2020

An Approach to Modeling Simulated Military Human-agent Teaming

Maartje Hidalgo
University of Central Florida

AN APPROACH TO MODELING SIMULATED
MILITARY HUMAN-AGENT TEAMING

by

MAARTJE HIDALGO, M.S.
Leiden University, 2007

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Industrial Engineering and Management Systems
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2020

Major Professor: Waldemar Karwowski

ABSTRACT

With the rise of human-agent teaming (HAT), a new cycle of scientific discovery commenced. Through scientific discovery, a number of theories of constructs in HAT were developed, however, an overarching model is lacking that elucidates the relative importance of these constructs in relation to human performance.

The main objective of this research was to develop a model of simulated military HAT and to validate it against selected empirical data. Experimental data borrowed from four simulated military HAT studies were used to test the proposed Core model. The Core model was assumed to be directly affecting task performance and consisted of constructs related to Task Composition, Task Perception, and the qualities that each team member (Human/Agent Qualities) brings to the team. The available experimental data were tested against the null model: everything, within and between these Core sections, are equal contributors to hit rate.

Furthermore, in order to validate the Core model, a validation approach was developed based on relative importance, wherein the outcome was a proportional value and followed a beta distribution (Ferrari & Cribari-Neto, 2004). This new modeling approach consisted of (1) application of dominance analysis (DA; Azen & Budescu, 2003; Budescu, 1993) to determine the most important contributors to task performance, (2) establishing robustness and generalizability of the dominance outcome through bootstrap procedures (Azen & Budescu, 2003; Efron, 1981), and (3) combining the dominant predictors into a full beta regression model to evaluate the fit and significance of the model (Ferrari & Cribari-Neto, 2004).

DA of all four experimental studies examined in this research led to rejecting the null hypotheses. Constructs in the proposed Core model were not equally important to performance in these simulated military HAT studies. Results showed consistently similar yet different dominance patterns in relation to human performance. Attempts were made to elucidate the most important predictors of task performance. Analyses unveiled the importance of taking task difficulty into consideration when assessing the relative importance within the proposed Core model.

This work is dedicated to my daughter Lilly. May this show you that anything is possible with discipline, determination, and the support of others.

We are naturally gifted with curiosity and a strong-willed nature: use it wisely.

ACKNOWLEDGMENTS

In this venture, on many occasions, the support of others is what got me through the growing pains of a dissertation. I could not have done this without the support of my husband and family. Thank you for filling the gap when I could not and for your patience with me and this process. In addition, Valarie Yerdon, thank you for your friendship, faith, and patience with me throughout this venture. More importantly, your willingness to introduce me to Dr. Peter Hancock incited this work.

Dr. Peter Hancock, thank you for giving me a chance and offering me a position as research assistant, allowing me to journey into a doctoral degree. It has been an honor to have you on my committee.

Dr. Grace Teo, thank you for teaching me invaluable skills early in my research career and for always offering support where needed.

Dr. Lauren Reinerman-Jones, thank you for your advisement, mentoring, and philosophical discussions that helped me shape my work and personal growth.

Dr. Ben Sawyer, thank you for your strategic advisement and keen guidance, allowing me to grow as a professional. You once said, “you cannot reward anyone more than with your time”; your time has been priceless to me.

Dr. Gerald Matthews, your expertise regarding modeling data has encouraged me to solve the problems you keenly saw early on. Thank you for your wisdom.

Dr. Daniel Barber, thank you for your patience with my endless questions regarding the data and giving me full reign.

Dr. Waldemar Karwowski, your time and willingness to support a student has made the impossible possible for me.

Dr. Paul Wiegand, thank you making the time to review parts of the work, just to support a student in need.

Keith MacArthur, your persistent willingness to teach me in R made this complicated puzzle solvable. Thank you for believing in my ability.

Rhyse Bendell, thank you for offering a listening ear and academic insight where needed.

And last but not least, a special thanks to Claudio Bustos, Department of Psychiatry and Mental Health, Faculty of Medicine, Universidad de Concepción, Chile. Your curiosity and eagerness to incorporate my findings into the dominanceanalysis package in R has made my work a tangible contribution to science.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS (or) ABBREVIATIONS

DA	Dominance Analysis
DARPA	Defense Advanced Research Projects Agency
HAT	Human-Agent Teaming
IRB	Institutional Review Board
MABA-MABA	Men Are Better At - Machines Are Better At
MMI	Multimodal Interface
RCTA	Robotics Collaborative Technology Alliance
SA	Situation Awareness
SAT	Situation Awareness-Based Agent Transparency model
ToM	Theory of Mind

CHAPTER ONE: THE RISE OF HUMAN-AGENT TEAMING

Standing at the verge of the fourth Industrial Revolution, technology is no longer a mere external tool; the lines between humans and technology will gradually blur (Davis, 2016; Schwab & Davis, 2018). Indeed, automation has integrated into most areas of human lives. Life without smart phones is unthinkable, smart homes emerge rapidly, and the majority of jobs rely on forms of intelligent systems. Such systems possess knowledge, can learn over time, have decision-making qualities, and can act upon the environment (Russell & Norvig, 2009). These intelligent systems are also called agents. Agents are either embodied or disembodied (Bradshaw et al., 2012; Fong et al., 2003; Sukthankar et al., 2012; Wiltshire et al., 2013). Advising software programs (e.g., Grammarly, 2019) are disembodied agents. Embodied agents can be physically present, such as robots, or virtually present, e.g., working remotely with an embodied agent. The present effort focuses on these physically embodied, intelligent systems and are referred to as agents.

This surge in agent development is reflected in the realm of science. Numerous systematic literature reviews have documented the incremental rise in agent-related research (Anjomshoae et al., 2019; Góngora Alonso et al., 2018; Mostafa et al., 2019; Pan et al., 2016; Savela et al., 2018). Moreover, in 2000, U.S. Congress passed a bill that required one-third of the aerial attack force to be unmanned and autonomous by 2010, and one-third of all ground combat vehicles to be unmanned by 2015 (Springer, 2013). This mandate incited new research and development toward transforming agents from tools to teammates at the squad level (Childers et al., 2016). Indeed, in 2012, the combined American military force actively used over 20,000 autonomous unmanned vehicles in the field (Singer, 2012). To meet the needs posed by the military demand, the U.S. Army funded a collaborative effort between industry, academia, and the military to progress agents from tools to teammates: the Robotics Collaborative Technology Alliance (RCTA; Childers et al., 2016). The RCTA also signified the need for scientific knowledge and theory development, as little was known about this new phenomenon of human-agent teaming (HAT). To understand and predict the performance of teams of a combination of humans and agents (human-agent teams), an overarching theoretical model is needed. The present study aims to develop such a model and introduces a validation approach to falsify the model.

The Emergence of Human-Agent Teaming

With the emergence of a new phenomenon in the natural world, researchers attempt to form theoretical models to understand the phenomenon. Development of a theoretical model for HAT begins with an assessment of the literature for a) vergence of definitions of core concepts, and b) the presence of validated theoretical models or theories. Some of the core components that require definitions are the notion of human-agent teaming and the operationalization of an agent in that context.

Defining Constructs of Human-Agent Teaming

Defining Human-Agent Teaming (HAT)

Human-agent teams are formed by one or more humans and intelligent agents that collaborate in a joint activity with a shared goal in mind (Barnes & Evans, 2010; Cuevas et al., 2007; Hoffman & Breazeal, 2004; Ososky et al., 2012; Rahimi & Hancock, 1986). It naturally follows that HAT is teamwork within a human-agent team. The essence of any teaming effort lies in collaboration, which signifies the committal activity of “working jointly with others or together in an intellectual endeavor” (“Merriam-Webster,” 2019). Collaboration is not merely a joint activity or working on a mutual goal. Collaborative behavior is intelligent in nature, where the intentions of others are weighed in the overall commitment to the joint goal, providing mutual support where needed (Grosz, 1996). These teaming requirements dictate the qualities of an agent in HAT, aside from intelligence and embodiment.

Agent Qualities

The most primitive foundation of an agent lies in automation. Automation is the process or task executed by a technology without the human’s intervention (Parasuraman & Riley, 1997). As autonomy or self-government of agents increased over time, the definition of automation was expanded in terms of agent requirements. Automation requires sensing qualities, data processing, and decision-making skills, psychomotor actors, and communication qualities (Sheridan & Parasuraman, 2005). The fluid transition of automation toward autonomy led to those terms frequently used interchangeably in the literature (e.g., Parasuraman, Sheridan, & Wickens, 2000). However, these concepts are distinct (Kaber, 2017). This distinction is important to address as it relates to agents' functionality in a team.

Sheridan and Verplank (1978) set forth a continuum of the degree of automation in support of the human, as presented in **Table *1***. The verbiage in this table is derived directly from their original work. In their description, the computer or agent gains decisive authority as the level of automation increases, thus implying the automation grows progressively more autonomous.

Table 1

Sheridan and Verplank's (1978) levels of automation.

Level of Automation	Description of Interaction
1	Human does the whole job up to the point of turning it over to the computer to implement.
2	Computer helps by determining the options.
3	Computer helps determine options and suggests one which human need not follow.
4	Computer selects action and human may or may not do it.
5	Computer selects action and implements if it human approves.
6	Computer selects action, informs human in plenty of time to stop it.
7	Computer does whole job and necessarily tells human what it did.
8	Computer does whole job and tells human what it did only if human explicitly asks.
9	Computer does whole job and tells human what it did and it, the computer, decides he should be told.
10	Computer does whole job if it decides it should be done, and if so tells human, if it decides he should be told.

Function Allocation

The utility of agents appears beneficial, but the benefit of pairing agents with humans is only as good as the complementary combined qualities that each brings to the team. The afforded qualities of the agent depend on the functions allocated to the agent (Fitts, 1951), which can be static or dynamic (Morris & Rouse, 1986; Rouse, 1994; Scerbo, 2007). The notion of function allocation stemmed from the 1950s when Paul Fitts and his colleagues proposed what functions should be allocated to machines (or agents) and humans in air navigation and air traffic control (Fitts, 1951). They posited that humans and machines are comparable information processing systems. The famous acronym MABA-MABA, Men Are Better At - Machines Are Better At, indicates that humans and machines have distinct strengths as information processors (**Table 2**).

Table 2

Fitts' list.

Men are better at	Machines are better at
Ability to detect small amount of visual or acoustic energy	Ability to respond quickly to control signals and to apply great force smoothly and precisely
Ability to perceive patterns of light or sound	Ability to perform repetitive, routine tasks
Ability to improvise and use flexible procedures	Ability to store information briefly and then to erase it completely
Ability to store very lare amounts of information for long periods and to recall relevant facts at the appropriate time	Ability to reason deductively, including computational ability
Ability to reason inductively	Ability to handle highly complex operations, i.e. to do many different things at once
Ability to exercise judgment	

*Note. A*dapted from Fitts (1951).

The driving principle is that functions in which machines are better should be automated. This work is valuable in capturing "the most important regularity of automation" (de Winter & Dodou, 2014, p.1), but has been criticized for its notion of comparability rather than complementarity to humans (Hancock, 2009; Jordan, 1963), the absence of the strength of human affect (Hancock, 2009), and limited application to static function allocation (Hancock, 2009). One thing to note is that while there is an area of work dedicated to affective robotics, given the nature of ruggedized work for military, search and rescue, and otherwise similar domains, anthropomorphic characterizations will not be a central focus in the present effort.

In a complex and dynamic environment, such as the battlefield, the functions an agent needs to execute should vary based on situational demand and task type, as no function allocation is optimal for all types of operations and situations (Feigh & Pritchett, 2014; Reinerman-Jones et al., 2017; Reinerman-Jones et al., 2011; Ross et al., 2008; Taylor et al., 2013). Therefore, dynamic function allocation is more appropriate for HAT.

Traditionally, dynamic function allocation was classified as either adaptive or adaptable (Rouse, 1994). In adaptive allocation, the intelligent system initiates changes in function assignment based on operator state and situational demand, while humans take this initiative in adaptable systems (Rouse, 1994; Scerbo, 2007). Thus, in these systems, the initiator is fixed. However, dynamic and complex environments require the partakers to fluidly adjust to changing environments to work most effectively as a team. This necessitates a dynamic adjustment of the initiator in the collaboration, also called mixed-initiative interaction (Allen et al., 1999).

Mixed-initiative interaction allows team members to flexibly interleave their initiative, control, and decision-making based on their strengths (Allen et al., 1999; Barnes et al., 2017; Jiang & Arkin, 2015), which is especially important in dynamic and complex environments (Jiang & Arkin, 2015). Embodied agents have been deployed to highly dangerous environments, such as disaster sites, to save and protect human lives. However, often, these agents were not successful due to mobility, communication, and perceptual limitations that required human intervention. When both human and agent are equipped with initiative and self-governance qualities, they will be more capable of effective teamwork.

To this point, it is now clear what the basic foundation of an agent is and is not, and what functions or tasks agents are better at than humans. However, in the recent decade, research of HAT focused increasingly on other aspects of teaming, such as shared understanding (Cooke, 2015; Cooke et al., 2013; Cuevas et al., 2007; Mathieu et al., 2000; Ososky et al., 2012), trust (Billings et al., 2012; Guznov et al., 2015; Hancock et al., 2011; Hanna & Richards, 2018; Sanders et al., 2014; Schaefer et al., 2019) and intent (Breazeal & Aryananda, 2002; Schaefer et al., 2017), while expanding agent communication possibilities through natural language (Chandarana et al., 2017; Harris & Barber, 2014) and multimodal communication (Baber et al., 2011; Barber, 2018; Barber, et al., 2015; Reinerman-Jones et al., 2017)

Agent Intent

To work as a member of a team, that is in part comprised of humans, it is important that the human teammate understands the agent's reasoning for its actions and interprets the agent's actions as beneficial to the teamwork (Schaefer et al., 2017). As such, the concept of agent intent is intertwined with transparency, or *what* the agent communicates (Chen et al., 2018; Lyons et al., 2017), also known as explainable agency (Anjomshoae et al., 2019; Langley et al., 2017). The quest of identifying the best means of communicating such intent has been based on research of human-human teaming (Breazeal, 2004; DeChurch & Mesmer-Magnus, 2010; Demir et al., 2016; Mesmer-Magnus et al., 2017; Scholtz, 2003). The ability to infer or reason about others' minds depends on detecting eye contact, recognizing what others are looking at, pointing behaviors to direct and share attention, and understanding that others may have different beliefs than our own (Lyons & Havig, 2014; Scassellati, 2001). Thus, inferences about the agent's intent have a basis in communication (Schaefer et al., 2017).

Embodied agents can be programmed with algorithms to infer and reason about their human counterpart's beliefs, desires, and state (Abich et al., 2013; Bainbridge et al., 2008; Barnes et al., 2019; Breazeal et al., 2016; Breazeal et al., 2010; Reinerman-Jones et al., 2011; Taylor et al., 2013). With these algorithms, agents are capable of learning from social signals, inferring intent of their teammate, and communicating without using explicit vocabulary (Barnes et al., 2019; Mutlu et al., 2009; Mutlu et al., 2016; Scassellati, 2002). Moreover, these social-cognitive behaviors have shown to enhance the sense of presence in HAT (Fiore et al., 2013). Without this sense of presence, humans could miss the foundation of perceiving the agent as a teammate (Bainbridge et al., 2008).

Communication

Aside from the importance of communication in intent inference, agents also need communication qualities in order to function as an equal peer in a team in terms of sharing information. In natural human form, communication occurs through verbal and nonverbal means (Berlo, 1960; Mehrabian, 1979). As such, agent teammates need the capability of both perceiving and interpreting verbal communication, as well as producing grammatically correct and meaningful language, to be able to interface with humans (Russell & Norvig, 2009). Currently, agents are equipped with technologies to detect and process verbal input through speech detection and natural language processing algorithms, and with technologies to allow them to express simple lexicons (Breazeal & Aryananda, 2002; Childers et al., 2016; Harris & Barber, 2014). However, agents also require the capability to express and process nonverbal communication, as humans convey messages through nonverbal elements as well (Mehrabian, 1979), even more so in operations wherein verbal communication is limited. Non-verbal agent-to-human communication can occur through visual and/or tactile form (Lackey et al., 2011), or through multiple modalities (Barber et al., 2016; Oviatt, 2012).

Scientific Discovery of Human-Agent Teaming

In scientific discovery in new and emerging fields, theories are created based on well-validated theories from relevant research. Hypotheses are generated from related fields and tested against empirical data. For instance, HAT involves teamwork or teaming, albeit with different entities than human teamwork. The diagram in **Figure 1** breaks down the notion of gleaning from related fields to further the science in an emerging research area. Here, there is a general domain of teaming, wherein human teaming and HAT are distinct sub-domains. HAT can be informed by validated theories in the subdomain of human teaming. Each subdomain is formed by theories, that contain components on which theories and models exist.

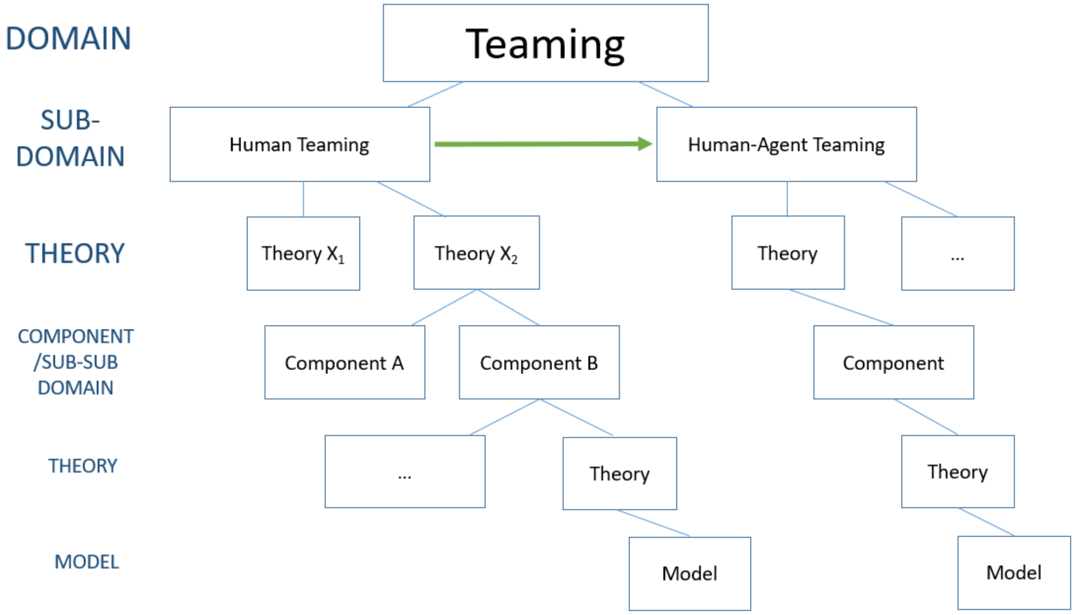


Figure 1. Diagram of scientific discovery by gleaning from related fields.

Note. This diagram visually explains the process of gleaning from related research from other subdomains (here, human teaming) to informing newly emerging phenomena in the natural world (human-agent teaming). Each subdomain is formed by overarching theories, that contain components on which theories and models exist.

Indeed, human teaming can inform HAT (Keebler et al., 2012; Wiltshire et al., 2013), as agents are designed around the human's needs and means of information processing (Bradshaw et al., 2004; Hancock, 2017). Several theories of human teaming exist (e.g., Driskell et al., 2018; Salas et al., 2005). Theories are a set of abstract structures, or models, that provide descriptive statements and/or representations of the phenomenon that aid in their understanding (Bailer-Jones, 2003; Cartwright, 1983; Giere, 1988; van Fraassen, 1987). One of the most comprehensive theories of human teamwork is developed by Salas, Sims, & Burke (2005), wherein they identify five core components of effective teamwork and three coordinating components that support the core components. This theory postulates that the core aspects of teamwork are leadership, mutual performance monitoring, back-up behavior, adaptability, and team orientation. The coordinating factors are shared mental models, mutual trust, and closed-loop communication; all necessary ingredients for effective teamwork. These constructs could be extended and empirically tested for its application to HAT.

Each of the components of the theory (see **Figure 1**), e.g., trust, mental model, closed-loop communication, are supported by theories and models. Models are descriptive statements and/or representations of a phenomenon, that are guided by theory, analogues to aspects of the observable world, and aid in understanding these phenomena (Bailer-Jones, 2003; Cartwright, 1983). One of these models, for example, suggests that closing the loop in communication (i.e., bidirectional communication) is effective (Barnlund, 1979; Schramm, 1954).

Thus far, there are no corroborated theories that apply to the sub-domain of HAT. However, a number of theoretical models have been developed for distinct components or constructs that are important in HAT.

Existing Theoretical Models for Components of HAT

Situation Awareness

Endsley (1995) developed a theoretical model of situation awareness (SA) that has been applied to many forms of human-automation interaction, some of which may be considered a form of teaming. SA refers to the ability of individuals to maintain updated knowledge of the state of a dynamic tasking environment (Endsley, 1988, 1995). The definition is a tripartite conceptualization: "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1995, p. 36). The first portion refers to Level 1 SA, the perception of elements in the environment. Level 2 SA reflects on a deeper understanding of the meaning and significance of the observed factors. Lastly, the projection to or prediction of the status in the near future is Level 3 SA. All levels of SA require both attentional and working memory processing, which can be deteriorated under highly loaded dynamic circumstances.

SA does not merely exist within individuals; SA can exist in teams (Endsley & Jones, 2001). Endsley (1995) posits that in a team formation, each individual should maintain SA for their own requirements, which can overlap partially, or be shared with, with others' SA.

Transparency

Within the subdomain of HAT, a model of transparency was created for disembodied agents by Lyons and colleagues (Lyons, 2013; Lyons et al., 2017; Lyons & Havig, 2014). For physically embodied agents, which is most relevant to the present effort, Chen and colleagues (2014, 2018) developed a situation awareness-based agent transparency (SAT) model to describe the information that both teammates need to convey about their decision-making process. Here, transparency was defined as "the descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process" (Chen et al., 2014, p. 2). The original SAT model emphasized the level and type of information that the agent should communicate, as depicted in **Table 3**.

Table 3

Situation awareness-based agent transparency model.

SA Level	SAT Category	Description
Level 1: Goals & Actions	Agent's current status/actions/plans	• Purpose: Desire (goal selection) • Process: Intentions (planning/execution); Progress • Performance • Perception (environment/teammates)
Level 2: Reasoning	Agent's reasoning process	• Reasoning process (belief/purpose) • Motivations, environmental and other constraints/affordances
Level 3: Projections	Agent's projections/predictions; uncertainty	• Projection of future outcomes • Uncertainty and potential limitations; likelihood of success/failure • History of performance

Note. Adapted from (Chen et al., 2014).

Later, they emphasized the importance of bidirectional transparency, hence, the components of **Table 3** are extended to the human as well (Chen et al., 2018). The levels of SA in the SAT-model refers to a higher level of information that is shared: the current status/action/plans (Level 1 SA), reasoning processing (Level 2 SA), and projections/predictions and level of uncertainty (Level 3 SA). Indeed, research shows that agent transparency through Level 3 SA leads to higher human SA and trust in the agent, compared to lower levels of SA transparency (Selkowitz et al., 2017), as well as improved human-agent team performance (Mercado et al., 2016).

Trust

Even though trust in automation is a difficult construct to define (Schaefer et al., 2019), the most accepted definition is “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” (Lee & See, 2004, p. 54). For HAT in a complex environment, trust refers to the attitude that an agent will help achieve the team’s goals, rather than the individual’s goal (Hancock et al., 2011). Herein, trust is the guiding mechanism for reliance on the agent (Lee & See, 2004). Trust needs to be adequately calibrated, as both overreliance and underreliance on the agent can lead to critical failures (de Visser et al., 2019; Parasuraman & Riley, 1997). The question of what determines human trust in an agent was answered by a meta-analytic study that reviewed human-related, agent-related, and environmental factors (Hancock et al., 2011), also known as the three factor model of trust (Schaefer et al., 2016). The strongest correlation was found for performance-related factors of the agent, followed by a moderate correlation with environmental factors, and little influence from human-related factors. This signifies the importance of a well-functioning agent in HAT. Without proper agent performance, including adequate communication and transparency (Barnes et al., 2014; Chen & Barnes, 2014), it will be difficult for a human to trust an intelligent agent.

Summary

It is evident that a new phenomenon has entered the natural world: human-agent teaming (HAT), or collaborative teamwork between intelligent entities, including human and physically embodied agents. The U.S. is making great efforts to implement these agents in the military force. Researchers have attempted to learn more of this new phenomenon based on research and theoretical models from related scientific fields. What is needed is an overarching theoretical model that helps to understand, explain, and predict HAT performance to facilitate the implementation of agents at the squad level. After a review of key constructs and theories, it is evident that several theoretical models of HAT exist. However, these apply only to components of HAT, rather than to the overarching subdomain of HAT (**Figure 1**). Without an integration of such models and constructs into a comprehensive model, the relative importance of these components in presence of the other constructs remains unknown. This weighted understanding is needed to optimize experimental design and prediction of HAT performance.

Goal Statement

In order to robustly predict performance in HAT, a clear understanding is needed of the most important contributors to this performance. The present effort aims to fill this gap by proposing a theoretical model of HAT for dynamic and complex environments, such as the military, that integrates key constructs identified in HAT research. Military HAT research utilizes mainly simulated agents (e.g., Mercado et al., 2015). Thus, the model will be developed specifically for simulated military HAT. The goal is to test (part of) the model against empirical data.

CHAPTER TWO: A MODEL OF SIMULATED MILITARY HUMAN-AGENT TEAMING

In this Chapter, a model of simulated military HAT is developed. As shown in **Figure 2,** the model centers around task performance and consists of three layers: the Core, a Relationship Layer, and an Environmental Layer that interconnect through a transactional interaction. Each of these sections will be discussed in this Chapter. As will be elaborated upon, the Core model is the primary focus of the present effort. Moreover, the model is here applied to simulated military HAT missions yet could be potentially extended to other dynamic and complex environments wherein humans and agents collaborate.

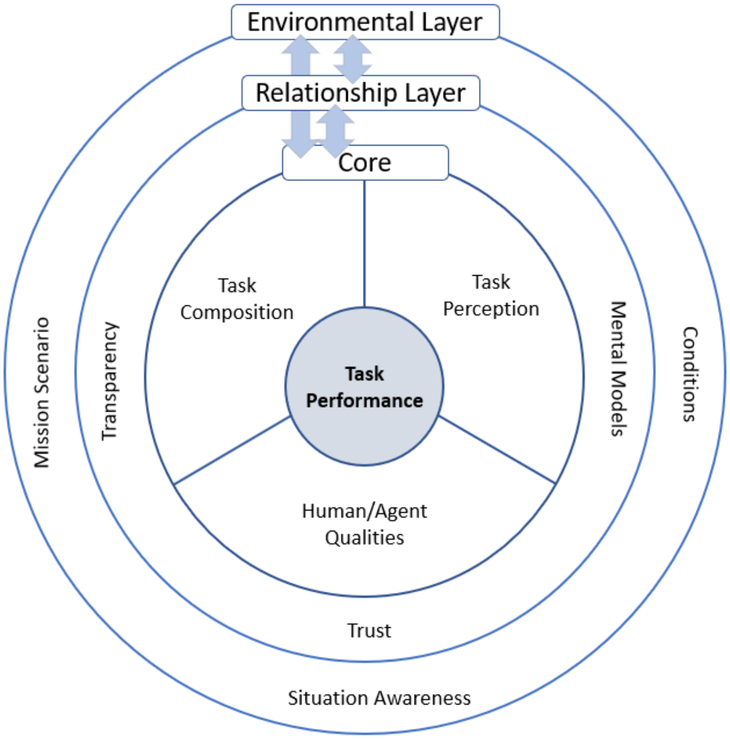


Figure 2. Model of simulated military human-agent teaming (HAT) that centers around task performance.

Note. The model of simulated military HAT consists of three layers. The outer Layer has the least direct impact on Task Performance: the Environmental Layer. This Layer consists of environmental variables, such as the scenario in which the mission takes place, environmental conditions, and overall awareness of the task, relationship, environment, and performance (situation awareness). The Relationship Layer focuses on the relationship between the human and agent teammate(s), with constructs as mutual trust, mental models, and transparency. The Core model directly impacts Task Performance and consists of Task Components, Task Perception, and the Qualities the Human/Agent bring to the team. The current effort focuses on the Core model. Lastly, the layers are transactional, as represented by the two-way arrows. The variables in one layer can affect the variables in the other layer and Task Performance, although a threshold may need to be reached.

