	88.25	86.55
	HH-HR	87.30	86.02
	PEER	**90.50**	**90.35**
Overall	HH	92.69	**96.77**
	HH-HR	92.40	95.70
	PEER	**94.78**	96.55

Bold represents highest accuracy per column

The Pearson's correlation coefficients between the algorithms' estimates and IMPRINT Pro models are used to analyze the algorithms' ability to track workload shifts within and across workload conditions. The correlation coefficients are presented in Table A.3. The algorithm's estimates significantly correlate with the IMPRINT Pro workload models, demonstrating that each algorithm tracks workload shifts across and within workload conditions. The correlation coefficients for each algorithm are similar to each other; thus, illustrating that the H-H teaming data does not negatively impact the algorithm's ability to track workload shifts.

A.2 Cross-Interaction Paradigm Discussion

Incorporating human-human teaming data can increase the amount of available training data, but it must have a minimal impact on an adaptive teaming system's ability to

Table A.3: HH, HH-HR, and PEER Algorithms' Correlation Coefficients for Within and Across Workload Conditions.

Workload	**Training**	**Within**		**Across**
		Low	High	
Cognitive	HH	0.93*	0.81*	0.93*
	HH-HR	0.95*	0.83*	0.90*
	PEER	0.95*	0.85*	0.94*
Physical	HH	0.87*	0.91*	0.91*
	HH-HR	0.87*	0.90*	0.94*
	PEER	0.91*	0.92*	0.92*
Overall	HH	0.97*	0.91*	0.97*
	HH-HR	0.97*	0.91*	0.98*
	PEER	0.97*	0.91*	0.97*

classify workload for human-robot teaming scenarios. Hypothesis H_6 focuses on such impacts to the algorithm's classification accuracy, by stating that the HH-HR and HH trained algorithms' accuracy will be within 5% of the PEER trained algorithm's accuracy. The hypothesis is fully supported, which demonstrates that the human-human teaming data does not significantly impact the algorithm's classification accuracy for human-robot teaming paradigms. Further, the algorithm trained only on human-human teaming data still achieved high performance for human-robot teaming scenarios, as the two scenarios have similar workload levels.

The HH-HR trained algorithm also needs to track workload shifts within and across workload conditions. Hypothesis H_7 states that the HH and HH-HR trained algorithms' estimates will significantly correlate with the IMPRINT Pro workload models. The hypothesis is supported, which illustrates that incorporating human-human teaming data did not impact the algorithm's ability to track workload shifts within and across workload conditions. Similarly, an algorithm trained solely on human-human teaming data tracked workload shifts for a human-robot teaming scenario. Overall, the workload assessment algorithm trained on human-robot and human-human teaming data sets did not substantially decrease performance. Further, an algorithm trained on a human-human teaming scenario achieved high performance in a similar human-robot scenario.

Appendix B

Window Size Impact

The window size impact analysis investigated the affect window size has on algorithm performance. A 30 second window was used for all results presented in the previous sections; however, workload assessment algorithms use a variety of window sizes, (e.g., [2, 25, 32]). 1, 5, 15, and 60 second window sizes were analyzed, as they are the most common in the literature. It was expected that performance will increase as window size increases, but there will also be a point of diminishing returns. Algorithm performance was determined using classification accuracy across the three prior analyses: workload generalizability, population generalizability, and emulated real-world conditions. The same methods from the previous sections were used to divide the training and testing sets. Hypothesis H_6 predicted that the 30 second window size will achieve the highest classification accuracy for each workload component and condition.

Classification accuracy was used to analyze the impact of window size on algorithm performance. The classification accuracies for *workload generalizability* by workload component, condition, and window size are provided in Table B.1, where bolded values represent the highest accuracy. Multiple values are bolded if the values were within two percent of one another. The 30 second window achieved the highest accuracy for cognitive, physical, and overall workload for each condition, while the 60 second window achieved similar performance for cognitive and overall workload. There was a general increase in classification accuracy as window size increased for cognitive workload. The 5 second window achieved the highest auditory workload classification accuracy. Smaller window sizes tended to achieve higher auditory workload classification accuracy. The overall workload classification accuracies were similar to each other. The highest physical workload accuracies occurred when the window size was 30 seconds and the 1 second window size

Table B.1: Window Size Impact: Algorithm's Classification Accuracy (%) by Workload Component and Condition for Workload Generalizability.

Workload	Window Size	Underload	Normal Load	Overload
Cognitive	1	84.32	76.91	85.95
	5	93.90	84.64	89.43
	15	98.78	49.20	80.65
	30	**100**	**99.28**	**100**
	60	**100**	**100**	83.06
Auditory	1	98.11	94.51	91.57
	5	**99.39**	**97.40**	93.22
	15	98.33	95.01	95.94
	30	91.57	90.02	92.01
	60	76.44	77.56	**99.98**
Physical	1	98.54	82.25	95.19
	5	97.51	73.87	47.47
	15	99.23	**87.08**	55.01
	30	**99.56**	**86.67**	**100**
	60	**100**	75.55	96.29
Overall	1	85.70	99.25	97.45
	5	93.47	97.60	91.53
	15	98.81	99.25	97.53
	30	**100**	**99.83**	**100**
	60	**100**	**100**	97.19

Note: Bold represents highest accuracy per workload component.

achieved higher physical workload accuracy than the 5 and 15 second window sizes. There was a large decrease in physical workload accuracy from the 1 to 5 second window size algorithms in the overload condition.