Task Performance

The model identifies performance as the focal point within teaming paradigm. In military missions, and other dynamic and complex scenarios such as search and rescue missions, performance on the task is the most important criterion, with an accuracy standard of approximately 90% (Naval Education and Training Command, 2009). In critical military operations, where human lives are at stake, the relationship between team members is rudimentary, although the basic foundation of trust and taking ownership for the mission needs to be present.

The Core

The Core model is the primary focus of the present effort based on the assumption that the Core is the most important portion of the model in relation to task performance. The Core consists of task characteristics (Task Composition), qualities of the human or agent (Human/Agent Qualities), and their perception of the task (Task Perception). These three components are proposed to be of equal importance to task performance.

Task Composition

Research in various domains has consistently shown that characteristics of the task, or Task Composition, affect task performance (Green, 1993; Lu et al., 2013; See et al., 1995; Szalma et al., 2008). Here, some of the common analyzed components of task composition in relation to HAT performance are discussed, including event rate, signal probability, and modality (Teo et al., 2018).

Event Rate

Event rate is the rate at which stimuli, both targets and non-targets, are presented within a given time period (Wickens & Hollands, 2000). In general, higher event rate is more taxing on the human information processing system than low event rate (Barber et al., 2019; Wickens, 2008). However, low event rate can also be experienced as taxing if the likelihood that one of these stimuli is a target is low (Dillard et al., 2014; Grier et al., 2003; Hancock & Warm, 1989). A foundational example was published by Mackworth (1948) where he described the tendency of the Royal Air Force to miss critical but rare occurrences on the sonar and radar screen when attempting to detect enemy submarines during World War II. Despite operators' high motivation to detect the enemy, errors of omission were made.

Signal Probability

As mentioned, signal probability reflects the likelihood of a critical event, e.g., a target or threat, occurring (Warm & Jerison, 1980). The effects of signal probability on performance stems from the field of vigilance, starting with Mackworth's (1948) seminal work. Vigilance is a highly specialized psychophysical field focused on the study of the ability to maintain attention over a long period of time (Parasuraman & Davies, 1977). Herein, individuals monitor for a critical but very seldomly occurring signal (low signal likelihood), in a static environment such as in air traffic control or cybersecurity (Brookings et al., 1996; Sawyer et al., 2014). Performance is known to drop significantly over time, a phenomenon known as the vigilance decrement (Grier et al., 2003; See et al., 1995). In experimental studies focused on HAT for dynamic and complex environments, the environment has more dynamic movement and the task duration is often much shorter than the average 40 minutes in vigilance (e.g., for simulated HAT see Abich et al., 2013, Barber et al., 2019, and Bendell et al., 2019; for vigilance see See et al., 1995). Even though the HAT paradigm does not meet the standards of vigilance, the field of vigilance may inform HAT as the tasks both involve monitoring an environment for critical events. In a cordon-and-search mission, Soldiers monitor the dynamic environment for insurgents and contraband for a potentially prolonged period of time (Sutherland et al., 2010). Based on knowledge of cognitive processing resources, higher event rate and lower threat probability leads to lower performance than low event rate and higher threat probability (Wickens & Hollands, 2000). Vigilance research additionally suggests that low event rate and low threat probability can be detrimental for performance (Dillard et al., 2014; Grier et al., 2003).

Modality

In military missions, it is crucial that Soldiers can communicate their findings and keep each other in the loop to reduce threats to their squadron. Communication between the human and the agent can occur through a number of modalities: auditory in the form of speech, visual in the form of gestures and images, and tactile through meaningful haptic patterns (**Table 4**; Lackey, Barber, Reinerman-Jones, Badler, & Hudson, 2011).

Table 4

Communication modalities in human-agent interaction.

Modality	Delivery	Explicit	Implicit
Auditory	Speech, sounds	Language	Tone, rate, pitch
Visual	Posture, facial expression, gesture, gait, social distance, images through interface	Intentional pointing, hand signals, imagery	Unintentional body language, intensity, eye contact, talking with hands, emotions
Tactile	Belt, vest	Intentional touching, patterns	Pressure, patterns, shakiness

Note. Adapted from Lackey et al. (2011).

Auditory communication can be expressed in formal language and implicit alterations of such language, e.g., tone, rate, and pitch (Lackey et al., 2011), which is mainly of interest in social robotics. In general, communication through auditory modalities tends to be picked up faster by humans (Latorella, 1998; Wickens, Dixon, & Seppelt, 2005). Moreover, when auditory communication occurs during an ongoing visual task, the tendency to identify a stimulus as a threat becomes more conservative (Bendell et al., 2019a). In addition, new developments show that enhanced auditory cues, such as spatialized audio, are useful in providing spatial localization information while reducing workload (Kim et al., 2018).

The visual communication modality facilitates communication between human and agent teammates, even when auditory communication means are compromised. Visual agent-to-human communication can take place through the means displayed in **Table 4**. For dismounted operations, useful visual communication means are visual displays and gestures (Dumas et al., 2009; Harris & Barber, 2014). Gestures are useful when agents are in the line-of-sight of humans. If this is not available or not preferred, agents can communicate through interfaces for conveying visual representations of messages in the form of maps, pictures of objects, video feeds, and text. Moreover, visual display communication is effective in providing transparency of the agent's state (Mercado et al., 2015).

However, visual display communication may interfere with the human's continuous visual attention to the environment, especially on traditional displays where the user's head is down. Even heads-up displays with mission-critical information on the screen may be a distraction away from the primary task and may lead to performance degradation (Lewis & Neider, 2016; Sawyer, 2015; Wickens, 2017).

Another communication modality is tactile, which is less obtrusive, as it delivers information via a tactile belt/vest or wearable devices, through tactors that apply electromechanical vibration to the skin (Fitbit, 2019; White, 2010). These forms of communication facilitate the conveyance of simple messages, in the form of a tactile one- or two-word lexicon (Barber et al., 2015; Reinerman-Jones et al., 2017) or cues relating to spatial orientation and navigation (Ho et al., 2005; Prewett et al., 2012). Moreover, due to it inobtrusive nature, tactile cues are functional in the military battlefield.

Lastly, multimodal communication, i.e., communicating through more than one modality simultaneously (Dumas et al., 2009; Oviatt, 2012), is beneficial when accuracy is vital (Dobrišek et al., 2013; Huey & Wickens, 1993; Maurtua et al., 2017). However, this method affects multiple modality resources at the same time and increases workload (Lu et al., 2013; Wickens, 2002; Wickens et al., 2011). In the battlefield, accuracy makes the difference in life or death, in which case the need for accuracy may outweigh the increased workload from multimodal information presentation.

Task Perception

Task Perception refers to the way in which individuals perceive or experience the task. Task Perception may impact task performance, as it relates to compensatory strategies, or self-regulation, employed by individuals to modulate task performance (Hancock & Warm, 1989; Hockey, 1997; Negretti, 2012). Task Perception is conceptualized in terms of perceived workload and perceived stress.

Perceived Workload

Workload is a complex psychological construct that refers to a cognitive state indicating the load imposed on the human information processing system by the contextual environment (Matthews et al., 2019; Matthews & Reinerman-Jones, 2017; Stanton et al., 2017). Perceived workload is the individual's reflection of the cost incurred by the task and is measured with rating scales (Hart & Staveland, 1988). Two additional measurements of workload exist (Matthews & Reinerman-Jones, 2017; O'Donnell & Eggemeier, 1986). Performance measures of workload indicate the effect of a dual task on the cognitive information processing system. If secondary task performance drops, it is postulated that the primary task depleted the information processing resources. The third measurement of workload is formed by various physiological measures. Here, neurophysiological measures, such as cerebral blood flow velocity, signal the level of involvement of specific brain regions (Neubauer et al., 2013), while cardiovascular measures, such as heart rate variability, are more indicative of the level of effort (Thayer et al., 2012).

More recently, research showed that these three measures do not consistently converge, which may indicate a multidimensional rather than unitary workload construct (Hancock & Matthews, 2019; Matthews et al., 2015; Matthews et al., 2019; Matthews & Reinerman-Jones, 2017; Yeh & Wickens, 1988). In this debate of construct validity, subjective measurements of workload have received most criticism. Matthews, de Winter, and Hancock (2019) succinctly summarize the criticism into two fundamental concerns.

The first concern relates to the philosophical issues with quantification of a psychological experience. Questions such as what the appropriate scale is to use, how to define the construct, the effects of bias of memory due to the time lapse between task and evaluation, and the bias of contextual effects, mostly remain unanswered in this philosophical debate (Annett, 2002). However, for a number of reasons, the use of subjective rating scales continues for psychological constructs (e.g., de Winder, 2014). Measures of perceived experiences have value in understanding of a phenomenon if used with relevance to the study, wherein it is used as a representational measure (Annett, 2002; Hand, 1996). In addition, measures of perceived workload have shown to be useful in predicting performance in HAT (Abich et al., 2013; Abich et al., 2017), which emphasizes an operational use of the measure (Annett, 2002; Hand, 1996). In operationalism, “an attribute is defined by its measuring procedure” (Hand, 1996, p. 453), thus the measure is all one needs to know regarding the construct. The measure is the construct.

The divergence problem between perceived measures of workload with other measures of workload may reflect psychometric issues, which is the second fundamental concern of the use of subjective workload rating scales (Matthews et al., 2019). However, this notion does not necessarily invalidate the use of perceived measures of workload. Rather, the divergence may reflect a multifaceted construct of workload rather than a unitary construct (Matthews et al., 2015). For instance, perceived workload is suggested to be sensitive to the number of tasks being performed, while performance measures are sensitive to the modality used for both tasks, impacting the resource demand and availability (Vidulich & Tsang, 2012).

Matthews, de Winter, and Hancock (2019) suggested that subjective workload measures, such as the NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988), are important in terms of self-regulatory strategies. Through the perception of increased demand and potential drops in performance, individuals make a strategic decision in terms of up- or downregulating their information processing resources or effort toward the task (Hockey, 1997), which may be further regulated by differences in personality (Matthews & Campbell, 1998). The self-regulation hypothesis certainly would explain the dissociation often seen between subjective workload levels (e.g., high workload) and performance (e.g., maintained performance), and would validate the continued use of subjective rating scales for perceived workload if used appropriately.

Perceived Stress

Stress is “a particular relationship between the person and the environment that is appraised by the person as taxing or exceeding his or her resources and endangering his or her well-being” (Lazarus & Folkman, 1984; p. 19). This definition emphasizes the subjectivity of the experience of stress. Not every person responds in the same manner to identical stressors; it depends on the way in which the individual interprets the conditions (Lazarus & Folkman, 1984). Stress may impair performance by changing the individual’s adaptability to the task (Hancock & Warm, 1989). Similar to perceived workload, the perception of stress due to task demand is a regulator of effort in response to increased task demand (Hockey, 1997), which may be further moderated by personality differences (Matthews et al., 2019; Matthews & Campbell, 2009). Military personnel are exceptional in handling stressful environments and maintaining task performance, which may in part be due to personality differences. Indeed, military members have a different personality profile, characterized by lower scores on agreeableness, neuroticism and openness to experience (see **Table 5** for definitions) than non-enlisters prior to enlistment (Jackson et al., 2012). After enlistment, their military training subsequently alters these traits by lowering agreeableness further (Jackson et al., 2012).

Human/Agent Qualities

In the Core model, Human/Agent Qualities are included as constructs that affect task performance. A team is only as good as its constituents or the qualities that each entity brings to the team, which is conceptualized based on their personality traits and entity-specific qualities.

Personality

The most common theory of personality traits is the Big Five, which resulted from factor analyses indicating five general dimensions: neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness (**Table 5**; Costa & McCrae, 2008; Digman, 1990).

Table 5

Big Five personality traits (Costa & McCrae, 2008).

Big Five Trait	Description
Neuroticism	Level of emotional stability, indicating the ease of anxiety, frustration, worry, and irritability
Extraversion	Level of sociability, dominance, thrill seeking, and energy
Openness to experience	Level of creativity, imagination, and enjoyment of new activities and experiences
Agreeableness	Level of sympathy, altruism, and tenderheartedness
Conscientiousness	Level of goal-direct efficiency, planning/organization, and responsibility

Research indicates that higher levels of agreeableness and conscientiousness are beneficial for team performance, operationalized as a composite of various organizational work outputs (O’Neill & Allen, 2011; Peeters et al., 2006). Furthermore, higher levels of neuroticism predict impairments of cognition, including attentional resources and working memory, and a higher negative sensitivity to threats (Matthews et al., 2003).

In 2011, this research was extended from human teaming to human-agent teaming, in a study with a disembodied agent that served as a decision-making aid to the human (Szalma & Taylor, 2011). Their results showed that task factors posed stronger effects on task performance than personality traits. Neuroticism and conscientiousness significantly correlated with performance in opposite directions: high neuroticism corresponded with lower accuracy, while high conscientiousness correlated with higher accuracy. Furthermore, the effects of personality traits on perceived workload and stress were significant. High neuroticism significantly predicted higher perceived stress and workload than the other four traits. To date, research studies like these have not yet been extended to embodied HAT. However, since the results are congruent between human-teaming and HAT, similar results are likely to be found.

Entity-Specific Qualities

Human Qualities

Task performance has been linked to a number of human qualities that are difficult to separate, including age, gender, and experience. For instance, experience with a task increases over time due to repeated exposure and tends to improve performance. Since increased time and increase in age may coincide, age may be associated with experience. However, this is not necessarily the case, since cognitive decline is also related to an increase in age (Matthews et al., 2000). Moreover, expertise may compensate for age-related performance decline for domain-specific tasking (Morrow et al., 1994).

Similarly, video gaming experience seems beneficial for performance on simulated military HAT missions (Chen & Barnes, 2012). In general, youngsters tend to engage more in video gaming that older people. Another construct that is potentially interwoven in video gaming experience, is gender. Men tend to play more video games than women (Lin et al., 2015). Men also tend to have higher scores on spatial ability tests than women (Chen, 2010; Hyde, 1990; Maeda & Yoon, 2013). The question which of the two, gender or spatial ability, is more beneficial for performance remains unknown. While this answer is yet unknown, the discussed research indicates that these human qualities may be important for task performance.

Agent Qualities

The agent team member also brings qualities to the team, as explained by the notion of Men Are Better At – Machines Are Better At (MABA-MABA; Fitts, 1951; **Table 2**). Research shows that the level of automation assigned to the agent is beneficial for routine task performance, although it may also lead to problems with human take-over qualities and situation awareness (Onnasch et al., 2014; Sebok & Wickens, 2017). To this end, the importance of transparency arose.

Agent qualities have also been operationalized in terms of reliability, which indicate the capability of the agent to accurately perform its task, expressed in a percentage. This construct has been applied mostly when the agent is disembodied (a software agent) and serves as a decision-making aid (e.g., Chen & Terrence, 2009; Szalma & Taylor, 2011). Wickens and Dixon (2005) demonstrated that the cut-off for agent reliability was below 70%; below this point human task performance deteriorated significantly. Chen and Terrence (2009) applied this to HAT yet set the reliability level to 60% and compared the type of error made by the agent (false positive vs. missed). They found that agent unreliability affected performance in interaction with the human team member's capability, corroborating the notion that both must be considered in a model of HAT performance. Aside from impacting task performance, agent reliability also affects trust (Hancock et al., 2011), indicating a transaction with a construct in the Relationship Layer.

Agent qualities are not just conceptualized in terms of level of automation or reliability, but also in terms of affordances. Affordances are what an object or system naturally allows the user to do, e.g., flat surfaces at hip level invite us to sit on it (Norman, 2013). It was originally posed by Gibson (1979) as a term within ecological psychology, highlighting what the environment offers to the animal. An embodied agent is naturally afforded with more communication qualities than disembodied agents. Embodied agents can use gestures and movements to communicate a message, while disembodied agents can only rely on text messages for this purpose. If an embodied agent has a mouth (or speakers) it may also be afforded with the ability to speak. Thus, the physical form and structure of the agent, or morphology, naturally determines its qualities. The morphology of an agent also interacts with the human's interpretation of the agent (Fong et al., 2003).

Humans tend to anthropomorphize objects and entities they interact with; people 'humanize' entities, that is to ascribe human traits, attitudes, and emotions to an entity (Epley et al., 2007). Anthropomorphism aids human understanding and prediction of the entity's behavior, based on their own inherent knowledge-base (Duffy, 2003; Epley et al., 2007), and adds to the human’s mental models of the agent (Kiesler & Goetz, 2002; Powers & Kiesler, 2006; Sims et al., 2005; Talone et al., 2015). In general, people tend to find familiar forms more accessible, desirable, and expressive (Fong et al., 2003), which is important for the implementation of agents as social entities (Relationship Layer). However, there is a treacherous balance in the design of humanoid features and human acceptance of the agent. If an embodied agent is designed to be too similar to the human, it runs the risk of appearing creepy, a phenomenon known as the Uncanny Valley (Mori et al., 2012). This phenomenon is accentuated when movement is taken into account (Mori et al., 2012). Moreover, anthropomorphic design can lead to unrealistic human expectations of the agent, which may negatively impact trust and acceptance (Duffy, 2003; Hancock et al., 2011).

Another concern in relation to anthropomorphized agents, specifically in the military, is the creation of a social bond with the agent that may inhibit Soldiers to send the agent in the dangerous battlefield (Carpenter, 2016). This notion may be valid, as anthropomorphism has shown to affect empathy (Riek et al., 2009). However, other research indicates that military embodied agents are generally perceived as more machine-like than robot-like (Schaefer et al., 2012). Nevertheless, anthropomorphic agents do offer undeniable advantages in their communicative affordances, such as deictic gestures.

Relationship Layer

After discussing Task Performance as the center of the model and the Core model as directly impacting Task Performance, the next part of the model (**Figure 2**) to discuss is the Relationship Layer. The Relationship Layer contains construct that pertain to the relationship in a human-agent team based on HAT research: mental models, mutual trust, and transparency. These constructs have also been identified as important components in human teamwork (Salas et al., 2005).

Mental Models

Mental models refer to a heuristic type understanding that allow people to describe, explain, and predict the world around them (Rouse & Morris, 1986). They are a critical component of effective teaming (Cannon-Bowers et al., 1993; Klein et al., 2005) and may contain variable contents (Johnson-Laird, 1983). Four different mental models are proposed in relation to teamwork: models about technology/equipment, the task at hand, team interaction, and the team member's qualities and limitations (Cannon-Bowers, Salas, & Converse, 1993). Salas et al. (2005) model merges these four mental models into shared mental models. Shared mental models refer to the mutual understanding of the task goal, each team member's responsibilities, and the coordination required to achieve the goal. This is different than situation awareness, which refers to a presently updated perception and understanding of the progress of the task, team members, and environmental conditions (Endsley, 1995). In terms of a shared mental model, all parties need and understanding of and commitment to the task at hand, sharing the same goal and common ground (Klein et al., 2004, 2005). The mental model of team interaction refers to an understanding of the roles and responsibilities of each team member and the way in which to communicate (Cannon-Bowers et al., 1993; Mathieu et al., 2000). Lastly, all members need to be critically aware of their strengths and limitations to be able to provide appropriate back-up behavior.

Agent Mental Model of Human

Agent's mental models of the human teammate and task can be computed through machine learning and decision making algorithms (Adams, 2014; Jonker et al., 2010; Ososky et al., 2012; Scheutz et al., 2017). Herein, the agent's algorithm of the mental model emphasizes similarity of mental models between human and agent, as this 'sharedness' leads to mutually similar expectations for the task goal and team (Jonker et al., 2010).

Human Mental Model of Agent

Human mental models of the agent refer to the ideas that humans form of agents to support their predictions and understanding of agents (Mathieu et al., 2000; Ososky et al., 2012; Phillips et al., 2011). Human mental models of the agents are affected by the morphology and communication affordances of the agent (Phillips et al., 2011) and are based on extrapolation of existing knowledge and experiences (Lee et al., 2005). As such, mental models benefit from education and training (Nikolaidis & Shah, 2012; Ososky et al., 2012).

However, when explicit knowledge is unavailable, mental models can also be formed based on analogies (Bailer-Jones, 2002). For example, when computers first entered the market, humans lacked technological knowledge and comprehension of these systems. Microsoft bridged that gap with the introduction of a folder icon system to provide an analogy for file storage. Although this is not an accurate representation of information storage on a computer, it provides a sufficiently accurate understanding of ‘storage’ for laymen to understand how they can store and search for files. Similar to computers, people are generally unaware of the technical workings of an intelligent, embodied agent. Therefore, they create mental models based on their experience and existing knowledge to aid their understanding and prediction of them. A common criterion of effective mental models is the extent in which they aid in the understanding and predicting behavior of the agent (Norman, 2013). However, for military HAT, mental models need a higher degree of accuracy than conventional mental models, as the military battlefield is more extreme and dangerous (Phillips et al., 2011).

Trust

Another construct important to the relationship in HAT is mutual trust. In Chapter 1, trust in agents in HAT was defined as the attitude that an agent will help achieve the team’s goals, rather than the individual’s goal (Hancock et al., 2011). When team members trust each other, they understand that others monitor their performance with the task in mind, rather than being out to ‘get them’ (Salas et al., 2005). However, building trust is a process: trust needs to be developed and calibrated (Salas et al., 2005; Schaefer et al., 2019). Furthermore, human trust in the agent depends in part on appropriately formed mental models, accurate SA, and agent transparency (Schaefer et al., 2019).

Transparency

In transparency, the focus is on the information that teammates convey about their decision-making process. The SAT model (**Table 3**) explains the information that the agent and human teammates should communicate, in line with the three levels of SA: agent's current status/actions/plans, agent's reasoning process, and agent's projections/predictions and/or uncertainties. As such, agent transparency can enhance the three levels of SA as proposed by Endsley (1995). Since it is a relational action, transparency has its place in the Relationship Layer rather than the Environmental Layer. Moreover, transparency directly affects trust by increasing the human's understanding of the agent's actions (Schaefer et al., 2017; Selkowitz et al., 2017).

Environmental Layer

The outer layer of the model (**Figure 2**) is the Environmental Layer. Herein, several facets of the environment are covered, including environmental conditions (e.g., weather, day/night, extreme temperatures), mission scenario, and situation awareness (SA of the task, environment, and team).

Environmental Conditions

Differential weather circumstances scope the task and qualities of a dismounted military team, especially when these conditions are extreme (e.g., rain, ice, fog). Moreover, when working with an embodied agent, it is important to understand the effect of extreme conditions on the agent as well. During the search and rescue missions of the 9/11 terrorist attacks with embodied agents, issues were encountered due to unforeseen effects of the environment on the agent: tracks were melting (Murphy, 2004). Furthermore, extreme environmental circumstances can deteriorate the ability of the team to perform the mission. For instance, fog impacts visibility and thereby affects both the primary mission, if this is vision-based, but also the modality through which team members can communicate by limiting visual communication qualities.

Mission Scenario

The mission scenario determines the goal and criticality of the mission, and thereby affects the team's mental model of the task, the extent to which they need to rely on each other, and Task Perception. Here, the mission scenario is a dismounted military mission, wherein threats are identified. Misidentification of threats could result in life or death. Moreover, these military missions are dynamic; anything can change at any point in time. For instance, in a military operation, the number of individuals to monitor for threat identification may change. In such circumstances, the mission scenario affects Task Composition (event rate: number of characters available per given timeframe). Thus, the mission scopes the task at hand and may interact with the definition of the team's mental model of the task among other things.

Situation Awareness

The last component of the Environmental Layer to discuss is situation awareness (SA). As discussed in Chapter 1, SA refers to the perception of elements in the environment (Level 1), the meaning and understanding of this observation (Level 2), and the projection of the status in the near future (Level 3; Endsley, 1995). In the Environmental Layer, SA represents the bird's eye view that team members have over the Task (Composition), the qualities that each team member has (Human/Agent Qualities), the relationship between team members (Relationship Layer), and an awareness of the environment (Environmental Layer).

Transactional Effects

As shown in **Error! Reference source not found.**, the three layers of the model, the Core, Relationship Layer, and Environmental Layer interconnect through a transactional interaction as depicted by the two-way arrows between the layers. This represents the notion that the construct within each layer can affect the constructs in other layers. These transactional effects may ultimately impact task performance, although a threshold may need to be reached before this occurs.

The first transaction to discuss is the between the Relationship Layer and the Core. For instance, as discussed, agent morphology (Human/Agent Qualities within the Core) affects the human's mental model of the agent (Phillips et al., 2011). Research suggests that mental models affect human trust in the agent, which may affect task performance (de Visser et al., 2019). Furthermore, it was found that higher agent transparency (Relationship Layer) may improve HAT performance (Mercado et al., 2016), although the mechanism through which this works remains to be explained in the HAT field. Some research suggests that agent transparency may affect the human's perception of the task, in terms of stress and perceived workload, and as such performance may improve (Mercado et al., 2015).

In addition, there is a transaction between the Environmental Layer and the Relationship Layer. The connection between transparency (Relationship Layer) and situation awareness (Environmental Layer) has been explained by the SAT-model discussed in Chapter 1 (**Table 3**) and fortified by research. Studies indicate that the highest level of agent transparency leads to higher human SA compared to lower levels transparency (Selkowitz et al., 2017). Moreover, accurate SA (Environmental Layer) is said to play an essential role in development of trust (Relationship Layer; Salas et al., 2005; Schaefer et al., 2019). Trust and SA exchange transactional meaning through awareness of the task requirements, the actions each team member intends to perform and their reasoning for these decisions (Chen et al., 2014; Chen et al., 2018). If these actions and decisions are aligned with the mission goal and trust is well-calibrated, there is a beneficial effect on HAT. Lastly, the mission scenario (Environmental Layer) informs the team's mental model of the task, in terms of the goal and criticality of the mission.

The last transactional effect is between the Environmental Layer and the Core. As discussed, environmental circumstances may deteriorate the ability of the team to perform the mission. This occurs particularly when the conditions are extreme, thus the threshold to impact task performance may be relatively high. Additionally, the mission may affect task performance, through an interaction with the perception of the task (the Core), the qualities of the team members (the Core), and the accurate calibration of mutual trust and mental models (Relationship Layer). Accurate SA also updates the team's perception of the task and may subsequently contribute to self-regulatory strategies that may affect performance (Vidulich & Tsang, 2012).

Summary

The main objectives of the present effort were to develop a model of simulated military HAT and to propose an approach to validate the model with empirical data. The literature review elucidated components that contribute to HAT performance, that were integrated into a proposed model, wherein task performance is central. This model consists of three layers. The outer layer (Environmental Layer) contains environmental variables and a bird's eye view over the teaming paradigm. The middle layer is the Relationship Layer and pertains to constructs that affect the relationship between team members in HAT. The focus of this research effort is on the Core model. The Core includes components that directly affect task performance: Task Composition, Task Perception, and Human/Agent Qualities. All aspects of the Core will be tested against the null model, i.e., everything is equally important. A validation approach is presented in CHAPTER THREE: METHODS AND PROCEDURES and applied to validate the Core model (**Error! Reference source not found.**).

CHAPTER THREE: METHODS AND PROCEDURES

The goal of the present effort was twofold. The first objective was to develop a model of HAT performance, which was provided in CHAPTER TWO: A MODEL OF SIMULATED MILITARY HUMAN-AGENT TEAMING (**Figure 2**). The second goal of this effort was to propose an approach to validating the Core model, hence, developing a validation approach for models that imply relative importance of the components. This validation approach was used to falsify the Core model against experimental data borrowed from the RCTA (Childers et al., 2016). In the next sections, the borrowed data will be described, as well as the specific hypotheses and description of the methodology or validation approach.

Experimental Data

De-identified experimental data was taken from the past decade of research under the RCTA (Childers et al., 2016), as “Not Human Subjects Research” (APPENDIX L: IRB DETERMINATION DISSERTATION). Studies were selected based on the following inclusion criteria:

- Contains a signal/threat detection task.
- Contains an additional task that requires collaboration with an embodied agent, such as agent reporting.

Four studies of the RCTA, reported by Abich et al. (2017), Barber et al. (2017), Barber et al. (2019), Bendell et al. (2020), and Kopinsky (2017), met the above criteria. A full description of the studies can be found in APPENDIX A: DESCRIPTION BORROWED STUDY A through APPENDIX D: DESCRIPTION BORROWED STUDY D. Approval from the Institutional Review Board (IRB) of these studies is included in APPENDIX G: IRB FOR BORROWED EXPERIMENTAL STUDY A through APPENDIX J: IRB FOR BORROWED EXPERIMENTAL STUDY D.

In each of the utilized studies, the ongoing task was a simulated military cordon-and-search operation (Sutherland et al., 2010), wherein participants identified threats among the humanoid characters walking across the screen. Participants identified a threat by clicking on them with a pointing device. During the mission, participants worked with (a) simulated embodied agent(s) that conducted an independent search of a designated cordon. The agent reported back to the human about its findings. The content of the reports varied between the studies (see Appendices A through D). The modality through which the human teammate received the agent's report also varied between studies: auditory through headphones, visual through an interface, or dual (both simultaneously). To ensure engagement, participants were instructed to memorize these reports as they were later randomly probed. A data matrix is available in **Table 6**.

Table 6

Data matrix.

Manipulation	Experimental Study (Source: Abich et al., 2017; Barber et al., 2017; Barber et al., 2019; Bendell et al., 2020; Kopinsky; 2017)									
	Study A.1	Study A.2	Study B.1	Study B.2	Study B.3	Study C.1	Study C.2	Study C.3	Study D.1	Study D.2
Task Composition										
Event rate										
15 characters/min.	•	•		•		•*	•*	•*		
30 characters/min.			•		•					
60 characters/min.				•		•*	•*	•*	•	•
Signal likelihood										
0.09-0.10									•	•
0.12-0.13			•	•	•	•	•	•		
0.13-0.14	•	•								•
Task duration										
5 minutes				•						
10 minutes			•		•					
12 minutes	•	•								
15-16 minutes									•	
32 minutes						•	•	•		•
Agent Task Type										
Receive Report	•	•				•	•	•	•	•
Pull Report			•	•						
Visual Complexity										
Basic	•									
Enhanced		•								

•* Coded as NA

Manipulation	Experimental Study (Source: Abich et al., 2017; Barber et al., 2017; Barber et al., 2019; Bendell et al., 2020; Kopinsky; 2017)									
	Study A.1	Study A.2	Study B.1	Study B.2	Study B.3	Study C.1	Study C.2	Study C.3	Study D.1	Study D.2
Task Composition										
Agent Report Delivery Frequency										
Interval									•	
Immediate										•
Agent Report Modality										
Auditory									•	•
Visual	•	•	•	•	•				•	•
Single-Adaptive						•	•			
Dual								•		
Human/Agent Qualities										
Agent Type										
Legged	•									
Wheeled		•								
Demographics										
Age	•	•	•	•	•	•	•	•	•	•
Gender	•	•	•	•	•	•	•	•	•	•
Experience										
Military Experience			•	•	•					
Video Gaming Experience	•	•				•	•	•		
Task Perception										
Perceived Workload (NASA-TLX)										
Mental Demand	•	•	•	•	•	•	•	•	•	•
Physical Demand	•	•	•	•	•	•	•	•	•	•
Temporal Demand	•	•	•	•	•	•	•	•	•	•
Effort	•	•	•	•	•	•	•	•	•	•
Frustration	•	•	•	•	•	•	•	•	•	•
Performance	•	•	•	•	•	•	•	•	•	•

Manipulation	Experimental Study (Source: Abich et al., 2017; Barber et al., 2017; Barber et al., 2019; Bendell et al., 2020; Kopinsky; 2017)									
	Study A.1	Study A.2	Study B.1	Study B.2	Study B.3	Study C.1	Study C.2	Study C.3	Study D.1	Study D.2
Task Performance										
Threat Detection Accuracy										
Hit rate	•	•	•	•	•	•	•	•	•	•

Note. This data matrix describes the experimental data available from four studies that were borrowed from the RCTA (Childers et al., 2016) for the present research effort. The variables are categorized in accordance with the proposed Core model **Figure 2**. Task Composition variables pertain to characteristics of the task. Human/Agent Qualities include descriptors and qualities that human and agent team members bring to the teaming effort. Task Perception variables pertain to the human subjective experience of the task, which is here conceptualized in terms of the NASA-TLX (Hart & Staveland, 1988; APPENDIX E: NASA-TLX). Lastly, Task Performance was operationalized in terms of human performance on the threat detection task: hit rate. Hit rate was computed as the ratio of the correctly detected threats by the number of total available threats.

Hypotheses

The following null hypotheses were tested in relation to the Core model:

Hypothesis 1. Of the Human/Agent Qualities, all human/agent factors are equally important to task performance.

Hypothesis 2. All NASA-TLX subscales (Task Perception) contribute equally to task performance.

Hypothesis 3. All Task Composition variables contribute equally to task performance.

Hypothesis 4. Task Composition, Perception of Task, and Human/Agent Qualities are equally important to task performance.

Validation Approach

To validate the Core model, a method was selected that could unveil the factors that were most important to task performance. The method of choice was dominance analysis (DA). In DA, the dominance of a variable is established by comparing the unique additional contribution of the predictor to all possible subset (regression) models (Budescu, 1993). Understanding the unique contribution of the variables elucidates the underlying variable loadings onto the outcome and thereby facilitates prediction (Tighe & Schatschneider, 2014). DA has been widely used in fields of ophthalmology (Lips-Wiersma et al., 2018; Shakarchi et al., 2019), biomedicine (Nolan & Santos, 2019), clinical psychology (Shah et al., 2019), and education (Tighe &

Schatschneider, 2014). This method is also used in engineering fields and is there referred to as feature selection (Che et al., 2017; Kuhn & Johnson, 2013; Yu & Liu, 2004), which is a more "black box" approach to DA.

Dominance Analysis

Through comparison of the unique contribution each predictor yields to the response variable across different subset model sizes (DA), three levels of dominance can be determined: complete dominance, conditional dominance, and general dominance (Azen & Budescu, 2003; Budescu, 1993). A predictor is said to completely dominate the other predictors, if its additional contribution to each of the *k* model sizes exceeds the contribution of the other predictors on all subset model sizes (Budescu, 1993). Dominance of x_i over x_j in a subset (x_h) predictors is (Budescu, 1993)

$$\rho^2_{Y.x_ix_h} \geq \rho^2_{Y.x_jx_h} \tag{1}$$

or

$$(\rho^2_{Y.x_ix_h} - {\rho_{Y.x_h}}^2) \geq ({\rho_{Y.x_jx_h}}^2 - {\rho_{Y.x_h}}^2) \tag{2}$$

where $\rho^2_{Y.x_ix_h}$ is the squared multiple correlation of the model which includes the predictor x_i and the remaining predictors x_h, while excluding predictor x_j (Budescu, 1993).

Conditional dominance is established is a predictor's additional contribution within a specific model size is larger than the contribution of the others (Azen & Budescu, 2003; Budscu, 1993). The unique additional contribution of a predictor in terms of $\rho^2_{Y.x_i x_h}$ is expected to decrease monotonically as the subset models increase in size (*k* increases; Azen & Budescu, 2003).

General dominance is based on the average of all conditional values and is the lowest level of dominance (Azen & Budescu, 2003; Budescu, 1993). General dominance of x_i, with *p* additional predictors in model subset size *k*, with $C^{(k)}_{x_i}$ as the average additional unique contribution of x_i across all (*p* - 1) over *k* subset models, is computed as (Budescu, 1993)

$$C_{x_i} = \sum_{k=0}^{p-1} C_i^{(k)}/p \tag{3}$$

Budescu (1993) stipulated that dominance is transitive; that is, if x_i dominates x_2 and x_2 dominates x_3, then by definition x_1 dominates x_3. Moreover, if a predictor completely dominates all other predictors, this predictor will also have conditional and general dominance (Azen & Budescu, 2003).

Finally, the dominance pattern can be expressed in dominance indices (Azen & Budescu, 2003). If x_i dominates x_j, this is expressed as D_{ij} = 1. If the reverse is true, that x_j dominates X_i, then D_{ij} = 0. If dominance cannot be established for either predictor, D_{ij} = 0.5. Since DA does not yield statistical significance, these values are then bootstrapped, to determine the generalizability of the results as well as the internal reproducibility, with confidence interval computations (Azen & Budescu, 2003).

As shown in Equation (1) and (2), DA is based on comparative squared semi-partial correlations by running all possible subset ordinary least squares regression models. As such, the data needs to meet the assumptions or linear regression: normally distributed residuals, linearity, and independent errors (Pedhazur, 1973). However, in recent years, DA has been extended to logistic regression (Azen & Traxel, 2009; Tonidandel & LeBreton, 2010), hierarchical multilevel modeling (Luo & Azen, 2013), multivariate regression modeling (Azen & Budescu, 2006), and beta regression (Shou & Smithson, 2015; Smithson & Verkuilen, 2006). As the response variable in the present effort is operationalized as hit rate, i.e., the number of correctly detected threats divided by the number of available threats, the response variable is naturally double-bounded between 0 and 1 (Ferrari & Cribari-Neto, 2004; Smithson & Merkle, 2013). This data fits within the family of beta distributions. Therefore, DA was conducted based on beta regression models. With this distribution, parametric test statistics, such was squared semi-partial correlations, cannot be used to compare these models. Therefore, a more appropriate pseudo R^2 was selected for this effort.

Beta Regression

The density of y with $0 < y < 1$ is (Ferrari & Cribari-Neto, 2004; Shou & Smithson, 2015)

$$f(y|\mu,\theta) = \frac{\Gamma(\theta)}{\Gamma(\mu\theta)\Gamma(\theta(1-\mu))} y^{\mu\theta-1}(1-y)^{\theta(1-\mu)-1} \qquad (4)$$

wherein shape parameters are $\alpha > 0$ and $\beta > 0$, the precision parameter is $\theta = (\alpha + \beta)$ and the mean (μ) of y is (Ferrari & Cribari-Neto, 2004)

$$\mathrm{E(y)} = \mu \ = \frac{\alpha}{\alpha+\beta} \tag{5}$$

and the variance is

$$var(y) = \ \sigma^2 = \ \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\mu\,(1-\mu\,)}{1+(\alpha+\beta)} = \frac{\mu\,(1-\mu\,)}{1+(\,\theta)} \tag{6}$$

It follows that *var(y)* is a function of the mean.

For a random sample $y_1, \ldots, y_n$, with , with $y \sim B(\mu, \theta)$, $i = 1, \ldots, n$, the beta regression model is (Cribari-Neto & Zeileis, 2010)

$$g(\mu_i) = x_i^{\mathrm{T}} \beta_i = \eta_i \tag{7}$$

where $\beta = (\beta_1, \ldots, \beta_\mathrm{k})^\mathrm{T}$ is a k x 1 vector of unknown regression parameters ($k < n$), η_i is a linear predictor and $x_i = x_{i1}, \ldots, x_{ik})^\mathrm{T}$ is the vector of k regressors. The coefficients are estimated with maximum-likelihood estimators. Beta regression assumes linearity between the predictor and response variable through the link function. The link function between the linear predictor and the mean of the distribution function is (Cribari-Neto & Zeileis, 2010; Shou & Smithson, 2015)

$$g(\mu) = \log\frac{\mu}{1-\mu} \tag{8}$$

The residuals of a beta regression model are not estimated with $y_i - \widehat{\mu_i}$ due to the inherent heteroscedasticity of double-bounded variables (Smithson & Merkle, 2013). Ferrari and Cribari-Neto (2004) suggest the use of standardized ordinary residuals, defined as

$$r_{\mathrm{P},i} = \frac{y_i - \widehat{\mu_i}}{\sqrt{\widehat{\mathrm{VAR}}(y_i)}} \tag{9}$$

Where $\widehat{\mathrm{VAR}}(y_i) = \widehat{\mu_\iota}(1\text{-}\widehat{\mu_\iota})/(1+\widehat{\theta_\iota})$, $\widehat{\mu_\iota}=g_1^{-1}(x_i^{\mathrm{T}}\widehat{\beta})$, and $\widehat{\theta_\iota} = g_2^{-1}(z_i^{\mathrm{T}}\hat{\beta})$. Although the residuals are not necessarily normally distributed (Smithson & Merkle, 2013), they are assumed to be independent (Ferrari & Cribari-Neto, 2004).

Pseudo *R^2*

Since the assumptions of parametric goodness-of-fit estimators are not met within beta distributions (Smithson & Merkle, 2013), a more appropriate pseudo R^2 was tested and selected for this effort. Pseudo R^2 is used in other non-parametric models based on maximum likelihood estimators such as logistic regression. Azen and Traxel (2009) established four criteria in their effort to select an appropriate pseudo R^2 for DA on logistic regression (p. 324):

1. Boundedness: the goodness-of-fit measure is bounded between 0 and 1, wherein 1 indicates a perfect fit.
2. Linear invariance: the measure should be robust against linear transformations of the variable.
3. Monotonicity: the measure should increase when more predictors are added to the model.
4. Intuitive interpretability: the measure aligns with the scale of the intermediate values.

With these criteria in mind, Azen and Traxel (2009) selected and compared three log-likelihood-based pseudo R^2: McFadden's R_M^2 (1973), Nagelkerke's R_N^2 (1991), and Estrella's R_E^2 (1998). McFadden's R_M^2 is defined as

$$R_M^2 = \frac{\ln L_0 - \ln L_M}{\ln L_0} = 1 - \frac{\ln L_M}{\ln L_0} \tag{10}$$

wherein L_0 is the value of the likelihood function for a base model with 0 predictors and L_M is the likelihood for the estimated model, and ln() is the natural logarithmic value. R_M^2 met all four criteria set forth by Azen and Traxel (2009).

Nagelkerke's (1991) R_N^2 is based on Cox and Snell's (2018) pseudo R^2, which is

$$R_{CS}^2 = 1 - \left(\frac{L_0}{L_M}\right)^{2/n} \tag{11}$$

wherein the sample size is represented in n. One of the limitations of R_{CS}^2 is that the upper bound is smaller than 1.00; the upper bound is $1 – L_0^{2/n}$. Therefore, Nagelkerke (1991) adjusted for this limit by

$$R_N^2 = 1 - \frac{1-\left({L_0}/{L_M}\right)^{(2/n)}}{1-(L_0)^{(2/n)}} \tag{12}$$

R_N^2 satisfied all the criteria of an appropriate pseudo R^2 (Azen & Traxel, 2009).

Lastly, Estrella's (1998) measure is defined as

$$R_E^2 = 1 - \left[\frac{\ln L_M}{\ln L_0}\right]^{-(2/n)\ln(L_0)} \tag{13}$$

which is similar to McFadden's R^2, but raised to the power of $-(2/n)\ln(L_0)$. Estrella (1998) posits that this is needed to ensure that the derivative corresponds with the corresponding linear derivative. It is not as fluently interpretable as the other measures (Azen & Traxel, 2009).

In the current effort, Estrella's R_E^2 (1998) and Nagelkerke's R_N^2 (1991) cannot be used as they were designed for dichotomous outcome variables. In addition, McFadden's (1973) R_M^2 and

Nagelkerke's $R_N{}^2$ assume that the ML estimators are bounded between 0 and 1, which is not the case in beta regression (Shou & Smithson, 2015). However, Cox and Snell's (2018) $R_{CS}{}^2$ can be extended to regression models that use ML estimation (Allison, 2013) and allow for continuous maximum likelihood estimators (Shou & Smithson, 2015).

The interpretation of a pseudo R^2 is not as straightforward as the interpretation of an ordinary least squared regression R^2. The latter indicates the variance explained by the model. However, a pseudo R^2 can only be used to compare models ran on one dataset, wherein a higher R^2 indicates a better fit, i.e., prediction, of the model (Institute for Digital Research & Education Statistical Consulting, 2011).

In this effort, the R (R Core Team, 2013) code from Shou and Smithson (2015) to conduct DA on beta regression models was adapted, tested, and incorporated in the dominancenalysis, an R package, now available on CRAN (Bustos & Countinho, 2019). All four goodness-of-fit measures were compared and $R_{CS}{}^2$ was selected as the preferred pseudo R^2.

Validating Dominance Analysis

To validate the generalizability and reproducibility of the dominance indices (e.g., D_{ij} = 1 for dominance of x_i over x_j, D_{ij} = 0 for x_i not being dominant over x_j, or D_{ij} = 0.5 for an unestablished dominance pattern) were bootstrapped. Bootstrapping allows for inference about a the population based on random sampling with replacement of the sample (Efron & Tibshirani, 1986). The larger the N of sampling with replacement, the higher the change that all cases will be replicated at some point. Therefore, the bootstrap sample was set to S = 1000 bootstrap samples. Next, the dominance values were computed over the bootstrap sample, building a bootstrap distribution of the D_{ij} dominance values (Azen & Budescu, 2003). The average of these dominance values within the bootstrap sample is defined as the expected dominance of x_i over x_j in the population, with bounded values of (0,1) and is computed as

$$\overline{D_{\imath\jmath}} = \frac{1}{S}\sum_{s=1}^{S} D_{ij}^{s} \tag{14}$$

Then, the standard error of $\overline{D}_{ij}$ were calculated, based on (Azen & Budescu, 2003)

$$SE\left(D_{ij}\right) = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}(D_{ij}^{s} - \overline{D_{\imath\jmath}})^{2}} \tag{15}$$

which indicates the variability of the dominance index over the S bootstrap samples. Azen and Budescu (2003) set out guidelines for interpretation of the standard error (p. 140): "$\overline{D}_{ij}$ is 1 (and SE is 0) if, and only if, D_{ij} = 1 in all bootstrap samples. Conversely, $\overline{D}_{ij}$ is 0 (and SE is 0) if, and only if, D_{ij} = 0 in all bootstrap samples. Finally, $\overline{D}_{ij}$ is 0.5 if the distribution of dominance indices (D_{ij}^{s}) is symmetric in the sense that the number of cases in which x_i dominates x_j equals the number of cases in which x_j dominates x_i." Here, the SE depends on the number of indeterminate dominance values, wherein SE is 0 if, and only if, D_{ij} = 0.5 in all bootstrap samples.

Azen and Budescu (2003) proposed another method to evaluate the robustness of the results: a reproducibility value, based on three proportional measures reflecting the dominance indices in the S bootstrap samples, such that

$$P_{ij} = \Pr\,(D_{ij} = 1) \tag{16}$$

for the proportion of the S bootstrap samples that replicated the dominance index D_{ij} =1, i.e., that x_i dominates x_j,

$$P_{ji} = \Pr\,(D_{ij} = 0) \tag{17}$$

for the proportion of bootstrap samples that replicated findings of x_j dominating x_i, or D_{ij} = 0, and

$$P_{noij} = \Pr\,(D_{ij} = 0.5) \tag{18}$$

for the proportion of bootstrap samples that reproduced no dominance establishment for x_i over x_j. Lastly, a reproducibility value is computed that indicates the proportion of bootstrap samples that concur with the dominance results in the sample (Azen & Budescu, 2003). If a reproducibility value is 0.97, the researcher can be 97% confident of the dominance index (Azen & Budescu, 2003).

Model Fit Evaluation

Finally, regression analyses were conducted in the hierarchy of the established pattern of importance to determine the fit of the model. Herein, the beta regression that was previously discussed (Equation (3) – (6)) was applied and evaluated using the pseudo R^2 and χ^2 as the statistics for models based on log-likelihood (Tabachnick et al., 2013; Zeileis et al., 2019).

Summary Validation Approach

In summary, a validation approach was developed that is appropriate for testing models that are based on importance and have a proportion-based outcome variable. The validation approach consists of three consecutive steps:

1. Conduct dominance analysis on beta regression models to determine the most important contributors to the outcome variable (Azen & Budescu, 2003; Budescu, 1993).

2. Establish the robustness and generalizability of the dominance results by bootstrapping the dominance values (i.e., D_{ij} = 1, D_{ij} = 0, D_{ij} = 0.5; Azen & Budescu, 2003; Efron, 1981)

3. Combine the most important predictors into a hierarchical beta regression model and evaluate the fit of the model (Ferrari & Cribari-Neto, 2004).

This validation approach was applied to the borrowed data from the four experimental studies (Abich et al., 2017; Barber et al., 2018; Barber et al., 2019; Bendell et al., 2020; Kopinsky, 2017) under the RCTA (Childers et al., 2016). The used data is summarized in **Table 6**.

Software

The program R (R Core Team, 2013) was used for the analyses. Basic analyses were conducted with the user interface R Commander (Fox & Bouchet-Valat, 2019). More advanced analyses and visualizations were conducted with GGPlot2 (Wickham, 2016, 2016), Tidyverse (Wickham, 2017), Hmisc (Harrell, 2019), GGally (Schloerke et al., 2017), Betareg (Zeileis et al., 2019), and Candisc (Friendly et al., 2017). The package dominanceanalysis (Bustos & Countinho, 2019) was used and updated as part of the present study in collaboration with the author of the package.

Operationalization of Constructs per Study

Next, each of the studies is described with operationalization of the constructs in light of the proposed Core model (**Figure 2**).

Study A

Experimental data from Study A was borrowed from the RCTA (Childers et al., 2016; Kopinsky, 2017; IRB in APPENDIX G: IRB FOR BORROWED EXPERIMENTAL STUDY A). For a full description of this study, see APPENDIX A: DESCRIPTION BORROWED STUDY A. This study was a mixed design, with visual complexity (of the signal detection display and icons) as a between-subjects variable (two levels: low vs. high) and agent type as a within-subjects variable (two levels: legged vs. wheeled; APPENDIX A: DESCRIPTION BORROWED STUDY A). The order of presentation was coded; order of agent type presentation was counterbalanced and randomized in Study A. Each task duration was approximately 10 minutes.

Since there were repeated measures in the data set, a repeated-measures check was conducted. There was no significant difference between first and second instance, Welch' *F*(1, 128.95) = 1.16, *p* = .284. A total of *N* = 134 observations was maintained in the dataset.

Operationalization of the Core Constructs

Task Performance: Hit Rate

Task performance was operationalized as hit rate: the number of correctly detected threats divided by the number of available threats. Average hit rate was 0.97 (*SD* = 0.04), within the accepted performance standards imposed by the military (e.g., Naval Education and Training Command, 2009). The boxplot and histogram indicated a non-normal distribution of the data (**Figure 3**).

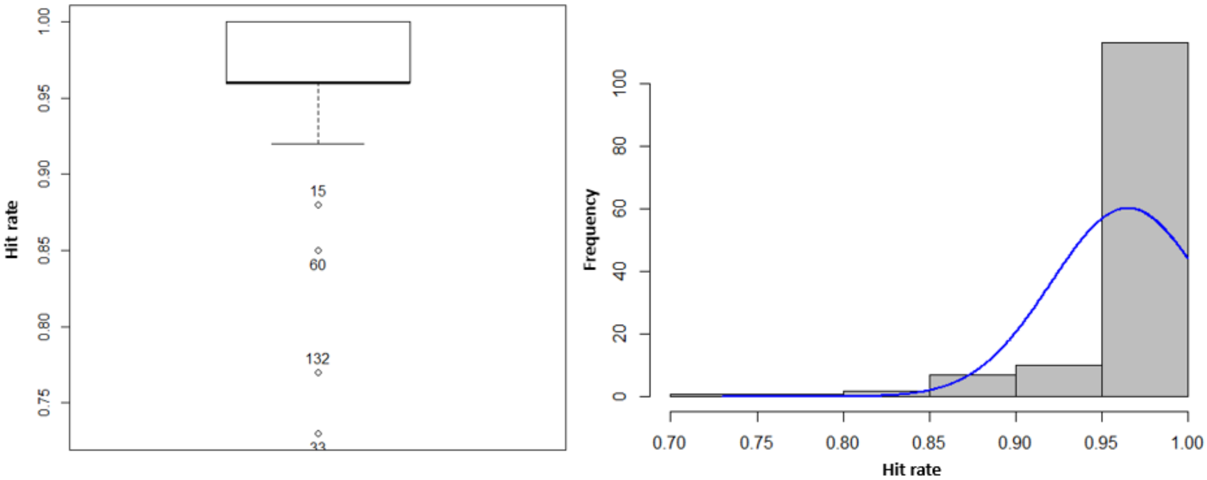


Figure 3. Distribution of hit rate taken from Study A.

Note. Hit rate was non-normally distributed in Study A, as identified in the boxplot and histogram with density plot. There was a negative skew in the data.

The non-normal distribution was related to a measurement scaling issue, as the response variable was naturally double bounded between a minimum value of 0.00 and maximum value of 1.00 (Smithson & Merkle, 2013). For these types of measurements, i.e., based on rates, the data follows a Beta rather than a Gaussian distribution (Ferrari & Cribari-Neto, 2004). The response variable y in Beta distributions is bounded, $0 \leq y \leq 1$, and the shape parameters are $\alpha > 0$ and $\beta > 0$, with a density function described as (Ferrari & Cribari-Neto, 2004)

$$f(y|\alpha,\beta) = \frac{(y)^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \qquad (19)$$

where Γ denotes the gamma function. Depending on the values of α and β, the distribution can have different shapes. As $0 \leq y \leq 1$ for Beta distributions, all hit rate values equal to 1.00 were winsorized to 0.995, creating a new response variable “winsorized hit rate”.

Furthermore, since a number of observations were >3 *SD*, these values were winsorized to the minimal value of 3 *SD* (0.85). See **Figure 4** for the distribution of winsorized hit rate.

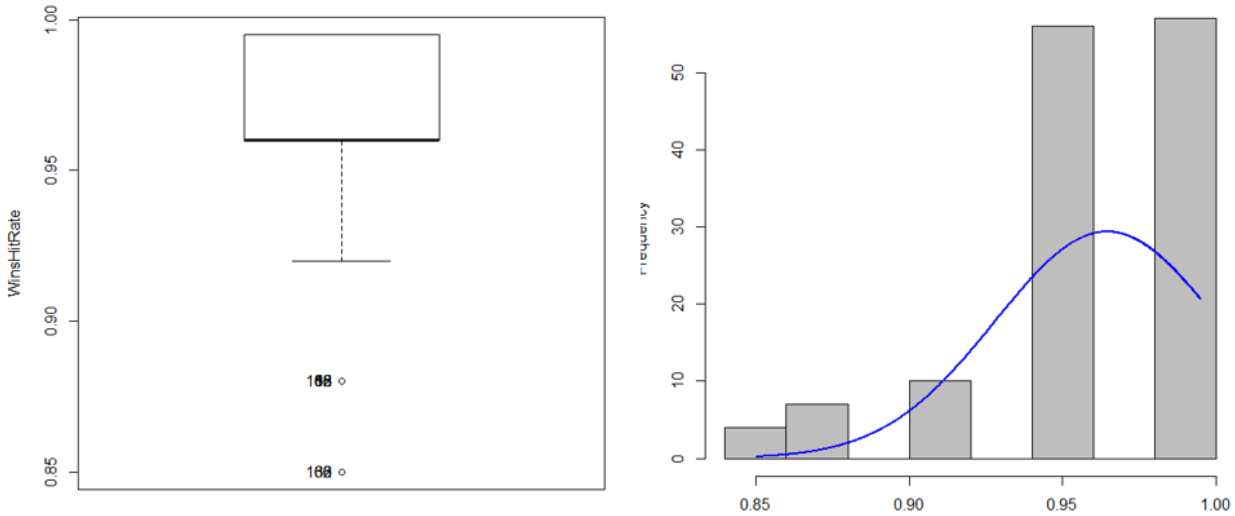


Figure 4. Distribution of winsorized hit rate taken from Study A.

Note. The distribution of winsorized hit rate is shown in this boxplot and histogram with density curve. Observations of hit rate equal to 1.00 were winsorized to 0.995 (Ferrari & Cribari-Neto, 2004). Observations < 3 *SD* of the mean were winsorized to the value of 3 *SD* of the mean. The remaining outliers were not due to technical or otherwise identifiable errors and were maintained.

Human/Agent Qualities

Of the human qualities, age, gender, military experience, and video gaming experience were included. The simulated agent was manipulated to be presented as legged or wheeled, yet was otherwise simulated to be a fully autonomous, 100% reliable, intelligent, and embodied. The agent scouted the outer cordon for threats and contraband. As within-subjects variable, agent morphology type was manipulated as legged (zoomorphic) or wheeled (machine-like).

Participants were recruited from the University of Central Florida's undergraduate psychology pool in exchange for course credit (*N* = 67; IRB in APPENDIX G). The characteristics of the sample in Study A are presented in **Error! Reference source not found.**.

Table 7

Sample characteristics Study A, as conducted by (Kopinsky, 2017).