Population generalizability is important, given that it is impractical to train an algorithm on each human team member. The workload assessment algorithm's classification accuracies for population generalizability by workload component, condition, and window size are provided in Table B.2. The 30 second window achieved the highest cognitive and overall workload classification accuracy for each workload condition, while the 60 second window achieved similar results. The 30 and 60 second windows achieved the highest physical workload classification accuracies, while the 1 second window achieved similar physical workload accuracies for the underload condition. The 1 and 5 second windows achieved the highest auditory workload classification for the underload and normal load conditions, while the 60 second window achieved the highest accuracy for the overload condition. There was an increase in cognitive workload accuracy as the window size increased, while there was a decrease in auditory workload accuracy as window sized increased. There was no discernible trend between window size and physical workload accuracy. Each window size achieved similar overall workload classification accuracies; although, there was an increase in classification accuracy for the underload condition.

Lastly, it is important to analyze the affect window size has under *emulated real-world conditions*. The classification accuracies by workload component, condition, and window size for emulated real-world conditions are provided in Table B.3. The 30 second window achieved the highest classification accuracy for cognitive and overall workload for the underload and normal load conditions, while the 60 second window achieved the highest accuracy for the overload condition. The highest physical workload classification accuracies were achieved with the 30 second window. Smaller window sizes achieved higher auditory workload classification accuracies. There was an increase in cognitive and overall workload classification accuracy for the underload condition, as the window size increased, while auditory workload classification accuracy decreased.

B.1 Window Size Impact Discussion

Examining the affect window size on algorithm performance allows adaptive workload system designers to determine how much workload metric data is needed to obtain a desired performance level. Hypothesis H_5 stated that a 30 second window size will produce the highest classification accuracies for workload generalizability, population generalizability, and the emulated real-world conditions. The hypothesis was supported for cognitive and

Table B.2: Window Size Impact: Algorithm's Classification Accuracy (%) by Workload Component and Condition for Population Generalizability.

Workload	Window Size	Underload	Normal Load	Overload
Cognitive	1	88.93	80.08	88.88
	5	94.42	81.86	91.84
	15	98.26	76.48	88.18
	30	**99.42**	**97.74**	**99.53**
	60	**99.73**	**97.48**	90.02
Auditory	1	**99.53**	**98.50**	97.59
	5	**99.86**	**98.56**	95.60
	15	98.61	95.39	96.19
	30	93.94	91.50	92.01
	60	87.94	75.26	**99.99**
Physical	1	**99.00**	82.64	96.46
	5	97.73	81.31	81.69
	15	95.56	81.85	86.04
	30	**98.83**	81.20	**100**
	60	**99.27**	**87.05**	86.01
Overall	1	89.25	98.48	98.31
	5	93.72	98.44	97.11
	15	97.56	97.63	97.97
	30	**99.79**	**99.92**	**99.73**
	60	**99.73**	**99.83**	97.55

Note: Bold represents highest accuracy per workload component.

overall workload classification for each algorithm analysis and was partially supported for physical workload, as the 30 second window size did not achieve the highest physical workload classification for the population generalizability analysis. However, the physical workload accuracy was above 80%. The hypothesis is not supported for auditory workload across each analysis, as smaller window sizes tended to achieve higher accuracy, which is due to relying solely on noise-level for the auditory workload estimation. A task's auditory demands can change rapidly, which produces immediate changes in noise-level. Larger window sizes did not capture these rapid changes.

Table B.3: Window Size Impact: Algorithm's Classification Accuracy (%) by Workload Component and Condition for Emulated Real-World Conditions.

Workload	Window Size	Underload	Normal Load	Overload
Cognitive	1	88.65	76.24	97.93
	5	88.82	55.76	98.76
	15	**98.24**	86.42	57.35
	30	**98.57**	**91.95**	88.80
	60	**98.05**	83.27	**99.40**
Auditory	1	97.33	96.19	**98.84**
	5	**98.33**	**98.53**	95.34
	15	**98.57**	94.43	84.93
	30	90.76	84.14	88.37
	60	81.10	63.22	78.43
Physical	1	97.59	72.53	98.05
	5	95.27	77.38	75.83
	15	89.93	71.18	92.37
	30	**99.08**	**79.94**	**99.60**
	60	98.04	59.95	96.42
Overall	1	87.28	**96.85**	**99.28**
	5	88.18	**96.09**	**99.65**
	15	95.54	**96.93**	76.92
	30	**98.52**	**96.66**	97.58
	60	96.71	91.22	**99.80**

Note: Bold represents highest accuracy per workload component.