Sample Categorization	*N*	Age (*M, SD)*	Video gaming experience (*M, SD)*
Male			
Student	65	20.66 (5.20)	4.85 (1.09)
Military	4	40.50 (12.12)	4.00 (2.31)
Overall	69	21.81 (7.33)	4.43 (1.70)
Sample Categorization	*N*	Age (*M, SD)*	Video gaming experience (*M, SD)*

Sample Categorization	*N*	Age (*M, SD)*	Video gaming experience (*M, SD)*
Female			
Student	65	21.11 (5.51)	2.60 (1.38)
Military	0	NA	NA
Overall	65	21.11 (5.51)	2.60 (1.38)
Overall			
Student	130	20.88 (5.34)	3.72 (1.68)
Military	4	40.50 (12.12)	4.00 (2.31)
Overall	134	21.47 (6.50)	3.86 (2.00)

Note. This table shows the sample characteristics including the number of observations (*N*), age (mean, standard deviation), and video gaming experience (**Table 47**), per gender (male, female) and military experience (student, military). The military participants were significantly older than students, Welch' *F*(1, 3.04) = 10.41, *p* = .048. Men played video games significantly more frequent then women, Welch' *F*(1, 126.37) = 97.50, *p* < .0001.

Task Perception

Task perception was operationalized as perceived workload as measured with the NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988), containing six subscales. A description of the subscales is found in **Error! Reference source not found.Error! Reference source not found.**. The rating scales range from 0 to 100 (see APPENDIX E: NASA-TLX). The performance subscale traditionally needed rescoring but was adapted to account for this (see APPENDIX E: NASA-TLX).

Table 8

Description of NASA-TLX subscales.

Scale	Description
Mental demand	The amount of mental and perceptual activity required during the task
Physical demand	The amount of physical activity required during the task
Temporal demand	The amount of experienced time pressure due to rate or pace of the task (elements)
Frustration	The amount of experienced frustration during the task
Effort	The amount of experienced (mental and physical) effort to accomplish the level of performance
Performance	A rating of how successful you perceived you were in accomplishing the task to standard

Note. Adapted from (Hart & Staveland, 1988, p. 32).

In Study A, the NASA-TLX was offered after each scenario. The average scores on the NASA-TLX subscales are shown in **Error! Reference source not found.**. The highest mean score was found for mental demand and the lowest score for performance. However, for each of the subscales the standard deviation was high relative to the mean. Such a high variability complicated interpretation of the scales.

Table 9

Average of NASA-TLX scores taken from Study A (Kopinsky, 2017).

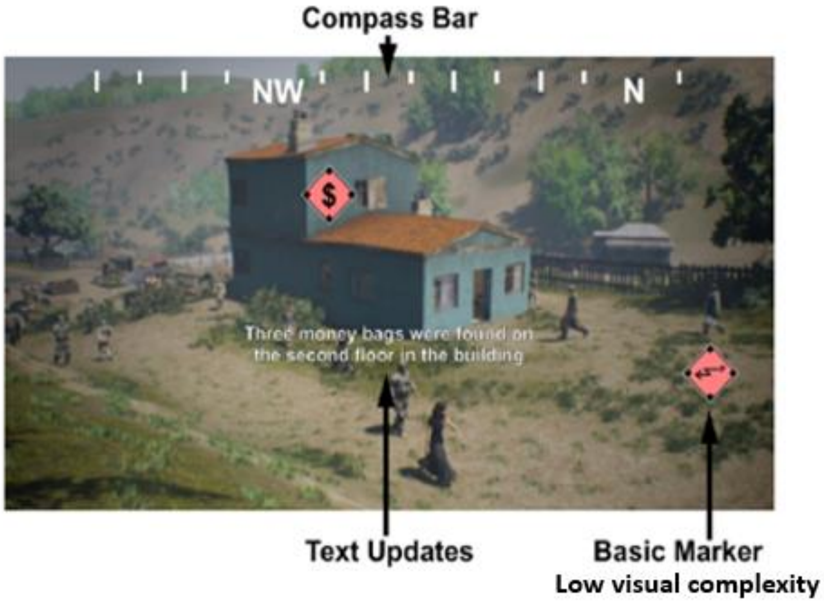
NASA-TLX Scale	Mean	SD
Global	33.23	19.15
Mental Demand	56.68	30.18
Physical Demand	30.34	23.63
Temporal Demand	33.58	26.33
Effort	44.18	27.98
Frustration	28.81	27.32
Performance	15.82	19.45

Note. In study A, the highest mean score was found for mental demand and the lowest score for performance. However, for each of the subscales the standard deviation was high relative to the mean, complicating interpretation of the scales.

Task Composition

The ongoing threat detection task was conducted at an event rate of 15 characters/minute on screen, with a signal probability of 0.13-0.14. During this task, the autonomous agent sent reports to the participant with information of what it found and where this was found.

Visual Complexity. The agent reports were sent visually, wherein the visual complexity of the report was manipulated as between-group variable. In low visual complexity (**Figure 5**), participants saw the visual report with a Compass bar and symbols identifying what was found.



Three money bags were found on the second floor in the building

Figure 5. Low visual complexity condition in Study A (Kopinsky, 2017; Copyright in APPENDIX K: COPYRIGHT).

Note. This figure shows the gaming environment in Study A during the low visual complexity condition. Agent reports are sent to the participant visually through text updates, a compass bar, and a symbol or marker with basic elements. These symbols are identifiers of what the simulated agent found.

In high visual complexity, the symbols were enhanced with the quantity of the items that were found. Additionally, a minimap was offered to provide an additional view of the location of the found items within the environment (**Figure 6**).

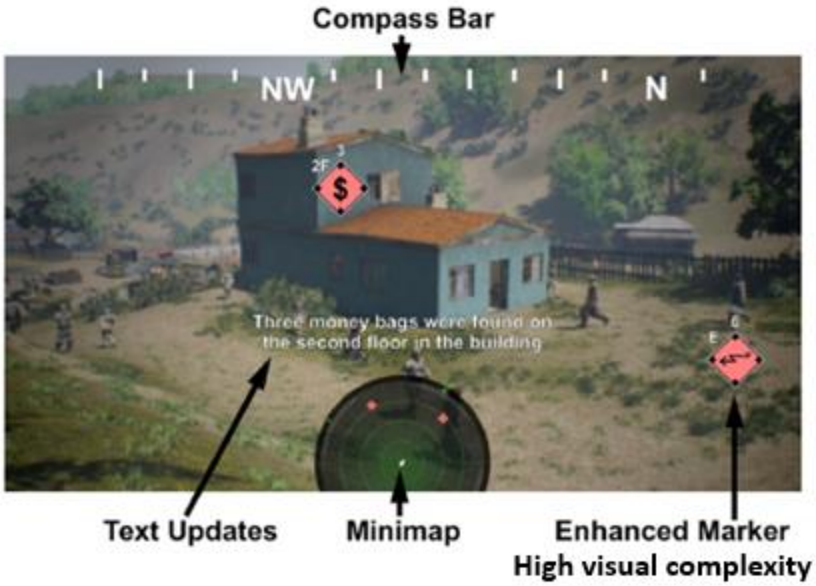

Three money bags were found on the second floor in the building

Figure 6. High visual complexity condition in Study A (Kopinsky, 2017; copyright in APPENDIX K: COPYRIGHT).

Note. This figure shows the gaming environment in Study A during the low visual complexity condition. Agent reports are sent to the participant visually through text updates, a compass bar, a minimap showing the agent's location, and a symbol or marker with enhanced elements. These symbols are identifiers of what the simulated agent found.

Study B

A full description of Study B is in APPENDIX B: DESCRIPTION BORROWED STUDY B, with the IRB in APPENDIX H: IRB FOR BORROWED EXPERIMENTAL STUDY B. Study B consisted of three within-subject conditions (**Figure 7**). Participants actively pulled agent reports, but under constant or changing event rate (B.1 and B.2 conditions). There was an additional condition (B.3 condition) that was conducted under constant event rate, wherein participants received agent reports. Each condition lasted approximately 10 minutes.

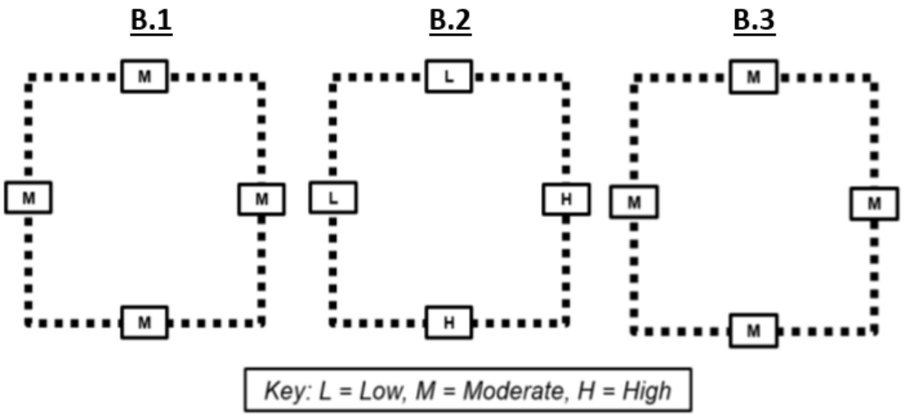


Figure 7. Experimental design of Study B (Abich et al., 2017; Barber et al., 2018).

Note. Study B consisted of three conditions: B.1 was conducted at constant medium event rate (30 characters/minute), B.2 was conducting at a changing event rate (low: 15 characters/minute, high: 60

characters/minute), and B.3 was a different reporting task than B.1 and was conducted at a constant medium event rate.

Since the overall dataset (*N* = 332) had repeated measures, a check was conducted. The repeated occurrences were not significantly different, Welch' *F*(2, 123.92) = 0.57, *p* = .566, thus the observations were all maintained.

Operationalization of the Core Constructs

Task Performance: Hit Rate

Task performance was operationalized as hit rate: the number of correctly detected threats divided by the number of available threats. Average hit rate was 0.95 (*SD* = 0.07), within the accepted performance standards imposed by the military (e.g., Naval Education and Training Command, 2009). The boxplot and histogram indicated a non-normal distribution of the data with a number of outliers (**Figure 8**).

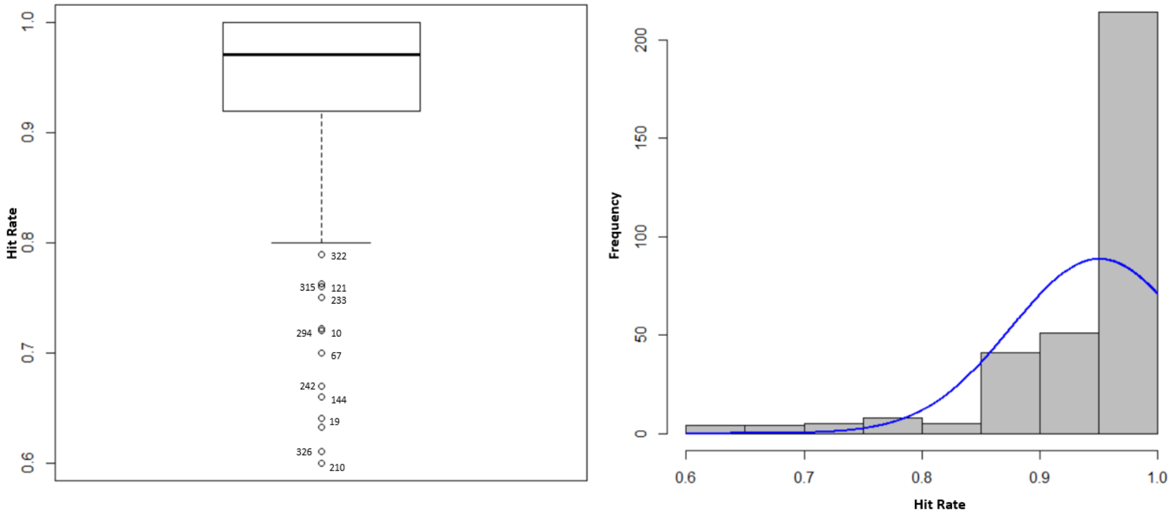


Figure 8. Distribution of hit rate taken from Study B.

Note. Hit rate was non-normally distributed in Study B, as identified in the boxplot and histogram with density plot. There was a negative skew in the data.

The distribution was a beta-distribution (Ferrari & Cribari-Neto, 2004), with $0 \leq y \leq 1$. Therefore, all hit rate values equal to 1.00 were winsorized to 0.995, creating a new response variable "winsorized hit rate". The outliers > 3 *SD* were winsorized to a minimal acceptable value of 0.74 (= 3 *SD*). The distribution of winsorized hit rate is shown in **Figure 9**.

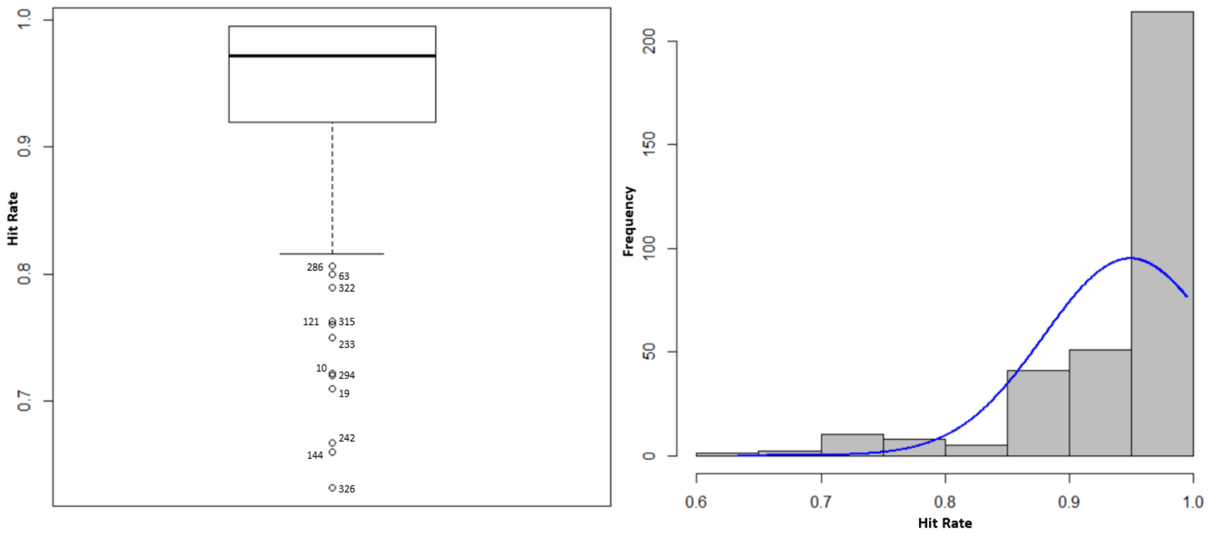


Figure 9. Distribution of winsorized hit rate taken from Study B.

Note. The distribution of winsorized hit rate is shown in this boxplot and histogram with density curve. Observations of hit rate equal to 1.00 were winsorized to 0.995 (Ferrari & Cribari-Neto, 2004). Observations < 3 *SD* of the mean were winsorized to the value of 3 *SD* of the mean. The remaining outliers were not due to technical or otherwise identifiable errors and were maintained.

Human/Agent Qualities

This study did not manipulate agent variables. The out-of-sight agent was simulated to be a fully autonomous and 100% reliable, intelligent, and embodied that scouted the outer cordon for threats and contraband. Of the human qualities, age, gender, military experience, and video gaming experience were included.

Two samples were utilized in this study. One sample were undergraduate students from the University of Central Florida (*N* = 56), that were recruited through the Psychology resource pool for course credit (IRB in APPENDIX H: IRB FOR BORROWED EXPERIMENTAL

STUDY B). The other sample were Soldiers from Ft. Benning’s officer school (*N* = 29, IRB in APPENDIX H). Soldiers volunteered and did not receive compensation for their participation. The characteristics of the sample in Study B are summarized in **Error! Reference source not found.**.

Table 10

Sample characteristics Study B (Abich et al., 2017; Barber et al., 2018).

Sample Categorization	*N*	Age (*M, SD)*	Video Gaming Experience (*M*, *SD*)
Male			
Student	295	19.95 (1.83)	4.74 (1.09)
Military	82	26.70 (3.41)	NA
Overall	213	22.55 (4.16)	4.74 (1.09)
Female			
Student	91	21.30 (5.55)	3.52 (1.52)
Military	28	26.71 (3.02)	NA
Overall	119	22.57 (5.56)	3.52 (1.52)
Overall			
Student	222	20.50 (3.87)	4.23 (1.41)
Military	110	26.70 (3.30)	NA
Overall	332	22.56 (4.70)	4.23 (1.41)

Note. This table shows the sample characteristics including the number of observations (*N*), age (mean, standard deviation), and video gaming experience (**Table 44**), per gender (male, female) and military experience (student, military). The military participants were significantly older than students, Welch’ *F*(1, 250.47) = 230.41, *p* < .0001. Male and female participants did not differ significantly in age, Welch’ *F*(1, 192.86) = 0.00, *p* = .970. Men played video games significantly more frequent then women, Welch’ *F*(1, 76.90) = 21.65, *p* < .0001.

Task Perception

The NASA-TLX (Hart & Staveland, 1988) was presented every 2.5 minutes in the B.1 and B.2 conditions. The B.3 condition did not have a NASA-TLX administration. The average of the perceived workload scales is presented in **Error! Reference source not found.**. The average highest score (mental demand) was below 50. The lowest score was found for physical demand. However, for each of the subscales the standard deviation was high relative to the mean. Such a high variability complicated interpretation of the scales.

Table 11

Average NASA-TLX scores taken from Study B.

NASA-TLX Scale	Mean	SD
Effort	35.67	27.52
Frustration	23.82	24.07
Mental Demand	40.65	31.55
Performance	22.10	23.73
Physical Demand	14.81	18.19
Temporal Demand	31.74	27.83
Global	28.14	19.87

Note. In study B, the highest score was for mental demand. However, for each of the subscales the standard deviation was high relative to the mean., complicating interpretation of the scales.

Task Composition

Event Rate. Event rate was manipulated as a within-subjects variable (Abich et al., 2017; Barber et al., 2018). In condition B.1 the ongoing threat detection task had a constant number of characters on screen per minute, which was set at 30 characters/minute. In B.2, the event rate changed halfway during the scenario. Half of the scenario ran in a low event rate, with 15 characters/minute, while the remainder ran in a high event rate, with 60 characters/minute. The order of the event rate shift, either from low-to-high or high-to-low, was counterbalanced within the design. Furthermore, a third condition, B.3, was present that was conducted at a medium event rate (30 characters/minute), wherein participants received agent reports. Signal probability across the three conditions was 0.12-0.13.

Task Type. As mentioned, in two conditions participants actively pulled agent reports, while in a third condition, participants received reports (Abich et al., 2017; Barber et al., 2018). In the pull-condition, Participants could pull a report from the agent teammate that contained information about the number of threats (critical, non-critical, and non-targets). A multimodal interface (MMI) could be brought up and a report was requested by clicking on text or image.

The information displayed in either report was identical. In the image report, boxes were shown around threats and critical threats, while the text report showed the number of threats, critical threats, and non-threats (not needed for probes). Participants also had the freedom to pull text and image reports sequentially.

In condition B.3, wherein participants received a report, the report was an assistance request from the agent that asked the participant to make a decision for them (A or B).

Study C

A full description of Study C is in APPENDIX C: DESCRIPTION BORROWED STUDY C, with the IRB in APPENDIX I: IRB FOR BORRWED EXPERIMENTAL STUDY C. In Study C, two within-subject factors were manipulated over three scenarios, that each lasted approximately 32 minutes (**Figure 10**; Barber et al., 2019). Event rate was manipulated as low (15 characters/minute on screen) vs. high (60 characters/minute), wherein the rate changed every eight minutes. An exception in this design, are the first and last blocks; these only lasted four minutes.

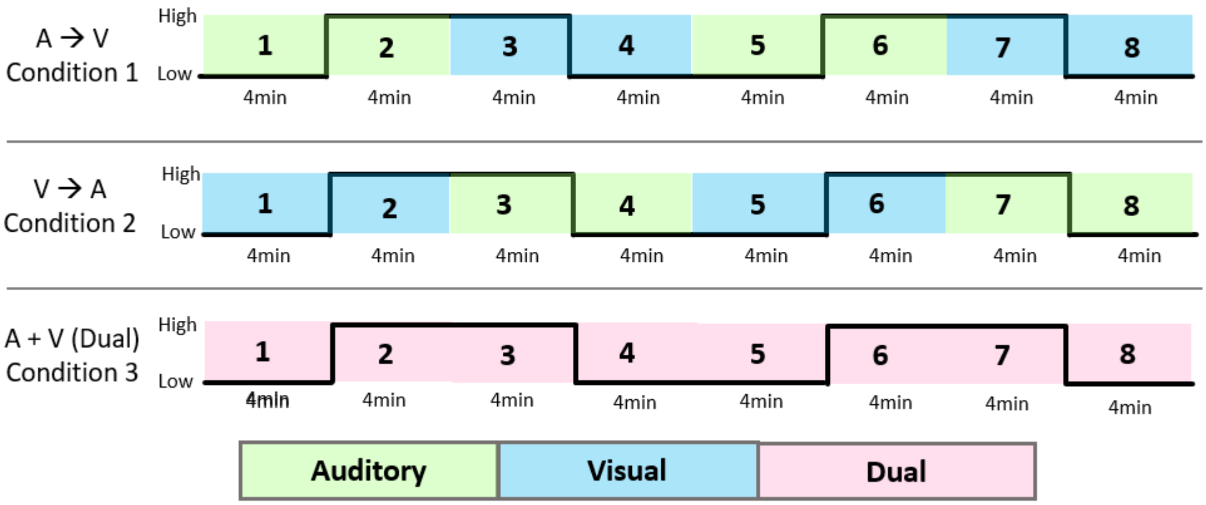


Figure 10. Experimental design of Study C (Barber et al., 2019).

Note. Study C consisted of three conditions, that all participants participated in. In condition 1 and 2 participants received reports from a simulated agent through a single modality, wherein the modality changed between auditory and visual (single adaptive modality). In condition 3 agent reports were sent through both modalities simultaneously (dual). Event rate changed within the conditions from low (15 characters/minute) to high (60 characters/minute).

Furthermore, agent report modality was manipulated between conditions. Condition 1 and 2 were both single-adaptive modalities, wherein condition 1 started in auditory modality and condition 2 started in visual modality. In the third condition, the reports were sent in two modalities simultaneously: auditory plus visual.

Since participants ran through all three conditions, a repeated-measures check was conducted. The three conditions were not significantly different, Welch' $F(2, 80.96) = 0.34$, $p = 0.713$. The sample contained $N = 126$ observations.

Operationalization of the Core Constructs

Task Performance: Hit Rate

Task performance was operationalized as hit rate: the number of correctly detected threats divided by the number of available threats. Average hit rate was 0.95 (*SD* = 0.07), within the accepted performance standards imposed by the military (e.g., Naval Education and Training Command, 2009). The boxplot and histogram indicated a normal distribution with a negative skew, with a number of outliers (**Figure 11**).

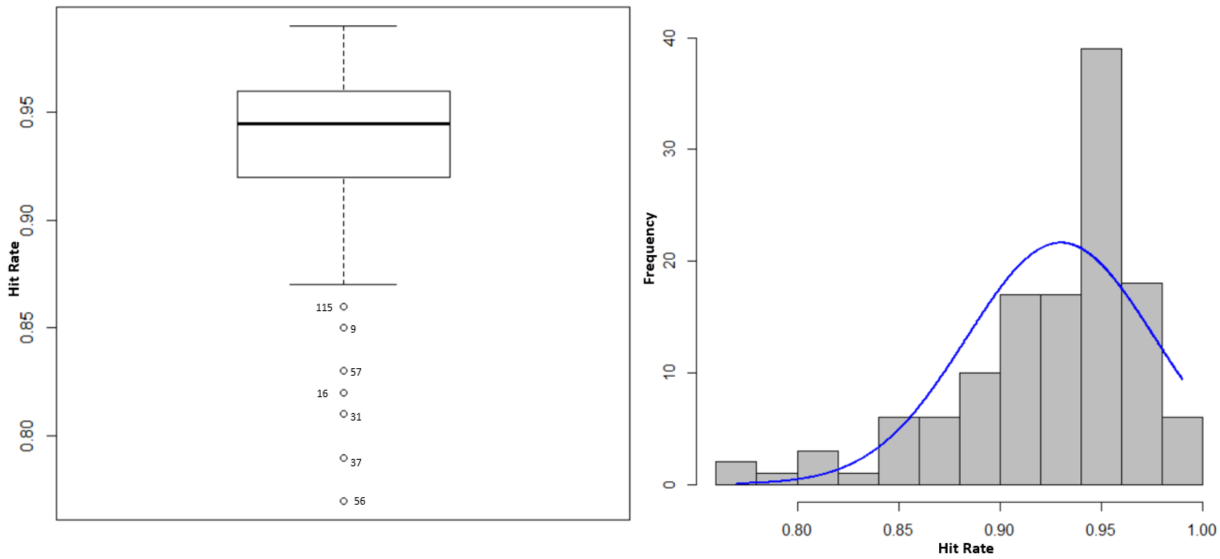


Figure 11. Distribution of hit rate taken from Study C.

Note. Hit rate was approximately normally distributed in Study A, but with a negative skew, as identified in the boxplot and histogram with density plot.

To fit the beta distribution (Ferrari & Cribari-Neto, 2004), hit rate was winsorized to a highest value of 0.995 and lowest value of 0.75 (= 3 *SD*), see **Figure 12** for the distribution. Trimming the outliers resulted in a more non-normal distribution.

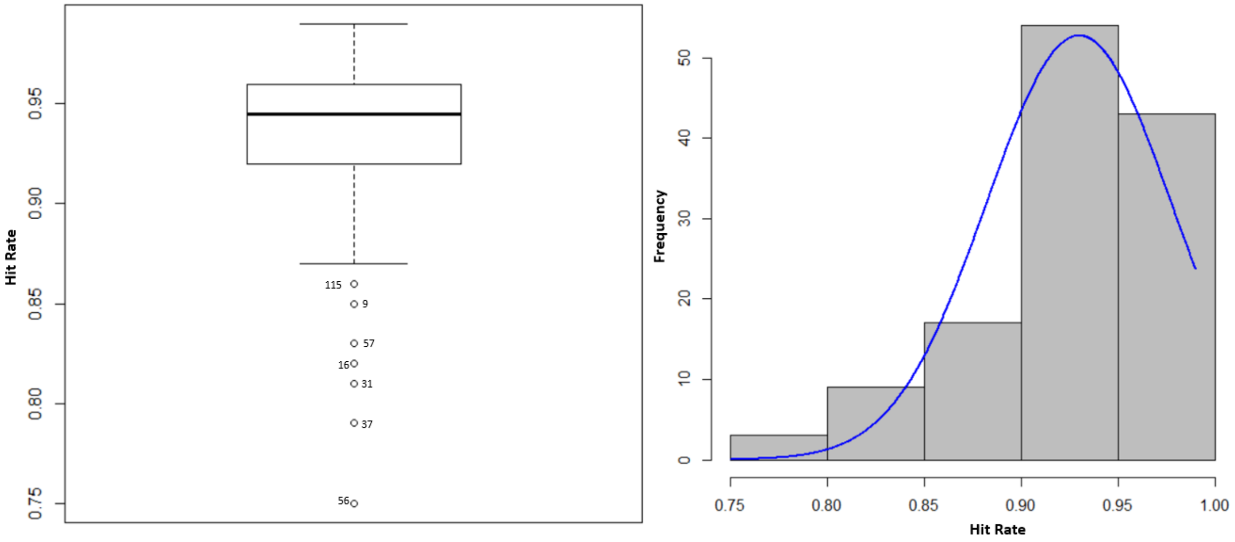


Figure 12. Distribution of winsorized hit rate taken from Study C.

Note. The distribution of winsorized hit rate is shown in this boxplot and histogram with density curve. Observations of hit rate equal to 1.00 were winsorized to 0.995 (Ferrari & Cribari-Neto, 2004). Observations < 3 *SD* of the mean were winsorized to the value of 3 *SD* of the mean. Trimming the outliers resulted in a more non-normal distribution. The remaining outliers were not due to technical or otherwise identifiable errors and were maintained.

Human/Agent Qualities

Study C did not manipulate agent variables. The out-of-sight agent was simulated to be a fully autonomous and 100% reliable, intelligent, embodied agent that scouted the outer cordon for threats and contraband. Of the human qualities, age, gender, military experience, and video gaming experience were examined. Participants were recruited from the University of Central Florida's undergraduate psychology pool in exchange for course credit (*N* = 42; IRB in APPENDIX I). Sample characteristics are presented in **Error! Reference source not found.**.

Table 12

Sample characteristics in Study C.

Sample Categorization	*N*	Age (M, SD)	Video gaming experience (*M, SD)*
Male			
Student	73	18.92 (2.40)	4.25 (1.37)
Military	3	20.00 (0.00)	5.00 (0.00)
Overall	75	18.96 (2.36)	4.28 (1.35)
Female			
Student	51	19.41 (2.16)	2.71 (1.65)
Military	0	NA	NA
Overall	51	19.41 (2.16)	2.71 (1.65)
Overall			
Student	123	19.12 (2.31)	3.61 (1.67)
Military	3	20.00 (0.00)	5.00 (0.00)
Overall	126	19.14 (2.28	3.64 (1.67)

Note. This table shows the sample characteristics including the number of observations (*N*), age (mean, standard deviation), and video gaming experience (**Table 44**), per gender (male, female) and military experience (student, military). One participant had military experience, leading to three observations. Men were not significantly older than women, Welch' *F*(1, 113.32) = 1.23, *p* = .270. Men played video games significantly more frequent then women, Welch' *F*(1, 92,82) = 31.80, *p* < .0001.

Task Perception

The NASA-TLX was conducted after each condition. The average scores are shown in **Error! Reference source not found.**. The highest scores were found for mental demand and effort, and the lowest score for physical demand. However, for each of the subscales the standard deviation was high relative to the mean. Such a high variability complicated interpretation of the scales.

Table 13

Average NASA-TLX scores taken from Study C.

NASA-TLX Scale	Mean	SD
Effort	77.21	21.29
Frustration	65.33	27.83
Mental Demand	86.50	14.62
Performance	51.08	28.29
Physical Demand	30.00	32.62
Temporal Demand	62.83	26.49
Global	62.17	15.87

Note. In study C, the highest mean scores were found for mental demand and effort. However, for each of the subscales the standard deviation was high relative to the mean, complicating interpretation of the scales.

Task Composition

Event Rate. Event rate was manipulated as low versus high. The blocks that were similar in their manipulations (e.g., auditory + low (Condition 1, block 5 and Condition 2, block 4)) could not be combined as significant differences were found (Barber et al., 2019). Thus, in the present effort, event rate was coded as constant (Condition 3) versus changing (Condition 1 and 2). Signal probability was 0.12-0.13.

Agent Report Modality. Since the blocks could not be combined, agent report modality was coded as single-adaptive (Condition 1 and 2) or dual (Condition 3).

Study D

A full description of Study D is in APPENDIX D: DESCRIPTION BORROWED STUDY D, with the IRB in APPENDIX J: IRB FOR BORROWED EXPERIMENTAL STUDY D. This study employed a mixed design, wherein two two-level factors were manipulated (Bendell et al., 2020). Each participant experienced two sensory modalities of agent report delivery (visual text vs. auditory speech) in two separate scenarios, each lasting approximately 16 minutes. The between-subjects variable was the timing of agent report delivery. Reports could be delivered regularly every minute (Condition D.1) or immediately, which was irregular (Condition D.2). This order for the scenarios was randomized and counterbalanced.

A repeated-measures check indicated that the two instances of the same participant did not significantly affect hit rate, Welch' *F*(1, 114.41) = 0.35, *p* = .556. The total analyzable sample was *N* = 117.

Operationalization of the Core Constructs

Task Performance: Hit Rate

Task performance was operationalized as hit rate: the number of correctly detected threats divided by the number of available threats. Average hit rate was 0.67 (*SD* = 0.11), well below the accepted performance standards imposed by the military (e.g., Naval Education and Training Command, 2009). The boxplot and histogram indicated an approximate normal distribution of the data with a number of outliers (**Figure 13**).

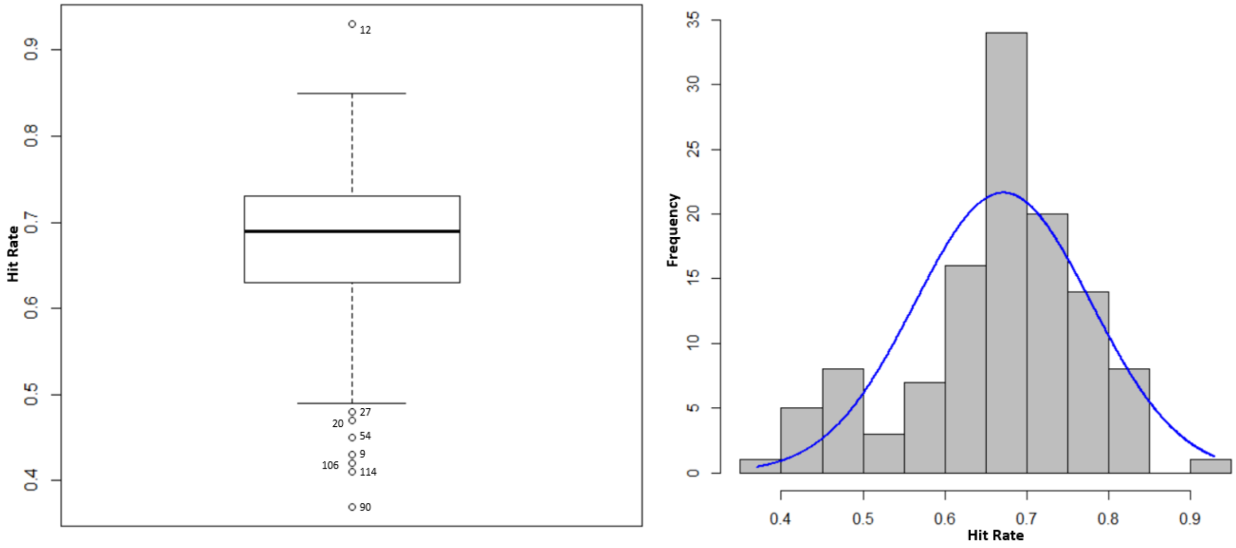


Figure 13. Distribution of hit rate taken from Study D.

Note. Hit rate was approximately normally distributed in Study D, as identified in the boxplot and histogram with density plot.

The data was more normally distributed, but still contained in a beta distribution due to the double-bounded response variable (Smithson & Merkle, 2013). Any hit rate values of 1.00 were winsorized to 0.995. The lower minimal value was winsorized to 3 *SD* of the mean (0.34). The distribution of winsorized hit rate was similar to the original variable (**Figure 14**).

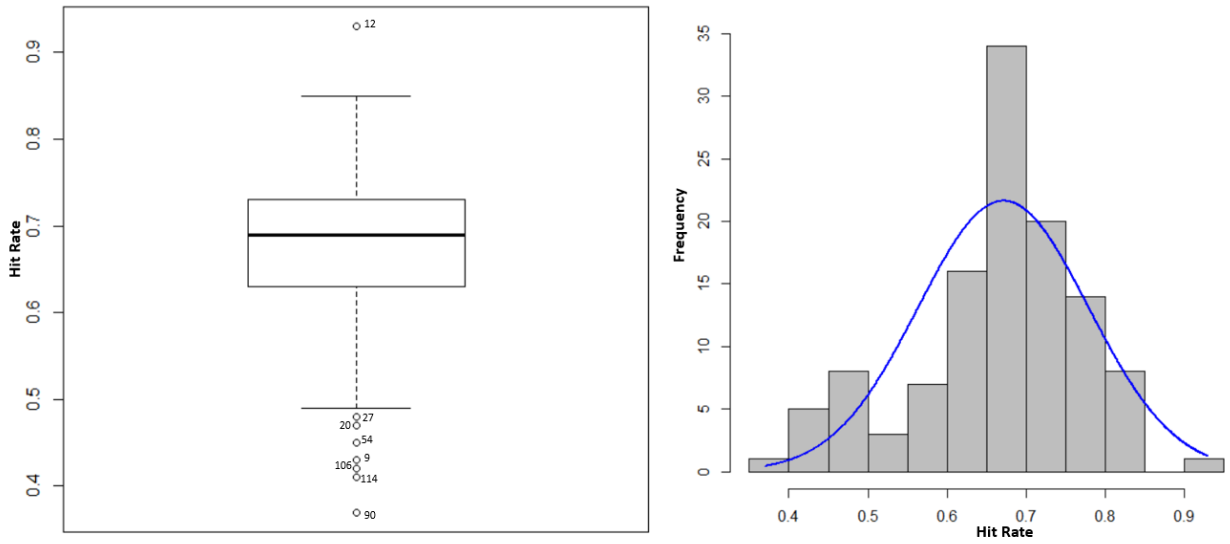


Figure 14. Distribution of winsorized hit rate taken from Study D.

Note. The distribution of winsorized hit rate is shown in this boxplot and histogram with density curve. Observations of hit rate equal to 1.00 were winsorized to 0.995 (Ferrari & Cribari-Neto, 2004). Observations < 3 *SD* of the mean were winsorized to the value of 3 *SD* of the mean. The distribution of winsorized hit rate was similar to the distribution of hit rate. The remaining outliers were not due to technical or otherwise identifiable errors and were maintained.

Human/Agent Qualities

This study did not manipulate agent variables (study description in APPENDIX D: DESCRIPTION BORROWED STUDY D, IRB in APPENDIX J: IRB FOR BORROWED EXPERIMENTAL STUDY D; Bendell et al., 2020). The out-of-sight agent was simulated to be a fully autonomous and 100% reliable, intelligent, and embodied that scouted the outer cordon for threats and contraband. Of the human qualities, age and gender were included. Participants were recruited from the University of Central Florida's undergraduate psychology pool in exchange for course credit (*N* = 59; IRB in APPENDIX J). None of the participants reported military experience and there was no data for video gaming experience. The sample characteristics are presented in **Error! Reference source not found.**.

Table 14

Sample characteristics in Study D (Bendell et al., 2020).

Sample Categorization	N	Age (*M*, *SD*)
NA	2	NA
Male		
Student	57	20.07 (4.95)
Military	0	NA
Overall	57	20.07 (4.95)
Female		
Student	58	19.24 (1.73)
Military	0	NA
Overall	58	19.24 (1.73)
Overall		
Student	117	19.65 (3.70)
Military	0	NA
Overall	117	19.65 (3.70)

Note. This table shows the sample characteristics including the number of observations (*N*) and age (mean, standard deviation) per gender (male, female) and military experience (student, military). This sample had no military experience. Men and women did not significantly differ in age, Welch' *F*(1, 69.28) = 1.43, *p* = .236.

Task Perception

The NASA-TLX was administered after each condition. The average scores are shown in **Error! Reference source not found.**. The highest score was for mental demand, followed by performance. Physical demand was the lowest score. However, for each of the subscales the standard deviation was high relative to the mean. Such a high variability complicated interpretation of the scales.

Table 15

Average NASA-TLX scores in Study D.

NASA-TLX Scale	Mean	SD
Effort	53.46	24.52
Frustration	34.96	28.04
Mental Demand	67.69	22.99
Performance	60.51	24.80
Physical Demand	16.32	16.91
Temporal Demand	45.38	25.22
Global	46.40	13.95

Note. In study D, the highest mean score was found for mental demand and performance. However, for each of the subscales the standard deviation was high relative to the mean, complicating interpretation of the scales.

Task Composition

Event rate was constant at 60 characters/minute on screen, at a threat probability of .09-.10 (Bendell et al., 2020).

Agent Report Modality. To ensure reports were attended to, an auditory tone alerted participants one second prior to release of each report (Bendell et al., 2020). There were 30 non-critical reports that contained information pertaining to the route, such as obstacles encountered. Four reports were critical and included an IED image review request. Report review was possible by clicking a button on the controller to pull up the MMI. They could raise the controller to bring the MMI up or keep the controller down to look down at the simulated MMI. The modality through which reports were delivered was auditory or visual. In the auditory condition, all non-critical reports were sent through speech alone. Critical IED review requests were still sent visually, as these required visual inspection. Contrary, in the visual report condition, all reports were solely transmitted through the MMI.

Agent Report Delivery Frequency. The delivery frequency of the agent reports was manipulated (Bendell et al., 2020). They could be delivered every minute (interval; Condition D.1), or immediately (Condition D.2).

CHAPTER FOUR: RESULTS

The hypotheses were tested for each of the studies against the proposed Core model (**Figure 2**), as separate falsifications of the model using the approach discussed in Chapter 3.

Study A

Hypotheses Study A

Study A manipulated agent (morphology) type (legged vs. wheeled) and visual complexity of the markers in the agent reports (basic vs. complex). The threat detection task was conducted under a low event rate of 15 characters/minute and high threat probability of 0.13-0.14. Agent qualities were not available. The collaborative agent was simulated to be a fully autonomous and 100% reliable, intelligent, embodied agent that scouted the outer cordon for threats and contraband. The predictors that were available in study A were tested against the null hypotheses of the Core model. This is represented in **Figure 15**.

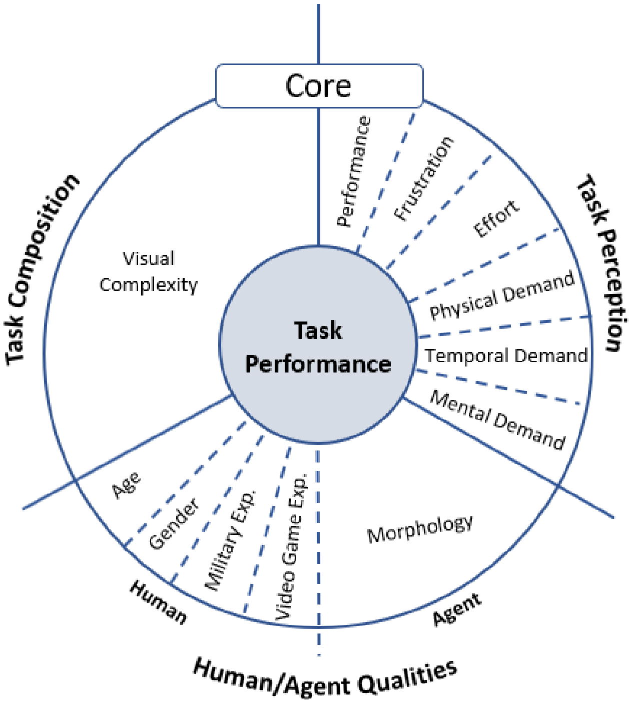


Figure 15. Visual representation of null hypotheses in Study A.

Note. The size of the sections of the pie represent the relative importance of the factor to task performance. The null hypotheses are that all factors and categories are of equal importance to task performance.

The null hypotheses are as follows:

Hypothesis 1. Of the Human/Agent Qualities, human and agent factors are equally important to task performance (hit rate).

Hypothesis 2. All NASA-TLX subscales (Task Perception) contribute equally to task performance.

Hypothesis 3. Task Composition, Perception of Task, and Human/Agent Qualities are equally important to task performance.

Dominance Analysis Study A

Linearity was established between predictors and the response variable (Appendix F, **Error! Reference source not found.**).

Human/Agent Qualities

Overall, the variables were not strong in predicting hit rate, since the average additional contribution of each predictor was very low (**Table 48**, Appendix F).

Complete Dominance

Agent type completely dominated all other predictors, as shown in Error! Reference source not found.. Age also dominated gender, video gaming experience, and military experience across all subset model sizes (*k*). Lastly, video gaming experience dominated gender.

Table 16

Complete dominance results Human/Agent Qualities Study A.

Variable	Agent Type	Age	Video Gaming Experience	Military Experience	Gender
Agent Type	0.5	1	1	1	1
Age	0	0.5	1	1	1
Video Gaming Experience	0	0	0.5	0.5	1
Military Experience	0	0	0.5	0.5	0.5
Gender	0	0	0	0.5	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

Conditional Dominance

Since agent type and age completely dominated the other predictors, they also conditionally dominated them **Figure 16**. The unique additional contribution of agent type remained fairly stable regardless of subset model size. However, the unique contribution of age decreased considerably as the subset model size increased. The additional contribution of gender and video gaming experience increased with subset model size, which was an indicator that these variables were potential suppressors. A suppressor variable improves prediction due to its collinearity with other predictors, rather than through a direct correlation with the response variable (Azen & Budescu, 2003; Smith et al., 1992).

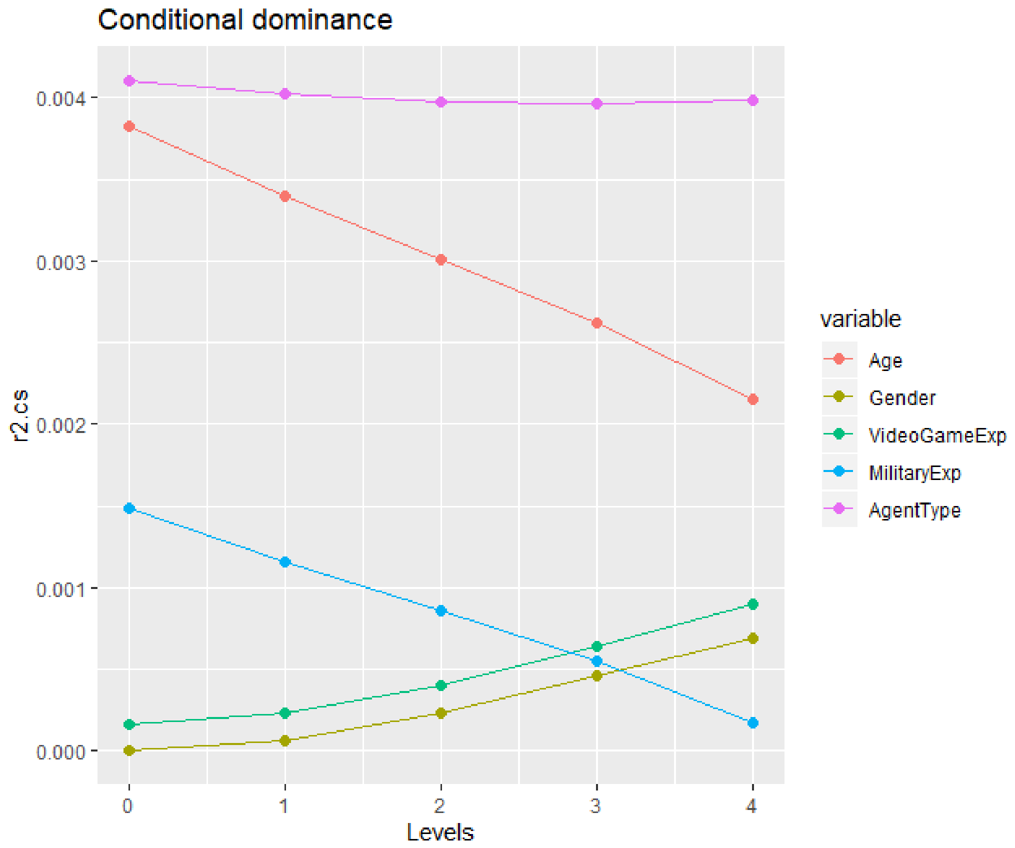


Figure 16. Conditional dominance results Human/Agent Qualities Study A.

Note. The plot shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model. The unique contribution of a predictor should monotonically decrease with increasing subset model sizes (Azen & Budescu, 2003; Budescu, 1993). An increase, such as here for video gaming experience and gender, indicates that these variables are potential suppressors, gaining importance through collinearity with other predictors in the model rather than through direct association with the outcome variable (Azen & Budescu, 2003).

General Dominance

As shown in **Figure 17**, gender did not (generally) dominate any other predictor. Aside from the completely dominating variables age and agent type, military experience generally dominated video gaming experience and gender.

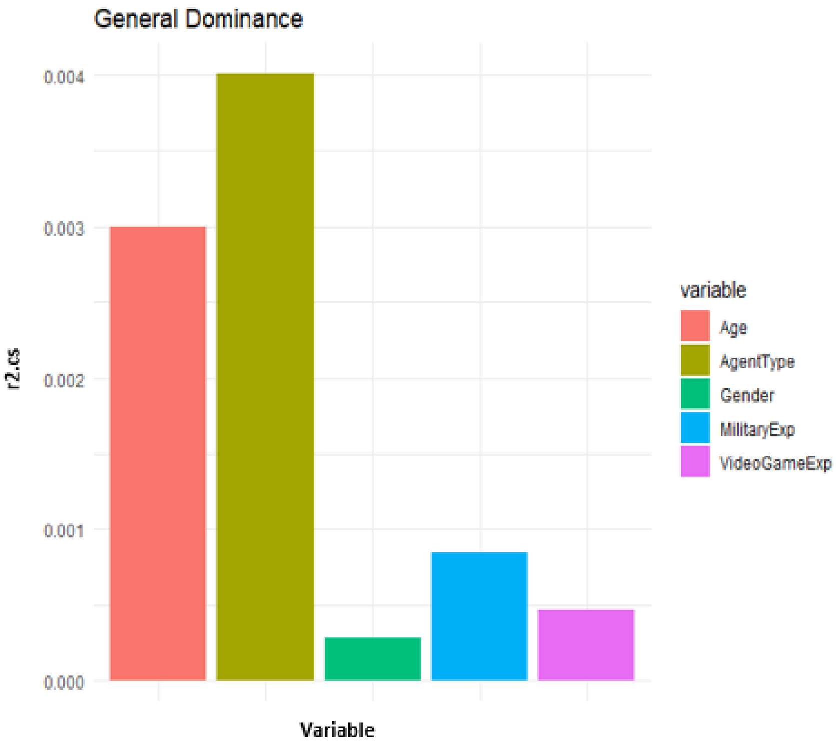


Figure 17. General dominance results Human/Agent Qualities Study A.

Note. The general dominance bar graph shows the unique contribution (in Cox & Snell’s (2018) pseudo R^2) averaged over all possible subset model sizes.

Task Perception

The global score on the NASA-TLX was removed as it was fully redundant with the six subscale scores. Overall, the variables were not strong in predicting hit rate, since the average additional contribution of each predictor was approximately 0.00 (Appendix F, **Table 49**).

Complete Dominance

As shown in **Error! Reference source not found.**, complete dominance was established for the performance subscales over all other scales, except the effort scale. Performance and effort were dominant over each other, depending on the subset model size. Dominance could not be established for performance over effort, or effort over performance (D_{ij} or D_{ji} = 0.5).

Table 17

Complete dominance results Task Perception in Study A.

Variable	Effort	Performance	Temporal Demand	Mental Demand	Physical Demand	Frustration
Effort	0.5	0.5	0.5	1	1	1
Performance	0.5	0.5	0.5	0.5	1	1
Temporal Demand	0.5	0.5	0.5	0.5	0.5	0.5
Mental Demand	0	0.5	0.5	0.5	0.5	0.5
Physical Demand	0	0	0.5	0.5	0.5	0.5
Frustration	0	0	0.5	0.5	0.5	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined

Conditional Dominance

As shown in **Figure 18**, performance dominated all other predictors when it was the only predictor in the model ($k = 0$) or with one other predictor ($k = 1$). For larger models, the effort subscale dominated. There was a monotonical increase for effort, mental demand, and temporal demand, which indicated that these variables were potential suppressors (Azen & Budescu, 2003). This indicated that these scales were not *unique* predictors of hit rate, as their predictive power was related to collinearity.

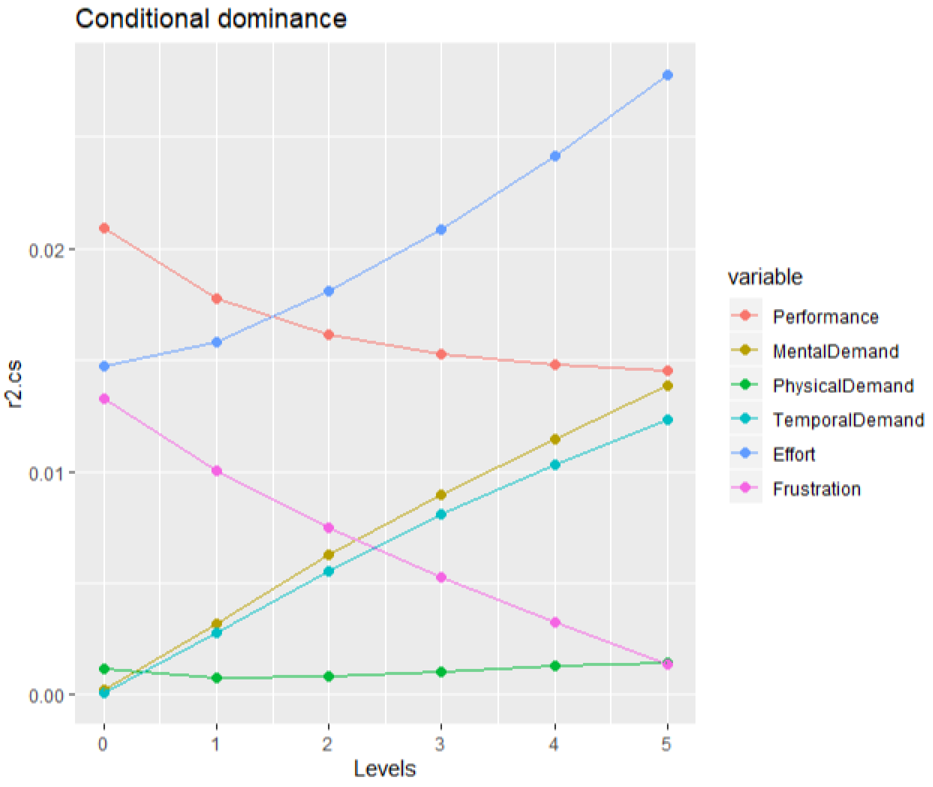


Figure 18. Conditional dominance results Task Perception in Study A.

Note. The plot shows the unique contribution (in Cox & Snell’s (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model. The unique contribution of a predictor should monotonically decrease with increasing subset model sizes (Azen & Budescu, 2003; Budescu, 1993). An increase, such as here for effort, mental demand, and temporal demand, indicates that these variables are potential suppressors, gaining importance through collinearity with other predictors in the model rather than through direct association with the outcome variable (Azen & Budescu, 2003).

General Dominance

The bar graph in **Figure 19** indicates that effort and performance generally dominated, which confirmed the transitive character of dominance (Budescu, 1993). Physical demand did not (generally) dominate any predictor.

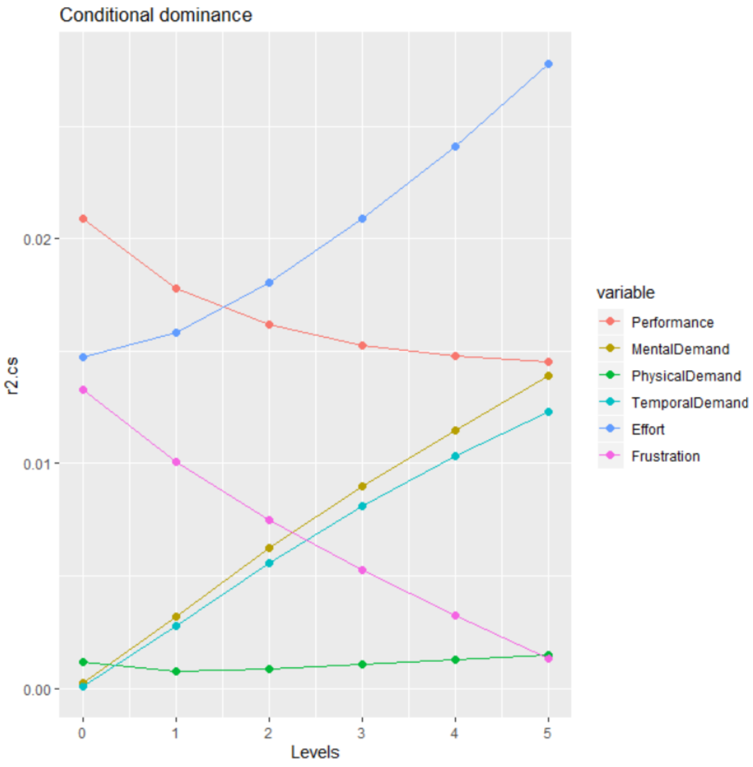


Figure 19. General dominance results Task Perception in Study A.

Note. The general dominance bar graph shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) averaged over all possible subset model sizes (levels).

Given the potential suppressing nature of the completely dominant subscale effort, the other completely dominant predictor, performance, was selected for inclusion in evaluation of the full model.

Full Model

Lastly, DA was conducted on the most important predictors, removing potential suppressors, such that:

$$winsorized\ Hit\ Rate \sim Gender + Military\ Experience + \ Age + Performance \\ + Agent\ Type + Video\ Gaming\ Experience$$

In the DA, the human predictors that were not important to hit rate, i.e., gender, military experience, and video gaming experience, were maintained, as they cannot be factored out in the natural world. However, they can be held constant and thereby accounted for, a method known as constrained DA (Azen & Budescu, 2003). The raw dominance analysis results can be found in Appendix F, **Table 50**.

Complete Dominance

Holding video gaming experience, military experience, and gender constant in the model, complete dominance was established for performance over all other predictors, followed by visual complexity (**Table 18**). Agent type also completely dominated age.

Table 18

Complete dominance results Full Model in Study A.

Variable	Performance	Visual Complexity	Agent Type	Age
Performance	0.5	1	1	1
Visual Complexity	0	0.5	1	1
Agent Type	0	0	0.5	1
Age	0	0	0	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

Since dominance is transitive (Budescu, 1993), conditional and general dominance followed the same pattern as complete dominance (**Error! Reference source not found.**, Appendix F). No additional dominance patterns were established.

Bootstrap

The results of *S* = 1000 bootstrap samples indicated that the confidence that the performance subscale and visual complexity would completely dominate in the actual population was low, varying from 39.3% to 58.5 % (

Table 19). This confidence increased slightly for the conditional dominance level, to around 60%, and to approximately 70% for the lowest level of dominance. This indicated that the robustness of the dominance results was not optimal.

Table 19

Bootstrap results for the full model in Study A.

Variable *i*	Variable *j*	D_{ij}	$\bar{D}_{\text{ij}}$	SE(D_{ij})	P_{ij}	P_{ji}	P_{noij}	Reproducibility
Complete Dominance								
Age	Agent Type Visual	0.5	0.423	0.358	0.191	0.345	**0.464**	**0.464**
Age	Complexity	0	0.316	0.320	0.088	**0.457**	0.455	**0.457**
Age	Performance Visual	0	0.263	0.343	0.111	**0.585**	0.304	**0.585**
Agent Type	Complexity	0	0.381	0.350	0.154	**0.393**	0.453	**0.393**
Agent Type Visual	Performance	0	0.319	0.384	0.179	**0.541**	0.280	**0.541**
Complexity	Performance	0	0.435	0.429	0.310	**0.441**	0.249	**0.441**
Variable *i*	Variable *j*	D_{ij}	$\bar{D}_{\text{ij}}$	SE(D_{ij})	P_{ij}	P_{ji}	P_{noij}	Reproducibility
Conditional Dominance								
Age	Agent Type Visual	0	0.434	0.410	0.278	**0.410**	0.312	**0.410**
Age	Complexity	0	0.273	0.371	0.151	**0.606**	0.243	**0.606**
Age	Performance Visual	0	0.247	0.361	0.136	**0.642**	0.222	**0.642**
Agent Type	Complexity	0	0.347	0.401	0.215	**0.521**	0.264	**0.521**
Agent Type Visual	Performance	0	0.307	0.410	0.218	**0.604**	0.178	**0.604**
Complexity	Performance	0	0.435	0.457	0.360	**0.491**	0.149	**0.491**
General Dominance								
Age	Agent Type Visual	0	0.478	0.500	0.478	**0.522**	0.000	**0.522**
Age	Complexity	0	0.218	0.413	0.218	**0.782**	0.000	**0.782**
Age	Performance	0	0.239	0.427	0.239	**0.761**	0.000	**0.761**
Variable *i*	Variable *j*	D_{ij}	$\bar{D}_{\text{ij}}$	SE(D_{ij})	P_{ij}	P_{ji}	P_{noij}	Reproducibility
	Visual							
Agent Type	Complexity	**0**	0.302	0.459	0.302	**0.698**	0.000	**0.698**
Agent Type Visual	Performance	**0**	0.286	0.452	0.286	**0.714**	0.000	**0.714**
Complexity	Performance	**0**	0.449	0.498	0.449	**0.551**	0.000	**0.551**

Note. D_{ij} is the dominance value of the original analyses, wherein $D_{ij} = 1 – D_{ji}$. Although each pair has two possible orders (*ij* and *ji*), only one order is shown to reduce redundancy

The $P_{..}$ values indicate the proportion of the S = 1000 bootstrap sample that replicated D_{ij}, such that $P_{ij} = Pr(D_{ij} = 1)$, $P_{ji} = Pr(D_{ij} = 0)$, $P_{noij} = Pr(D_{ij} = 0.5)$. The reproducibility value refers to the proportion of the bootstrap sample that replicated D_{ij}.

The bold values imply a reference to the dominance value from the sample (D_{ij}).

Model Fit Evaluation: Regression

The dominant predictors were combined into a hierarchical beta regression model to evaluate the model fit, such that:

$$\begin{aligned} winsorized\ Hit\ Rate \sim Performance + Visual\ Complexity + Agent\ Type + Age \\ + Military\ Experience + \ Gender + Video\ Gaming\ Experience \end{aligned}$$

The results are found in **Table 20**. The pseudo R^2 of the model is 0.038, $\chi^2(9) = 315.27$, p = 0.780. None of the variables were significant in the regression on (winsorized) hit rate.

Table 20

Results beta regression on Full Model in Study A.

	Beta Coefficient	SE	z-value	Probability(>|z|)
Intercept	3.37	0.74	4.56	**< 0.001**
Video Gaming Experience	-0.03	0.06	-0.52	0.606
Gender	0.05	0.21	0.26	0.769
Performance	0.00	0.00	-1.04	0.297
Visual Complexity	0.18	0.19	0.96	0.336
Agent Type	0.07	0.14	0.51	0.613
Age	0.00	0.01	0.03	0.974
Military Experience	-0.06	0.52	-0.11	0.913

Note. Significant values are in bold.

The poor fit of the model was confirmed by the predicted vs. observed values plot on the right side in **Error! Reference source not found.**. The plotted line is the fitted line based on maximum likelihood.

Hypotheses

Although the dominance analyses indicated a qualitatively different pattern in unique additional contribution for the predictors in the Core model, none of these differences were statistically significant. Therefore, all null hypotheses were rejected. If the differences would have been significant, the Core model should resemble **Figure 20**, with a primary contribution by Task Perception, driven by the NASA-TLX performance subscale, followed by Task Composition (visual complexity of symbols), and lastly Human/Agent Qualities, driven by agent morphology type.

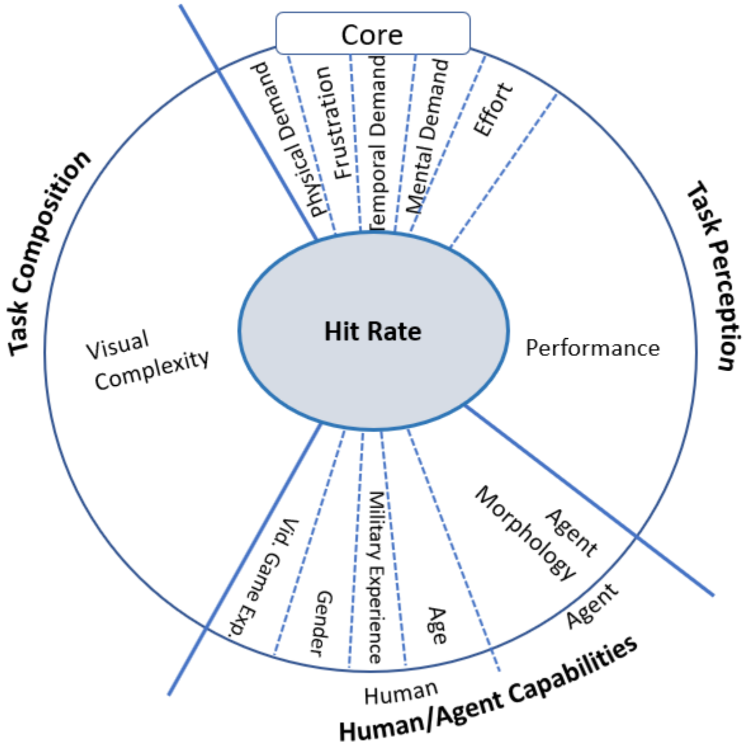


Figure 20. Updated Core model based on results Study A.

Note. The size of the sections of the pie represent the relative importance of the factor to task performance.

Study B

Hypotheses Study B

The predictors that were available in study B were tested against the null hypotheses of the Core model: everything is equal. This is represented in **Figure 21**. Task duration was included as well, since condition B.2 was divided in two blocks of five minutes: one with low event rate (15 characters/minute) and one with high event rate (60 characters/minute). Agent qualities were not available. The collaborative agent was simulated to be a fully autonomous and 100% reliable, intelligent, embodied, and out-of-sight agent that scouted the outer cordon for threats and contraband.

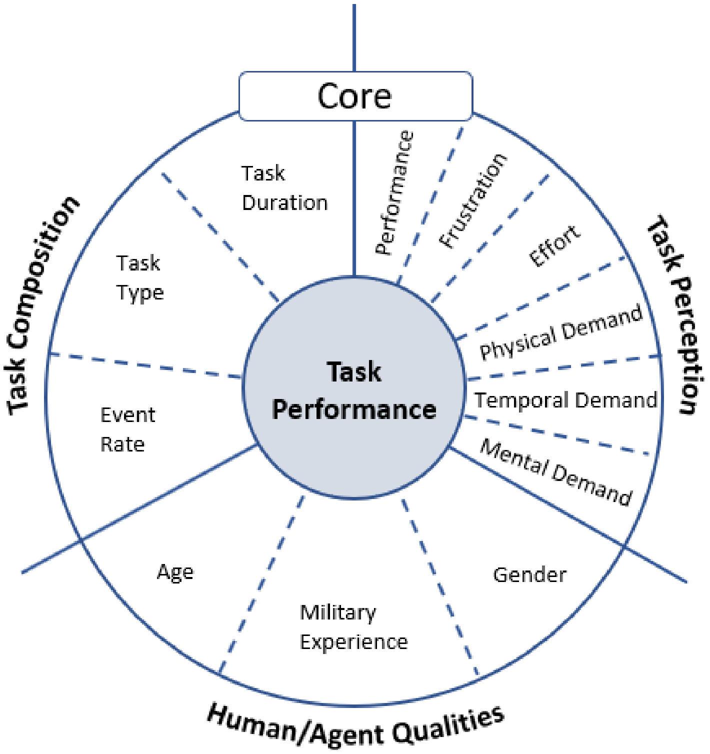


Figure 21. Visualization of hypotheses in Study B.

Note. The size of the sections of the pie represent the relative importance of the factor to task performance. The null hypotheses are that all factors and categories are of equal importance to task performance.

The null hypotheses were as follows:

Hypothesis 1. Of the Human/Agent Qualities, all human factors (age, gender, military experience, and video gaming experience) are equally important.

Hypothesis 2. All NASA-TLX subscales (Task Perception) contribute equally to task performance.

Hypothesis 3. All Task Composition variables contribute equally to task performance.

Hypothesis 4. Task Composition, Perception of Task, and Human/Agent Qualities are equally important to task performance.

Dominance Analysis Study B

Linearity was established between predictors and the response variable (Appendix F, **Error! Reference source not found.**).

<u>Human (/Agent) Qualities</u>

Agent qualities were not manipulated in Study B; the agent was simulated to be fully autonomous and 100% reliable. Overall, the variables were not strong in predicting hit rate, since the average additional contribution of each predictor was 0.000 – 0.014 (Appendix F, **Table 51**). Video gaming experience was excluded from the analyses due to a high percentage of missing values.

Complete Dominance

As shown in

Table 21, military experience completely dominated the other predictors in the model. Complete dominance for age over gender, or reversed, could not be established.

Table 21

Complete dominance results Human/Agent Qualities in Study B.

Variable	Military Experience	Age	Gender
Military Experience	0.5	1	1
Age	0	0.5	0.5
Gender	0	0.5	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

Conditional Dominance

Figure 22 shows that complete dominance could not be established for age and gender. Age only dominated gender in *k* = 0 subset models, while gender dominated age for larger subset models.

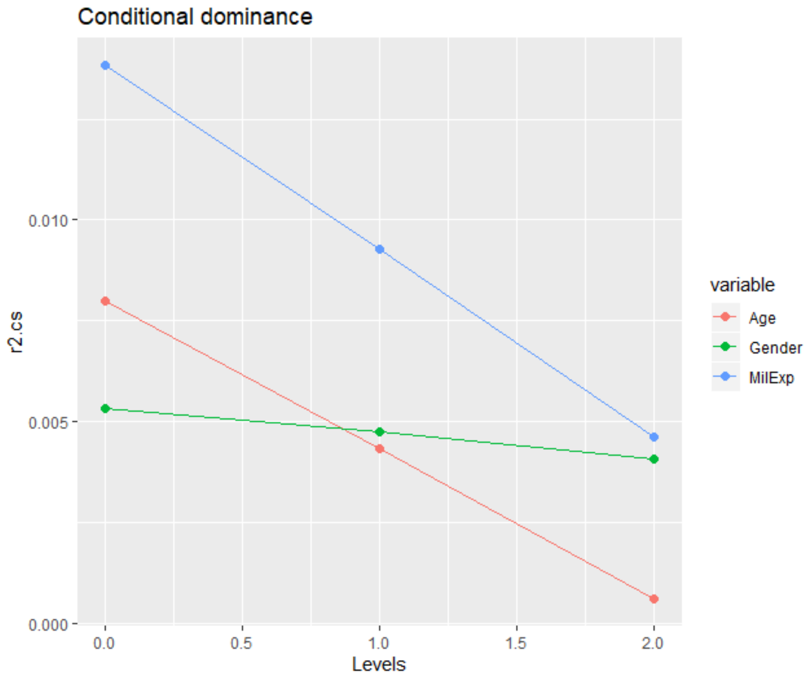


Figure 22. Conditional dominance results Human/Agent Qualities in Study B.

Note. The plot shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model.

General Dominance

In addition to the higher levels of dominance, general dominance was established for gender over age (**Figure 23**).

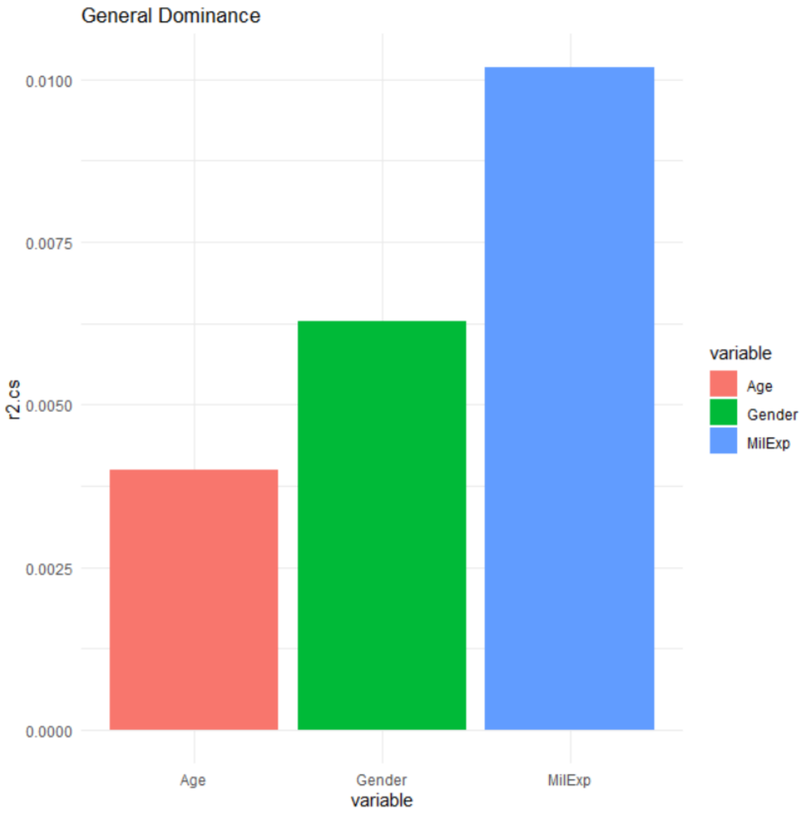


Figure 23. General dominance results Human/Agent Qualities Study B.

Note. The general dominance bar graph shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) averaged over all possible subset model sizes.

Task Perception

The global score on the NASA-TLX was removed as it was fully redundant with the six subscale scores. Overall, the variables were not strong in predicting hit rate, since the average additional contribution of each predictor was very low (Appendix F, **Table 52**).

Complete Dominance

Performance completely dominated mental demand, physical demand, effort, and frustration subscales (**Table 22**). Complete dominance was not established for performance over temporal demand, or vice versa. Temporal demand completely dominated the mental demand, physical demand, and effort subscales. Dominance of temporal demand over frustration could not be established.

Table 22

Complete dominance results Task Perception in Study B.

Variable	Performance	Temporal Demand	Frustration	Mental Demand	Physical Demand	Effort
Performance	0.5	0.5	1	1	1	1
Temporal Demand	0.5	0.5	0.5	1	1	1
Frustration	0	0.5	0.5	0.5	0.5	0.5
Mental Demand	0	0	0.5	0.5	0.5	0.5
Physical Demand	0	0	0.5	0.5	0.5	0.5
Effort	0	0	0.5	0.5	0.5	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

Conditional Dominance

Figure 24 shows the conditional dominance pattern of the predictors over all subset model sizes. Here, it was clear that complete dominance could not be established between performance and temporal demand. Performance was a stronger predictor for k = 0 and k = 1 subset models. However, for larger models, temporal demand grew increasingly more important. This effect indicates that temporal demand was potentially a suppressor variable, along with the effort subscale. In addition, effort and frustration dominated the mental and physical demand subscales. Frustration dominated effort for subset models up to *k* = 3, due to the suppressor effect of effort.

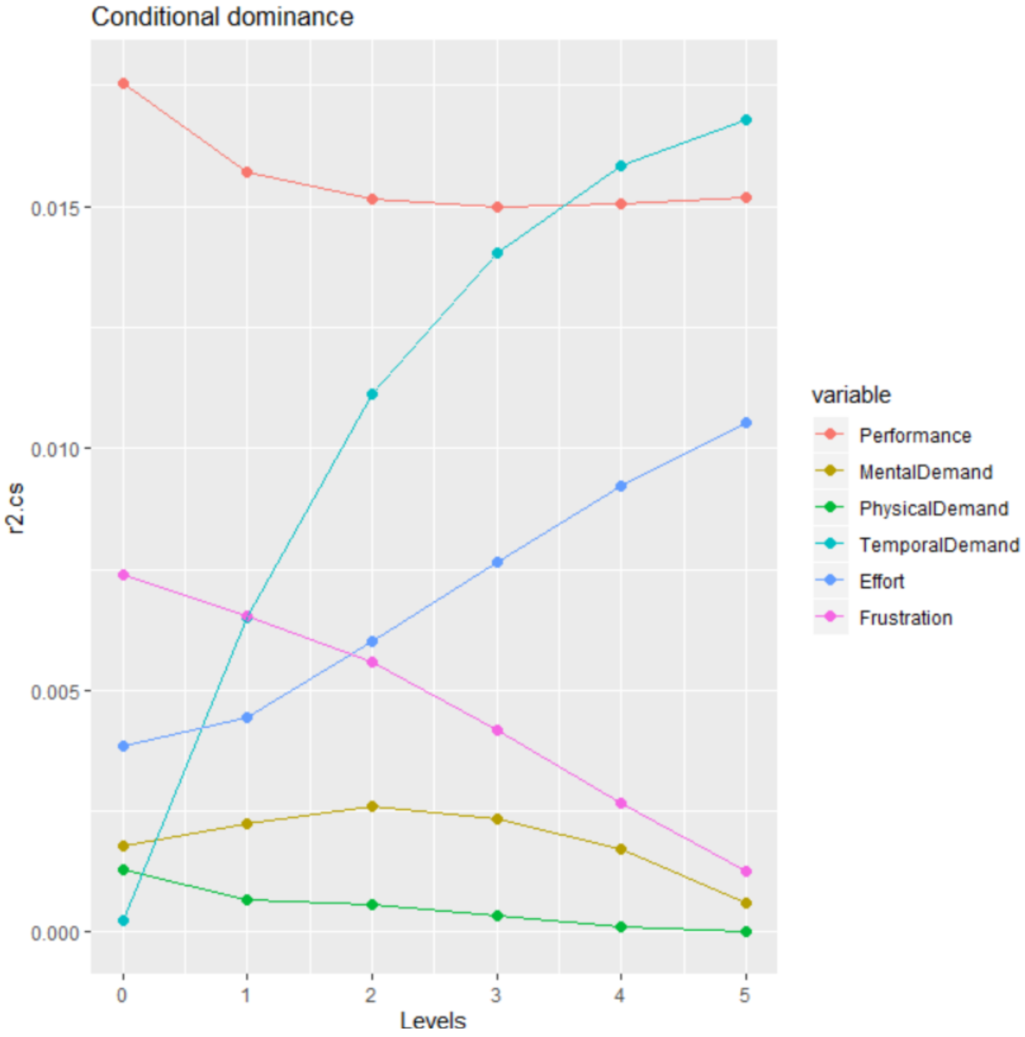


Figure 24. Conditional dominance results Task Perception in Study B.

Note. The plot shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model. The unique contribution of a predictor should monotonically decrease with increasing subset model sizes (Azen & Budescu, 2003; Budescu, 1993). An increase, such as here for temporal demand and effort, indicates that these variables are potential suppressors, gaining importance through collinearity with other predictors in the model rather than through direct association with the outcome variable (Azen & Budescu, 2003).

General Dominance

General dominance was established such that frustration > (i.e., dominated) effort > mental demand > physical demand (**Figure 25**).

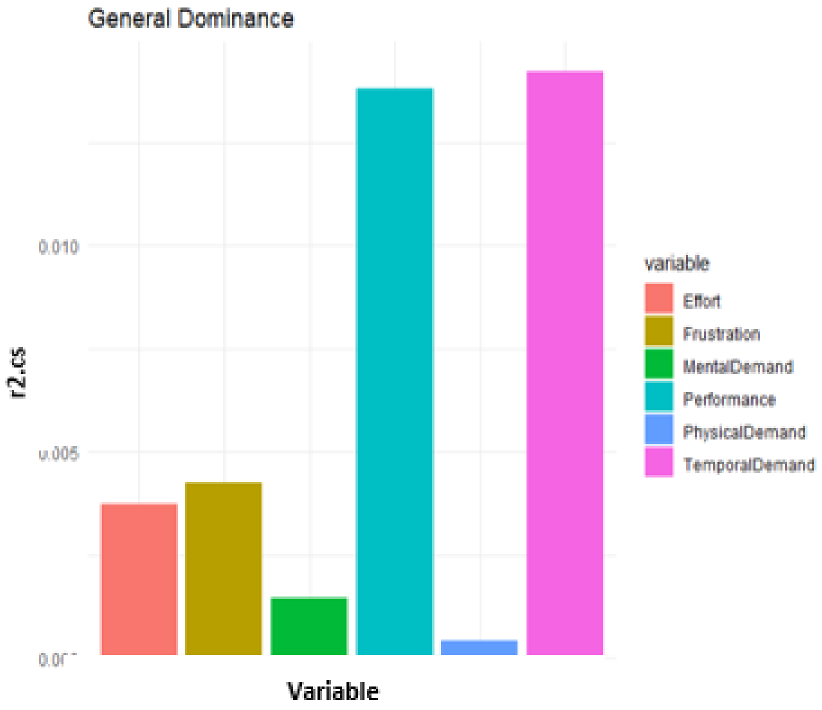


Figure 25. General dominance results Task Perception in Study B.

Note. The general dominance bar graph shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) averaged over all possible subset model sizes.

Task Composition

Type of event rate (changing vs. constant) was removed from the analyses due to redundancy issues with other predictors. Overall, the variables were not strong in predicting hit rate, since the average additional contribution of each predictor was very low (Appendix F, **Table 53**).

Complete Dominance

As presented in **Table 23**, event rate completely dominated task type and task duration. Complete dominance could not be established between task type and task duration.

Table 23

Complete dominance results Task Composition in Study B.

Variable	Event Rate	Task Duration	Task Type
Event Rate	0.5	1	1
Task Duration	0	0.5	1
Task Type	0	0	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

Conditional Dominance

Task duration was a more important predictor than task type for $k = 0$ and $k = 1$ subset models (**Figure 26**). When two other predictors were in the model ($k = 2$), dominance could not be established between the two.

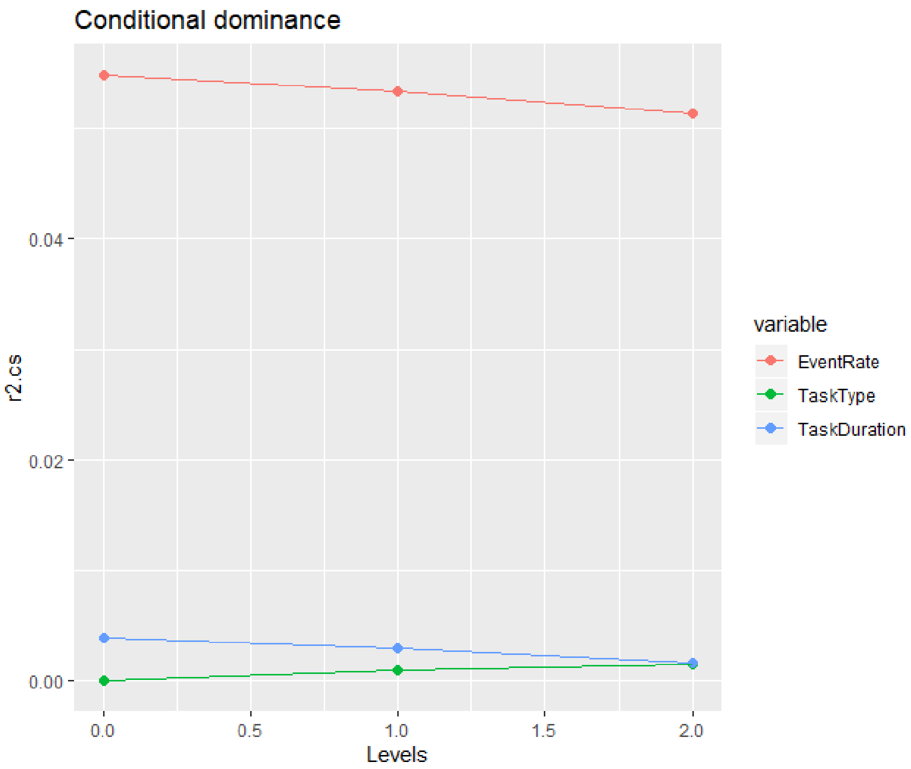


Figure 26. Complete dominance Task Composition in Study B.

Note. The plot shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model. The unique contribution of a predictor should monotonically decrease with increasing subset model sizes (Azen & Budescu, 2003; Budescu, 1993). An increase, such as here is slightly seen for the variable task type, indicates that this variable is a potential suppressor, gaining importance through collinearity with other predictors in the model rather than through direct association with the outcome variable (Azen & Budescu, 2003).

General Dominance

Aside from the overall dominance of event rate, on average over all subset models, task duration was generally a more predictor than task type (**Figure 27**).

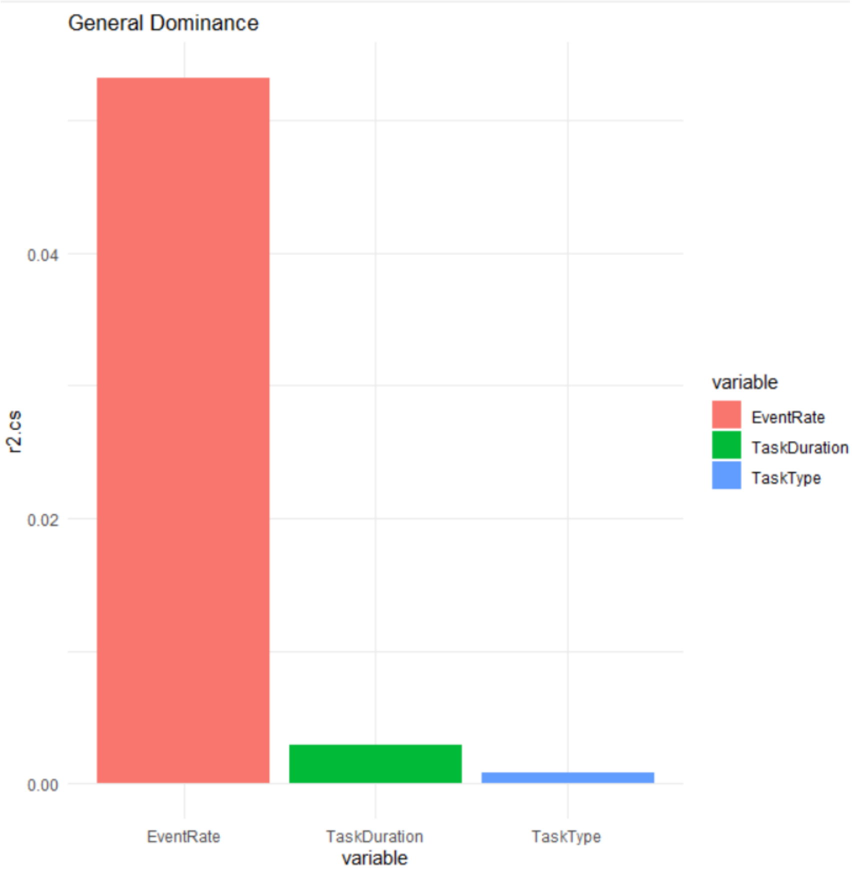


Figure 27. General dominance results Task Composition Study B.

Note. The general dominance bar graph shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) averaged over all possible subset model sizes.

Full Model

Lastly, DA was conducted on the full model, such that

$$Winsorized\ Hit\ Rate \sim Age + Gender + \ Military\ Experience +$$

$$Event\ Rate + Performance$$

In the DA, the human predictors that were not important to hit rate, age and gender, were held constant. The overall model pseudo R^2 was low (see **Table 54** in Appendix F).

Complete Dominance

As shown in **Table 24**, event rate completely dominated the NASA-TLX performance subscale and military experience. Military experience was dominated by the performance scale.

Table 24

Complete dominance results Full Model in Study B.

Variable	Event Rate	Performance	Military Experience
Event Rate	0.5	1	1
Performance	0	0.5	1
Military Experience	0	0	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

No additional conditional and general dominance patterns were established, since complete dominance was prevailing (**Error! Reference source not found.**, Appendix F).

Bootstrap

The results of *S* = 1000 bootstrap samples indicated that the confidence that event rate would dominate in the actual population was high, varying from 81.0% to 98.8 % (**Table 25**). The confidence that performance would dominate military experience ranged from 66.7 – 81.3 %.

Table 25

Bootstrap results Full Model in Study B.

Variable *i*	Variable *j*	D_{ij}	$\overline{D}_{ij}$	SE(D_{ij})	P_{ij}	P_{ji}	P_{noij}	Reproducibility
Compete Dominance								
Military Experience	Event Rate	0	0.015	0.099	0.005	**0.975**	0.020	**0.975**
Military Experience	Performance	0	0.196	0.299	0.059	**0.667**	0.274	**0.667**
Event Rate	Performance	1	0.860	0.309	**0.810**	0.090	0.100	**0.810**
Conditional Dominance								
Military Experience	Event Rate	0	0.015	0.105	0.008	**0.979**	0.013	**0.979**
Military Experience	Performance	0	0.182	0.317	0.085	**0.721**	0.194	**0.721**
Event Rate	Performance	1	0.864	0.318	**0.830**	0.103	0.067	**0.830**
General Dominance								
Military Experience	Event Rate	0	0.012	0.109	0.012	**0.988**	0.000	**0.988**
Military Experience	Performance	0	0.187	0.390	0.187	**0.813**	0.000	**0.813**
Event Rate	Performance	1	0.879	0.326	**0.879**	0.121	0.000	**0.879**

Note. D_{ij} is the dominance value of the original analyses, wherein $D_{ij} = 1 – D_{ji}$. Although each pair has two possible orders (*ij* and *ji*), only one order is shown to reduce redundancy

The $P_{..}$ values indicate the proportion of the S = 1000 bootstrap sample that replicated D_{ij}, such that $P_{ij} = Pr(D_{ij} = 1)$, $P_{ji} = Pr(D_{ij} = 0)$, $P_{noij} = Pr(D_{ij} = 0.5)$. The reproducibility value refers to the proportion of the bootstrap sample that replicated D_{ij}.

The bold values imply a reference to the dominance value from the sample (D_{ij}).

Model Fit Evaluation: Regression

The most important predictors were combined into a hierarchical beta regression model, wherein all human variables were preserved, as they would always be present in the natural world as well:

$$winsorized\ Hit\ Rate \sim Event\ Rate + Performance + \ Military\ Experience + Gender + Age$$

The beta coefficients and significance testing are presented in **Table 26**. The pseudo R^2 of the model is 0.107, $\chi^2(7) = 484.39$, $p = 0.008$. Even though the model was a poor fit, the full model significantly predicted hit rate. Only event rate was significant, although the performance subscale of the NASA-TLX approached significance.

Table 26

Results beta regression on Full Model in Study B.

	Beta Coefficient	SE	z-value	Pr(>|z|)
Intercept	3.30	0.49	6.78	**< 0.001**
Event Rate	-0.01	0	-3.48	**< 0.001**
Performance	-0.01	0	-1.91	0.056
Military Experience	0.19	0.18	1.04	0.298
Gender	-0.06	0.13	-0.46	0.643
Age	0.01	0.02	0.35	0.727

Note. Significant values are in bold.

The residuals did not show signs of dependence between the errors. The predicted vs. observed values plot confirmed the poor fit of the model (the plotted line is the fitted line based on maximum likelihood; see Appendix F, **Error! Reference source not found.**).

Hypotheses

The null Hypothesis 1, of the Human/Agent Qualities, all human factors (age, gender, military experience, and video gaming experience) are equally important, was rejected. Video gaming experience was excluded from the analyses. However, military experience was the most important predictor of hit rate in this study that contained a relatively larger number of military members (33.1%).

The null Hypothesis 2, all NASA-TLX subscales (Task Perception) contribute equally to task performance, was rejected. The performance subscale dominated all other predictors, even though the pseudo R^2 remained small. Temporal demand also showed importance. However, this importance increased with size of the subset model, indicating it was a potential suppressor variable, and therefore not included in the full model analyses.

The null Hypothesis 3, all Task Composition variables contribute equally to task performance, was rejected. Event rate dominated all other predictors, even though the pseudo R^2 remained small.

The null Hypothesis 4, Task Composition, Perception of Task, and Human/Agent Qualities are equally important to task performance, was rejected. Task Composition, in the form of event rate, was most important to hit rate, followed by Task Perception (NASA-TLX performance subscale) and lastly military experience.

Based on the analyses, the model is updated and represented in **Figure 28**. In study B, Task Composition was most important to hit rate, driven by event rate, followed by Task Perception (NASA-TLX performance subscale), and lastly Human(/Agent) Qualities, based on military experience.

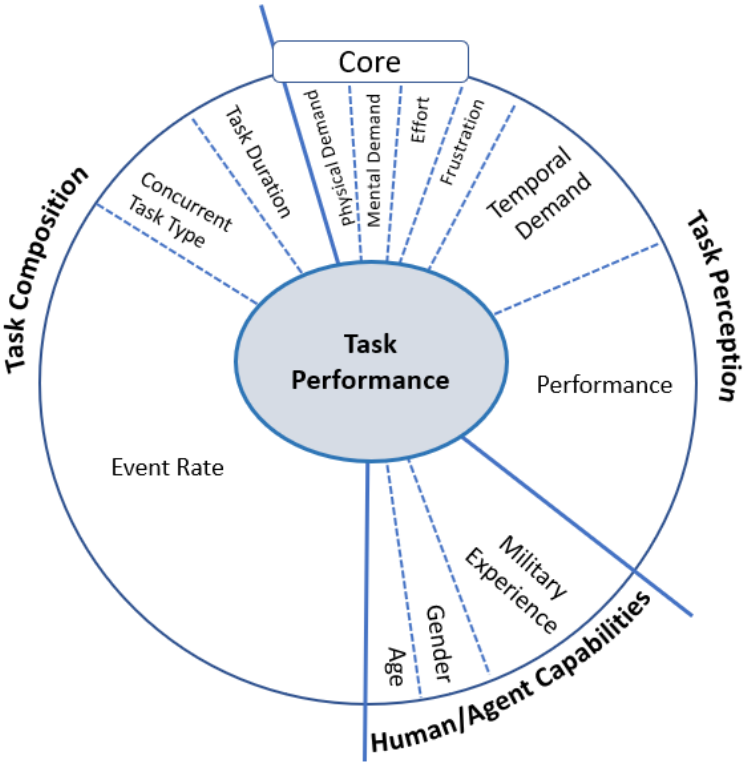


Figure 28. Updated Core model based on results in Study B.

Note. The size of the sections of the pie represent the relative importance of the factor to task performance.

Study C

Hypotheses Study C

Study C manipulated the agent report modality (single adaptive vs. dual) and event rate; however, event rate could not be analyzed since the scenarios were compared as a whole (see **Figure 10** for the experimental design). Similar blocks could not be individually combined as some were significantly different (Barber et al., 2019). Agent qualities were not available. The collaborative agent was simulated to be a fully autonomous and 100% reliable, intelligent, embodied agent that scouted the outer cordon for threats and contraband. The null hypotheses are visually presented in **Figure 29**.

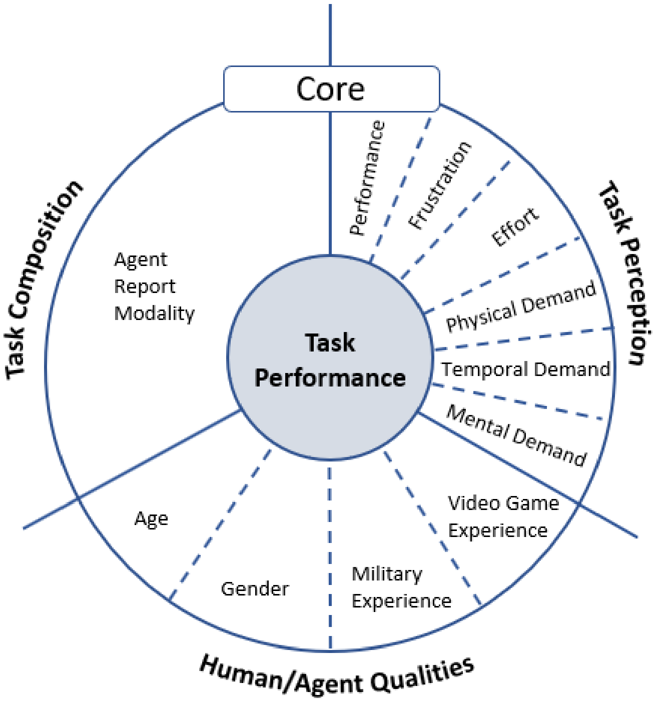


Figure 29. Visual representation of hypotheses in Study C.

Note. The size of the sections of the pie represent the relative importance of the factor to task performance. The null hypotheses are that all factors and categories are of equal importance to task performance.

The null hypotheses were as follows:

Hypothesis 1. Of the Human/Agent Qualities, all factors are equally important to task performance.

Hypothesis 2. All NASA-TLX subscales (Task Perception) contribute equally to task performance.

Hypothesis 3. Task Composition, Perception of Task, and Human/Agent Qualities are equally important to task performance.

Dominance Analysis Study C

Linearity was established between predictors and the response variable (Appendix F, **Error! Reference source not found.**).

Human (/Agent) Qualities

Agent qualities were not manipulated in Study C; the agent was simulated to be fully autonomous and 100% reliable. The overall model pseudo R^2 was low (see **Table 55** in Appendix F).

Complete Dominance

As shown in

Table 27, video gaming experience completely dominated gender, military experience, and age. In addition, military experience and gender completely dominated age. Complete dominance could not be established between military experience and gender.

Table 27

Complete dominance results Human/Agent Qualities in Study C.

Variable	Video Gaming Experience	Gender	Military Experience	Age
Video Gaming Experience	0.5	1	1	1
Gender	0	0.5	0.5	1
Military Experience	0	0.5	0.5	1
Age	0	0	0	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

Conditional Dominance

As shown in **Figure 30**, gender dominated military experience for models of size k = 0 and k = 1. However, for larger subset models, military experience dominated gender. Moreover, the increase in R^2 for military experience indicated this predictor was a possible suppressor variable, gaining importance due to collinearity with other predictors (Azen & Budescu, 2003).

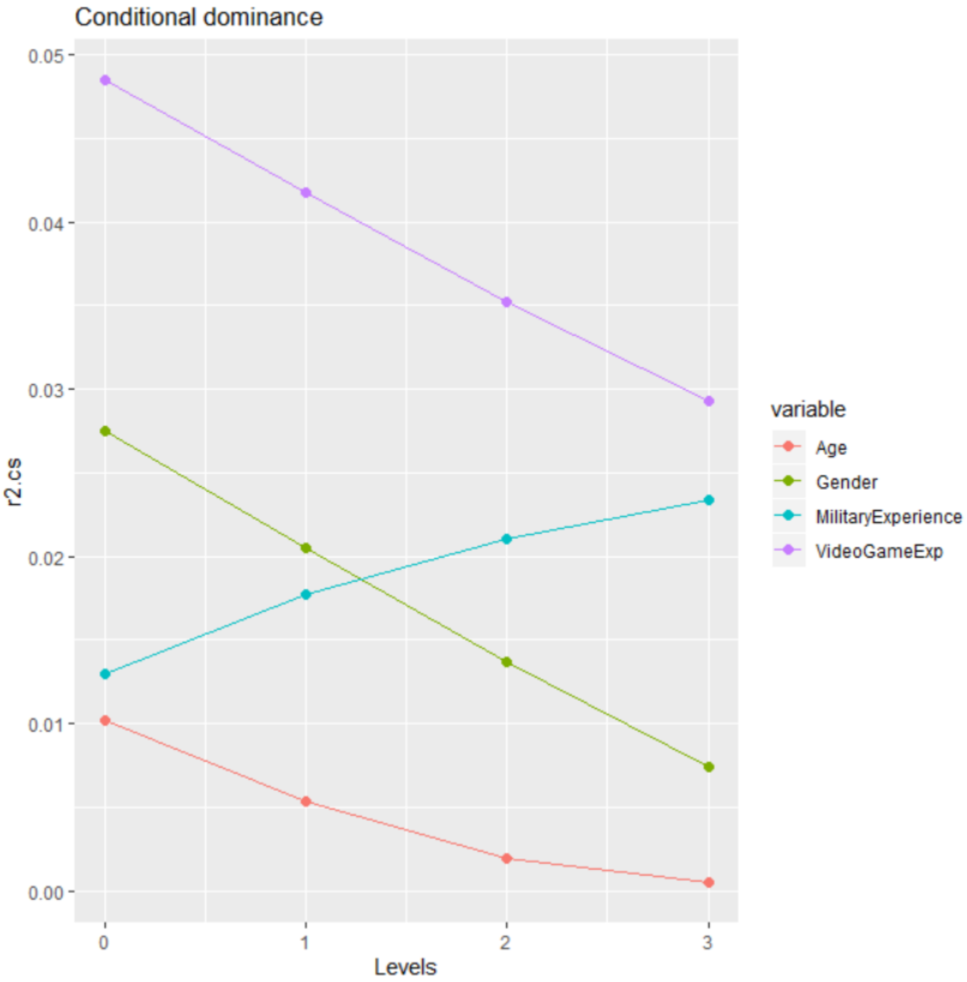


Figure 30. Conditional dominance results Human/Agent Qualities Study C.

Note. The plot shows the unique contribution (in Cox & Snell’s (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model. The unique contribution of a predictor should monotonically decrease with increasing subset model sizes (Azen & Budescu, 2003; Budescu, 1993). An increase, such as here for military experience, indicates that this variable is a potential suppressor, gaining importance through collinearity with other predictors in the model rather than through direct association with the outcome variable (Azen & Budescu, 2003).

General Dominance

General dominance, the lowest level of dominance, was not established for military experience over age and gender (**Figure 31**). Age did not generally dominate any other predictors.

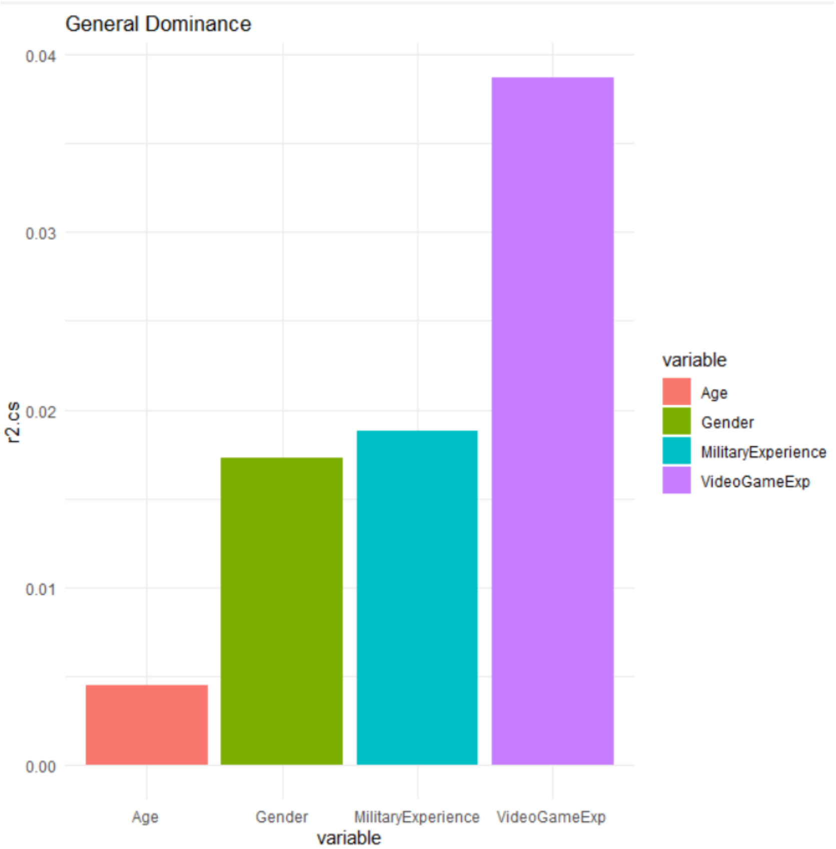


Figure 31. General dominance results Human/Agent Qualities in Study C.

Note. The general dominance bar graph shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) averaged over all possible subset model sizes.

Task Perception

The overall model's pseudo R^2 for Task Perception was low (see **Table 56** in Appendix F).

Complete Dominance

As shown in **Table 28**, temporal demand completely dominated all other subscales. Frustration completely dominated mental demand and physical demand completely dominated performance. Complete dominance between frustration and physical demand could not be established.

Table 28

Complete dominance results Task Perception in Study C.

Variable	Temporal Demand	Physical Demand	Frustration	Effort	Performance	Mental Demand
Temporal Demand	0.5	1	1	1	1	1
Physical Demand	0	0.5	0.5	0.5	1	0.5
Frustration	0	0.5	0.5	0.5	0.5	1
Effort	0	0.5	0.5	0.5	0.5	0.5
Performance	0	0	0.5	0.5	0.5	0.5
Mental Demand	0	0.5	0	0.5	0.5	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

Conditional Dominance

In addition to the complete dominance pattern, physical demand also conditionally dominated the effort subscale (**Figure 32**). Physical demand dominated the effort subscale for smaller subset models (up to *k* = 3), while frustration dominated physical demand for larger subset models (*k* > 3). This was an indication that the frustration subscale was a potential suppressor, increasing slightly in importance through collinearity with other predictors in the model.

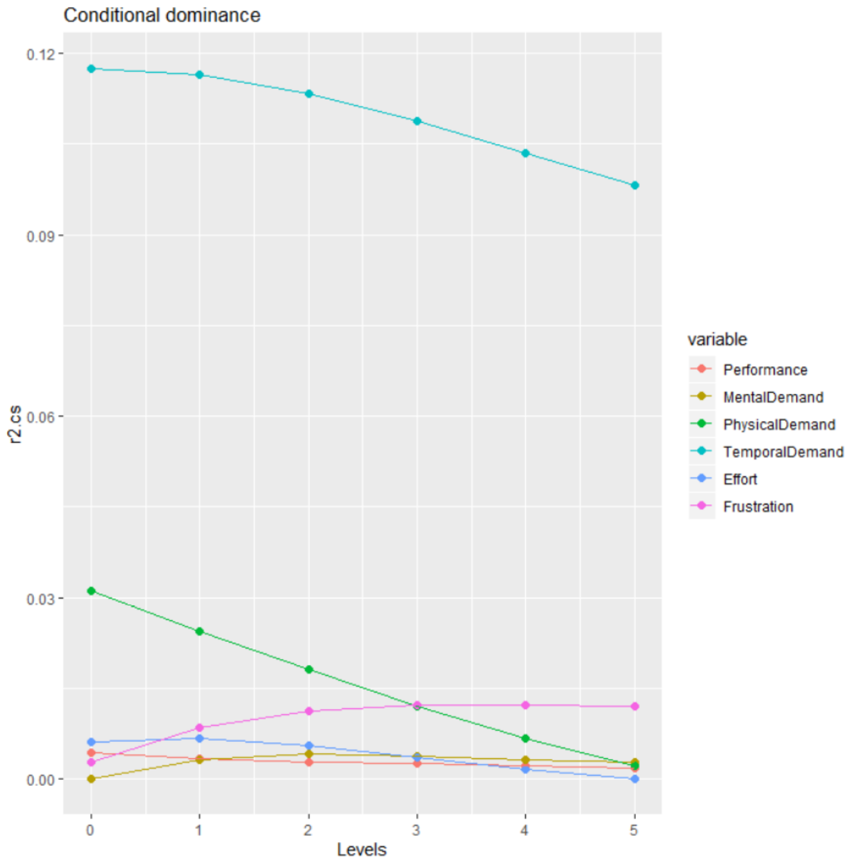


Figure 32. Conditional dominance results Task Perception in Study C.

Note. The plot shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model.

General Dominance

In addition to the complete and conditional dominance patterns, general dominance was established for physical demand over performance, mental demand, effort, and frustration (**Figure 33**). Frustration generally dominated performance and effort. Effort generally dominated performance and mental demand, while performance dominated mental demand. Mental demand was the least important predictor of hit rate.

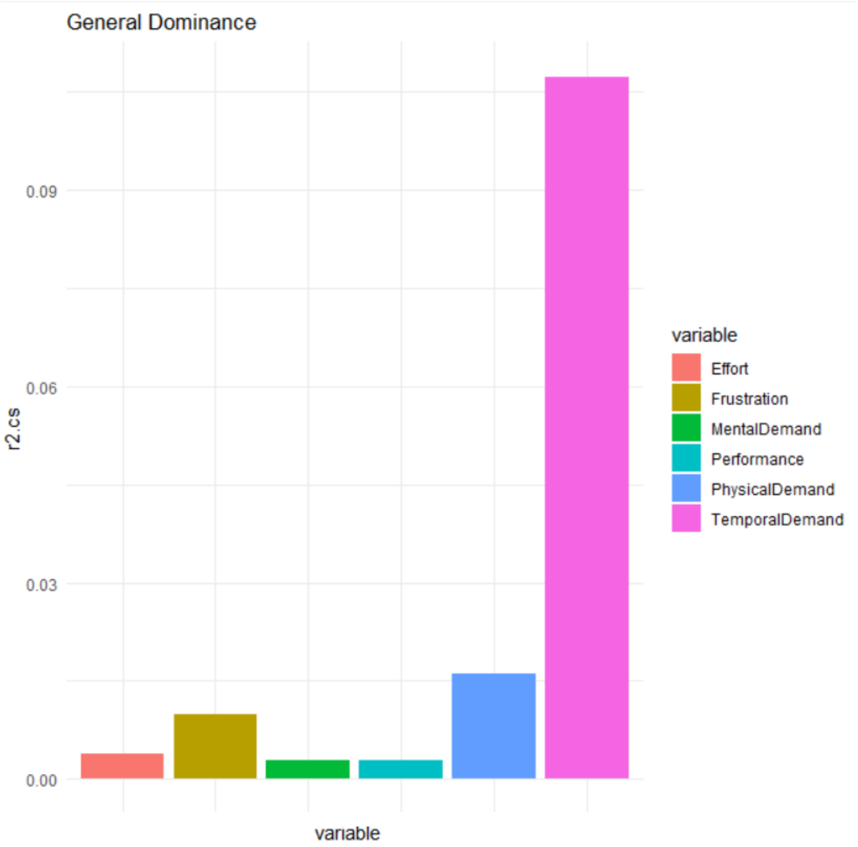


Figure 33. General dominance results Task Perception in Study C.

Note. The general dominance bar graph shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) averaged over all possible subset model sizes.

Full Model

Lastly, DA was conducted on the full model, such that

$$Winsorized\ Hit\ Rate \sim Age + Gender + Video\ Gaming\ Experience +$$

$$Agent\ Report\ Modality + Temporal\ Demand + Military\ Experience$$

In DA, the human predictors that were not important to hit rate, age, military experience and gender, were held constant. The overall model pseudo R^2 was low (see **Table 57** in Appendix F).

Complete Dominance

As shown in **Table 29**, temporal demand completely dominated all other predictors. Video gaming experience completely dominated agent report modality.

Table 29

Complete dominance results Full Model in Study C.

Variable	Temporal Demand	Video Gaming Experience	Agent Report Modality
Temporal Demand	0.5	1	1
Video Gaming Experience	0	0.5	1
Agent Report Modality	0	0	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

No additional levels of dominance (conditional or general) could be established, as the highest level of dominance prevailed the results (**Error! Reference source not found.**, Appendix F).

Bootstrap

The bootstrap does not handle a large number of missing values. Therefore, military experience was excluded from the bootstrap. The results of *S* = 1000 bootstrap samples indicated that the confidence that temporal demand would dominate age, video gaming experience, gender, and agent report modality in the actual population was high, around 80% (**Table 30**). This level of confidence grew higher as the level of dominance decreased to conditional and general dominance.

Table 30

Bootstrap results Full Model in Study C.

Variable *i*	Variable *j*	D_{ij}	$\overline{D}_{\mathrm{ij}}$	SE(D_{ij})	P_{ij}	P_{ji}	P_{noij}	Reproducibility
Complete Dominance								
Video Gaming Experience	Temporal Demand	0	0.129	0.289	0.071	**0.813**	0.116	**0.813**
Video Gaming Experience	Agent Report Modality	1	0.647	0.348	**0.432**	0.139	0.429	**0.432**
Temporal Demand	Agent Report Modality	1	0.879	0.279	**0.821**	0.064	0.115	**0.821**
Video Gaming Experience	Temporal Demand	0	0.118	0.292	0.081	**0.846**	0.073	**0.846**
Video Gaming Experience	Agent Report Modality	1	0.657	0.356	**0.459**	0.145	0.396	**0.459**
Temporal Demand	Agent Report Modality	1	0.889	0.277	**0.844**	0.067	0.089	**0.844**

Variable *i*	Variable *j*	D_{ij}	$\overline{D}_{\mathrm{ij}}$	SE(D_{ij})	P_{ij}	P_{ji}	P_{noij}	Reproducibility
General Dominance								
Video Gaming Experience	Temporal Demand	0	0.126	0.332	0.126	**0.874**	0.000	**0.874**
Video Gaming Experience	Agent Report Modality	1	0.670	0.470	**0.670**	0.330	0.000	**0.670**
Temporal Demand	Agent Report Modality	1	0.898	0.303	**0.898**	0.102	0.000	**0.898**

Note. D_{ij} is the dominance value of the original analyses, wherein $D_{ij} = 1 – D_{ji}$. Although each pair has two possible orders (*ij* and *ji*), only one order is shown to reduce redundancy

The $P_{..}$ values indicate the proportion of the S = 1000 bootstrap sample that replicated D_{ij}, such that $P_{ij} = Pr(D_{ij} = 1)$, $P_{ji} = Pr(D_{ij} = 0)$, $P_{noij} = Pr(D_{ij} = 0.5)$. The reproducibility value refers to the proportion of the bootstrap sample that replicated D_{ij}.

The bold values imply a reference to the dominance value from the sample (D_{ij}).

The dominance of video gaming experience was less robust. The results indicated that the confidence that this dominance pattern would occur in the population was 43.2% to 71.5% and only increased slightly under lower levels of dominance.

Model Fit Evaluation: Regression

To evaluate the model, the most important predictors were combined into a hierarchical beta regression model, wherein all human variables were preserved, as they would always be present in the natural world as well. Military experience was preserved in the hierarchy.

$$winsorized\ Hit\ Rate \sim Temporal\ Demand + Video\ Gaming\ Experience +$$

$$Agent\ Report\ Modality + Military\ Experience + \ Gender + Age$$

The beta coefficients and significance testing are presented in **Table 31**. The pseudo R^2 of the model was 0.189, wherein the model was significantly better than the null model, $\chi^2(8) = 236.87$, $p < 0.001$. The full model significantly predicted hit rate, based on temporal demand and video gaming experience. The other predictors were non-significant.

Table 31

Results beta regression on Full Model in Study C.

	Beta Coefficient	SE	z-value	Pr(>|z|)
Intercept	3.36	0.66	5.09	**< 0.001**
Temporal Demand	0.01	0.00	3.71	**< 0.001**
Video Gaming Experience	-0.08	0.04	-2.29	**0.022**
Agent Report Modality	-0.21	0.11	-1.86	0.062
Military Experience	-0.67	0.39	-1.72	0.086
Gender	-0.07	0.12	-0.63	0.530
Age	0.00	0.02	-0.07	0.944

Note. Significant values are in bold.

The residuals did not show signs of dependence between the errors. The predicted vs. observed values plot confirmed the poor fit of the model (the plotted line is the fitted line based on maximum likelihood; Appendix F **Error! Reference source not found.**).

Hypotheses

The null Hypothesis 1, of the Human(/Agent) Qualities, all factors are equally important to task performance, was rejected. Video gaming experience was the most important contributor to hit rate, followed by military experience and gender. Age was the least important predictor.

Hypothesis 2, all NASA-TLX subscales (Task Perception) contribute equally to task performance, was rejected. Temporal demand was the most important contributor to hit rate within this subset. This was not surprising since each participant ran through three 32-minute scenarios.

The null Hypothesis 3, Task Composition, Perception of Task, and Human/Agent Qualities are equally important to task performance, was rejected. In Study C, the most important predictor was formed by Task Perception, specifically the perceived load related to time (NASA-TLX temporal demand subscale), followed by Human (/Agent) Qualities (video gaming experience), and lastly Task Composition (agent report modality).

Based on the analyses and bootstrap, the hypothesized Core model in Study C is presented in Error! Reference source not found. In study C, Task Perception, driven by the temporal demand subscale of the NASA-TLX, was the most important contributor to hit rate, followed by Human(/Agent) Qualities, based on video gaming experience, and lastly Task Composition (agent report modality).

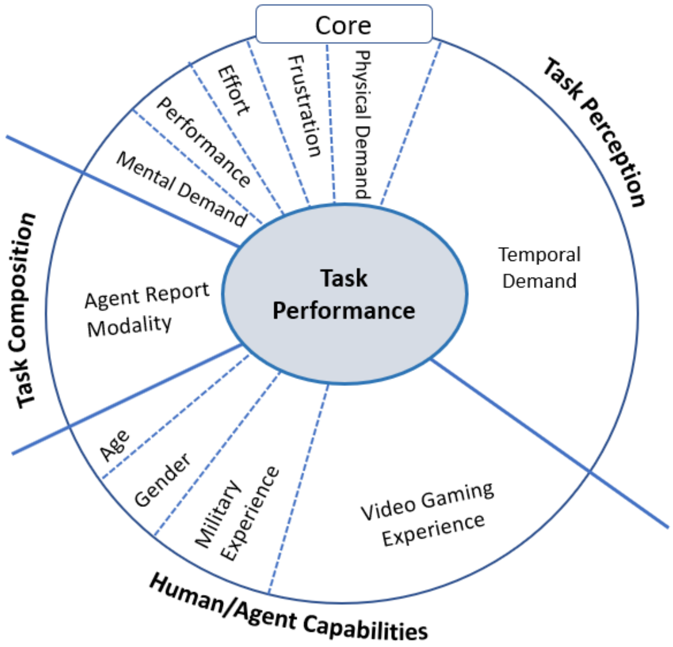


Figure 34. Updated Core model based on results in Study C.

Note. The size of the sections of the pie represent the relative importance of the factor to task performance.

Study D

Hypotheses Study D

Study D manipulated the delivery frequency of agent reports (immediate vs. interval) and the modality through which the report was delivered (auditory vs. visual). The threat detection task occurred at a constant event rate of 60 characters/minute with a low threat probability of 0.09-0.10. Average hit rate was 0.67 (*SD* = 0.11), which was significantly lower than hit rate in Study A, B, and C (see Appendix F. Data regarding video gaming were not available and all participants were non-military/students. Additionally, agent qualities were not available. The collaborative agent was simulated to be a fully autonomous and 100% reliable, intelligent, embodied agent that scouted the outer cordon for threats and contraband. The null hypotheses are visually presented in **Figure 35**.

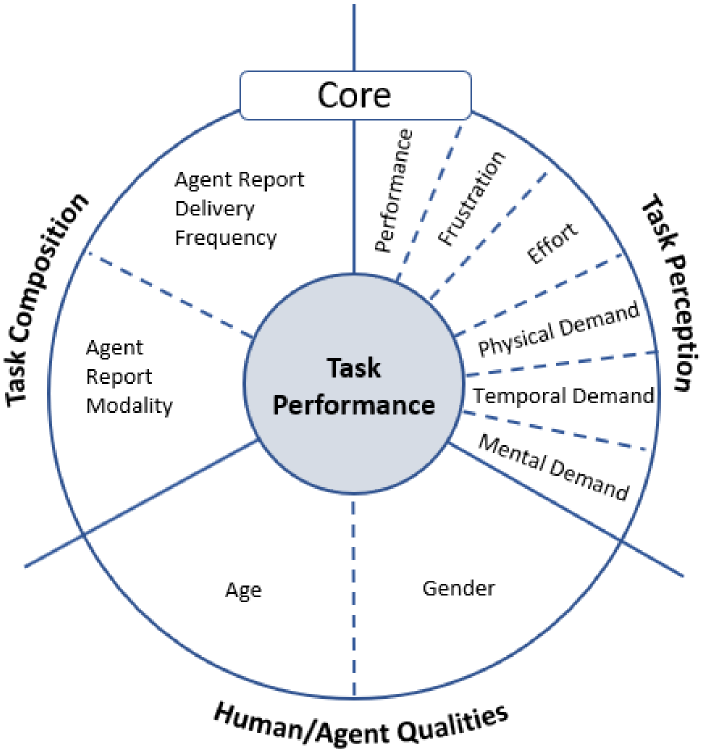


Figure 35. Visual representation of hypotheses in Study D.

Note. The size of the sections of the pie represent the relative importance of the factor to task performance. The null hypotheses are that all factors and categories are of equal importance to task performance.

The null hypotheses were as follows:

Hypothesis 1. Of the Human/Agent Qualities, all factors are equally important to task performance.

Hypothesis 2. All NASA-TLX subscales (Task Perception) contribute equally to task performance.

Hypothesis 3. The Task Composition factors contribute equally to task performance.

Hypothesis 4. Task Composition, Perception of Task, and Human/Agent Qualities are equally important to task performance.

Dominance Analysis Study D

Linearity was established between predictors and the response variable (Appendix F, **Error! Reference source not found.**).

Human (/Agent) Qualities

The Human(/Agent) Qualities' overall model's pseudo R^2 was low (see **Table 58** in Appendix F). Only age and gender were compared in the dominance analysis.

Complete Dominance

The dominance analysis pattern was clear for Human(/Agent) Qualities in Study D. Gender completely dominated age (**Table 32**), which indicated that gender also dominated gender over lower dominance levels, i.e. conditional and general dominance (Error! Reference source not found., Appendix F).

Table 32

Complete dominance results Human/Agent Qualities in Study D.

Variable	Gender	Age
Gender	0.5	1
Age	0	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

Task Perception

The overall model's pseudo R^2 for Task Perception was low (see **Table 59** in Appendix F).

Complete Dominance

As shown in **Table 33**, of the NASA-TLX subscales, mental demand completely dominated all other subscales and the performance subscale dominated frustration and physical demand. Lastly, the effort scale dominated the physical demand scale.

Table 33

Complete dominance results Task Perception in Study D.

Variable	Mental Demand	Performance	Effort	Temporal Demand	Frustration	Physical Demand
Mental Demand	0.5	1	1	1	1	1
Performance	0	0.5	0.5	0.5	1	1
Effort	0	0.5	0.5	0.5	0.5	1
Temporal Demand	0	0.5	0.5	0.5	0.5	0.5
Frustration	0	0	0.5	0.5	0.5	0.5
Physical Demand	0	0	0	0.5	0.5	0.5

Conditional Dominance

In addition to the complete dominance pattern, temporal demand and frustration also dominated physical demand. **Figure 36** shows that temporal demand and effort were potential suppressors, as the additional contribution increased as the *k* size of the subset models grew.

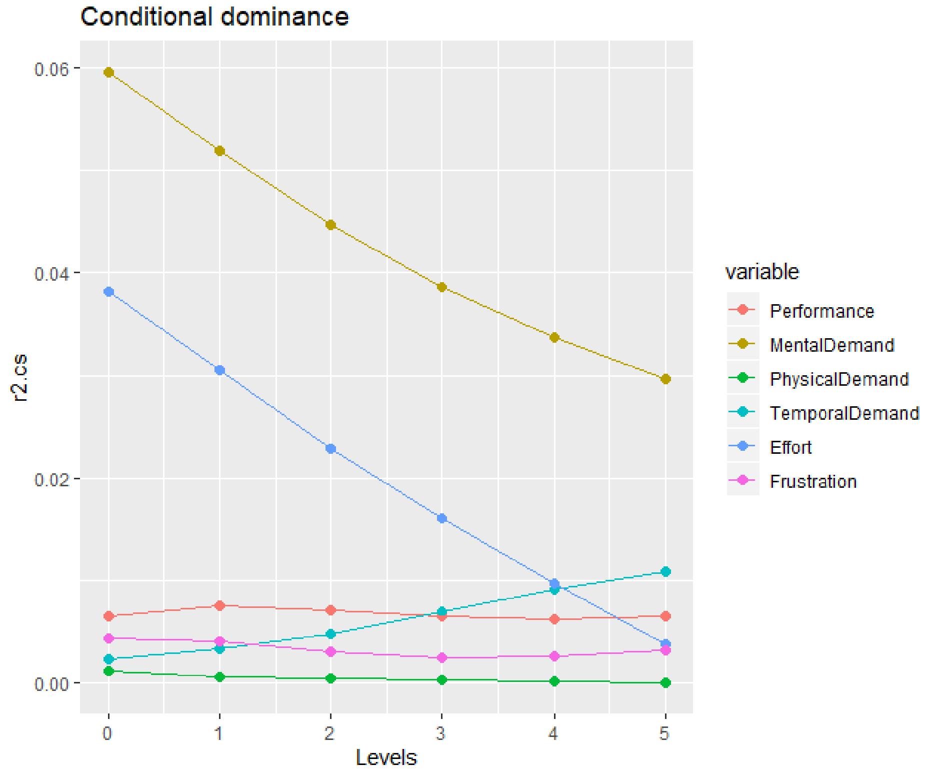


Figure 36. Conditional dominance results Task Perception in Study D.

Note. The plot shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model. The unique contribution of a predictor should monotonically decrease with increasing subset model sizes (Azen & Budescu, 2003; Budescu, 1993). An increase, such as here for temporal demand, indicates that this variable is a potential suppressor, gaining importance through collinearity with other predictors in the model rather than through direct association with the outcome variable (Azen & Budescu, 2003).

General Dominance

The general dominance values, as plotted in **Figure 37**, indicated no additional dominant predictors. The subscale of lowest importance to hit rate was physical demand.

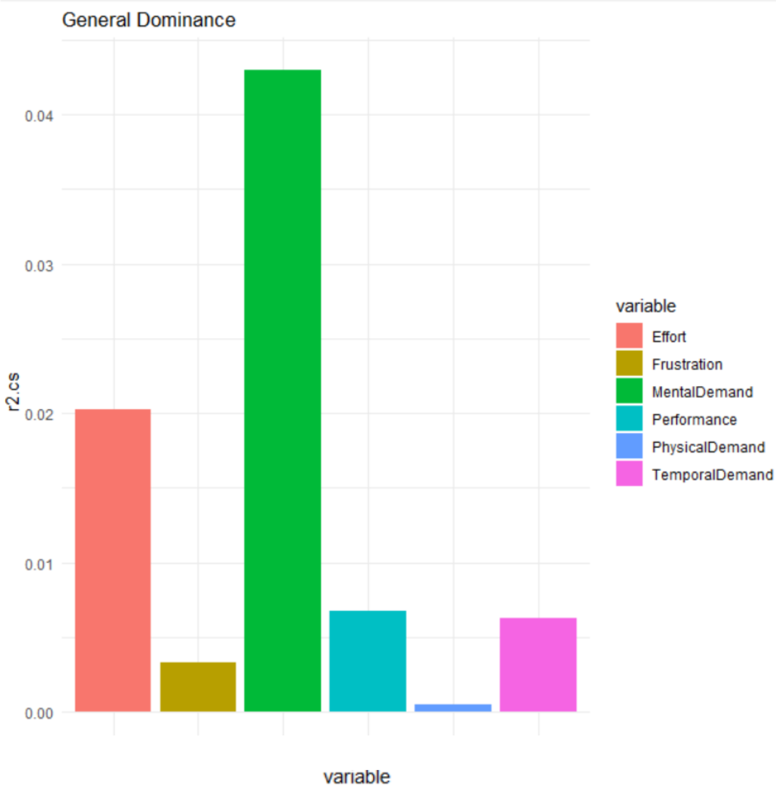


Figure 37. General dominance results Task Perception in Study D.

Note. The general dominance bar graph shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) averaged over all possible subset model sizes.

Task Composition

The overall model’s pseudo R^2 of Task Composition was low (see **Table 60** in Appendix F). Only two predictors were manipulated in study D and thus compared in the dominance analysis.

Complete Dominance

Agent report modality completely dominated the delivery frequency of the reports (**Table 34**). This indicated that agent report modality also dominated delivery frequency over lower dominance levels, i.e. conditional and general dominance (**Error! Reference source not found.**, Appendix F).

Table 34

Complete dominance results Task Composition in Study D.

Variable	Agent Report Modality	Delivery Frequency
Agent Report Modality	0.5	1
Delivery Frequency	0	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

Full Model

Lastly, DA was conducted on the full model, such that

$$Winsorized\ Hit\ Rate \sim Age + Gender + Mental\ Demand + Agent\ Report\ Modality \\ + Delivery\ Frequency$$

In DA, the human predictor that was not important to hit rate, i.e., age, was held constant. The overall model pseudo R^2 was low (see **Table 61** in Appendix F).

Complete Dominance

As shown in **Table 35**, gender completely dominated all other predictors in the model, followed by mental demand. Agent report modality also completely dominated agent report delivery frequency. Since this confirmed the earlier finding reported in Human/Agent Qualities, delivery frequency was not further evaluated and dropped from analyses.

Table 35

Complete dominance results Full Model in Study D.

Variable	Gender	Mental Demand	Agent Report Modality	Delivery Frequency
Gender	0.5	1	1	1
Mental Demand	0	0.5	1	1
Agent Report Modality	0	0	0.5	1
Delivery Frequency	0	0	0	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

Since complete dominance was established between all predictors, such that gender > mental demand > agent report modality > delivery frequency, the conditional and general dominance analyses did not yield any additional results (Error! Reference source not found., Appendix F).

Bootstrap

Delivery frequency was not further evaluated and dropped from analyses, since it consistently was not an important predictor of hit rate.

The results of *S* = 1000 bootstrap samples indicated that the confidence that gender would dominate mental demand and agent report modality in the actual population was high, varying from 73.2% to 88.9% (

Table 36). This level of confidence grew higher as the level of dominance decreased to conditional and general dominance. Since complete dominance was established between all predictors, such that gender > mental demand > agent report modality > delivery frequency, the bootstrapped conditional and general dominance values did not yield any additional results.

Table 36

Bootstrap results Full Model in Study D.

Variable *i*	Variable *j*	D_{ij}	$\overline{D}_{\mathrm{ij}}$	SE(D_{ij})	P_{ij}	P_{ji}	P_{noij}	Reproducibility
Complete Dominance								
Gender	Mental Demand	1	0.769	0.400	**0.732**	0.195	0.073	**0.732**
Gender	Agent Report Modality	1	0.918	0.248	**0.889**	0.054	0.057	**0.889**
Mental Demand	Agent Report Modality	1	0.730	0.405	**0.663**	0.203	0.134	**0.663**
Conditional Dominance								
Gender	Mental Demand	1	0.767	0.404	**0.734**	0.201	0.065	**0.734**
Gender	Agent Report Modality	1	0.919	0.252	**0.897**	0.059	0.044	**0.897**
Mental Demand	Agent Report Modality	1	0.737	0.414	**0.690**	0.217	0.093	**0.690**
General Dominance								
Gender	Mental Demand	1	0.775	0.418	**0.775**	0.225	0.000	**0.775**
Gender	Agent Report Modality	1	0.917	0.276	**0.917**	0.083	0.000	**0.917**
Mental Demand	Agent Report Modality	1	0.736	0.441	**0.736**	0.264	0.000	**0.736**

Note. D_{ij} is the dominance value of the original analyses, wherein $D_{ij} = 1 – D_{ji}$. Although each pair has two possible orders (*ij* and *ji*), only one order is shown to reduce redundancy

The $P_{..}$ values indicate the proportion of the S = 1000 bootstrap sample that replicated D_{ij}, such that $P_{ij} = Pr(D_{ij} = 1)$, $P_{ji} = Pr(D_{ij} = 0)$, $P_{noij} = Pr(D_{ij} = 0.5)$. The reproducibility value refers to the proportion of the bootstrap sample that replicated D_{ij}.

Bold values imply a reference to the dominance value from the sample (D_{ij}).

Model Fit Evaluation: Regression

The most important predictors were combined into a hierarchical beta regression model, wherein all human variables were preserved, as they would always be present in the natural world as well:

$$winsorized\ Hit\ Rate \sim Gender + Mental\ Demand + Agent\ Report\ Modality + Age$$

The pseudo R^2 of the model was 0.243 and was significantly better at predicting hit rate than the null model, $\chi^2(6) = 111.34$ $p < 0.001$. Moreover, the important predictors, i.e., all except age, were significant as shown in **Table 37**.

Table 37

Results beta regression on Full Model in Study D.

	Beta Coefficient	SE	z-value	Pr(>|z|)
Intercept	0.00	0.14	-0.01	**0.991**
Gender	0.37	0.08	4.59	< **0.001**
Mental Demand	0.01	0.00	3.66	< **0.001**
Agent Report Modality	0.2	0.08	2.52	**0.012**
Age	0.01	0.01	0.39	0.698

Note. Significant values are in bold.

The residuals did not show signs of dependence between the errors (Appendix F, **Error! Reference source not found.**). The predicted vs. observed values plot showed a large number of observations that were deviated from the fitted line based on maximum likelihood. However, the model looked superior compared to the models of studies A, B, and C.

Hypotheses

The null Hypothesis 1, of the Human/Agent Qualities, all factors are equally important to task performance, was rejected. Gender completely dominated age.

The null Hypothesis 2, all NASA-TLX subscales (Task Perception) contribute equally to task performance, was rejected. The most important variables to hit rate in terms of Task Perception was the mental demand subscale.

The null Hypothesis 3, the Task Composition factors contribute equally to task performance, was rejected. Agent report modality was more important than the delivery frequency.

The null Hypothesis 4, Task Composition, Perception of Task, and Human/Agent Qualities are equally important to task performance, was rejected. In study D, Human/Agent Qualities (gender) were most important to hit rate, followed by Task Perception (mental demand), and Task Composition (agent report modality) last.

Based on the analyses and bootstrap, the hypothesized Core model in study D is presented in **Figure 38**. In study D, Human Qualities (gender) was most important to hit rate, closely followed by Task Perception (mental demand), and lastly Task Composition (agent report modality).

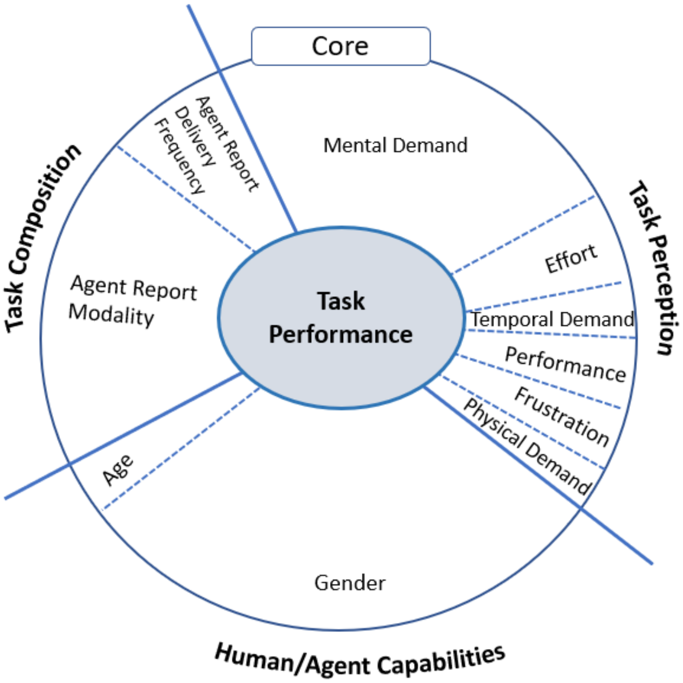


Figure 38. Updated Core model based on the results in Study D.

Note. The size of the sections of the pie represent the relative importance of the factor to task performance.

<u>Overall Results</u>

The overall results of the most importance factors of hit rate are captured in **Table 38**, wherein the darkness of the color indicates the level of importance. The results were very different between studies. This may be in part due to the different independent variables that were manipulated within each study. For instance, study A manipulated visual complexity, whereas study D manipulated agent report delivery frequency. However, the studies also differed in the content of the agent reports, threat criterion, and design of the humanoid character models (see Appendix A through D). These factors could not be accounted for in the present research effort, as they were either fully nested between the studies or unidentified (in case of agent report content for Study D). Other differences between the studies were in terms of event rate and threat probability, two factors that influence task difficulty (Wickens & Hollands, 2000).

Table 38

Data matrix with dominance results in color.

Manipulation	Experimental Study (Source: Abich et al., 2017; Barber et al., 2017; Barber et al., 2019; Bendell et al., 2020; Kopinsky; 2017))									
	Study A.1	Study A.2	Study B.1	Study B.2	Study B.3	Study C.1	Study C.2	Study C.3	Study D.1	Study D.2
Task Composition										
Event rate										
15 characters/min.	•	•		•		•*	•*	•*		
30 characters/min.			•		•					
60 characters/min.				•		•*	•*	•*	•	•
Signal likelihood										
0.09-0.10									•	•
0.12-0.13			•	•	•	•	•	•		
0.13-0.14	•	•								•
Task duration										
5 minutes				•						
10 minutes			•		•					
12 minutes	•	•								
15-16 minutes									•	
32 minutes						•	•	•		•
Agent Task Type										
Receive Report	•	•				•	•	•	•	•
Pull Report			•	•						
Visual Complexity										
Basic	•									
Enhanced		•								

•* Coded as NA

Note. Darker hue indicates higher importance.

Manipulation	Experimental Study (Source: Abich et al., 2017; Barber et al., 2017; Barber et al., 2019; Bendell et al., 2020; Kopinsky; 2017)									
	Study A.1	Study A.2	Study B.1	Study B.2	Study B.3	Study C.1	Study C.2	Study C.3	Study D.1	Study D.2
Task Composition										
Agent Report Delivery Frequency										
Interval									•	
Immediate										•
Agent Report Modality										
Auditory									•	•
Visual	•	•	•	•	•				•	•
Single-Adaptive						•	•			
Dual								•		
Human/Agent Qualities										
Agent Type										
Legged	•									
Wheeled		•								
Demographics										
Age	•	•	•	•	•	•	•	•	•	•
Gender	•	•	•	•	•	•	•	•	•	•
Experience										
Military Experience			•	•	•					
Video Gaming Experience	•	•				•	•	•		
Task Perception										
Perceived Workload (NASA-TLX)										
Mental Demand	•	•	•	•	•	•	•	•	•	•
Physical Demand	•	•	•	•	•	•	•	•	•	•
Temporal Demand	•	•	•	•	•	•	•	•	•	•
Effort	•	•	•	•	•	•	•	•	•	•
Frustration	•	•	•	•	•	•	•	•	•	•
Performance	•	•	•	•	•	•	•	•	•	•

The results of the conducted analyses in this effort suggest that these differences in task difficulty matter. Of the three studies that all had higher hit rate (> 0.90), event rate was significant in the study that manipulated this variable (study B). To understand the importance of predictors in light of task difficulty differences between studies, another DA was conducted wherein these factors were kept constant. This method ensured that the error associated with event rate and signal probability was accounted for. Age was also held constant, as the variable cannot be factored out in the real world.

However, even though event rate and signal probability were nested between the studies, a new variable could be created to account for their variance. Event rate and signal probability were combined into a new independent variable: threat conspicuity (**Table 39**). Threat conspicuity refers to the ease of perceiving a threat under conditions of event rate and threat probability. Low threat conspicuity was defined by high event rate (60 characters/minute) and low threat probability (0.09-0.10). High threat conspicuity was defined by low event rate (15 characters/minute) and high threat probability (0.13-0.14). Anything in between was defined as medium threat conspicuity.

Table 39

Operationalization of threat conspicuity.

Threat Conspicuity Level	Event Rate	Threat Probability
Low threat conspicuity	60 characters/min.	0.09-0.10
Medium threat conspicuity	30 characters/min. OR alternating 15 – 60 characters/min.	0.12-0.13
High threat conspicuity	15 characters/min.	0.13-0.14

Combined Studies: Constrained DA

Hypothesis

Similar to the individual studies, the null hypothesis in the combined studies was: Task Composition, Perception of Task, and Human/Agent Qualities are equally important to task performance (**Figure 39**).

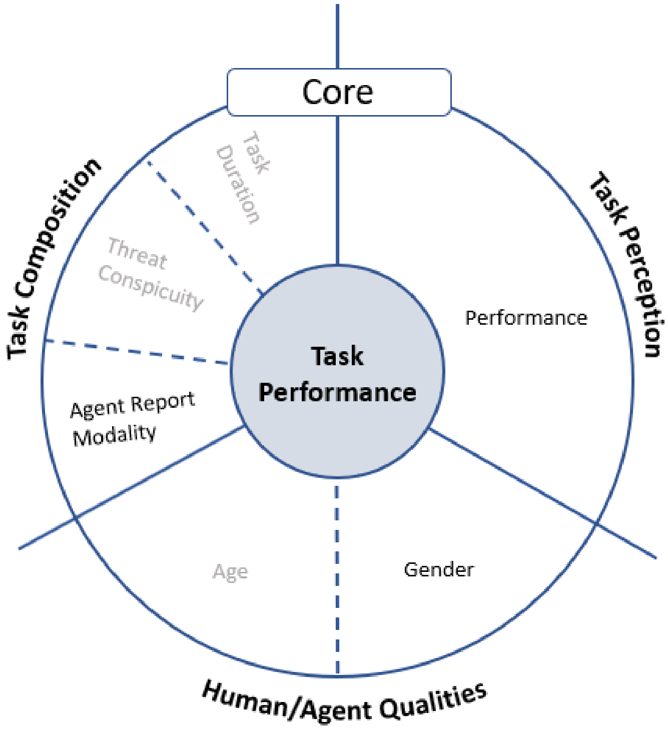


Figure 39. Visual representation of hypothesis in combined studies.

Note. The size of the sections of the pie represent the relative importance of the factor to task performance. The null hypotheses are that all factors and categories are of equal importance to task performance. Constant factors are grayed out.

Herein, based on the DAs on the individual studies, only the most important predictors of hit rate were included. The potential suppressors from the individual studies, temporal demand, mental demand, and military experience, were indeed also suppressors in the combined analyses (see **Table 62** and **Error! Reference source not found.** in Appendix F). These suppressor variables were dropped from the overall analyses, to gain insight into the most important predictors.

Furthermore, due to the large number of missing values between studies, agent type, visual complexity, agent report delivery frequency, video gaming experience, and task type were excluded from the dominance analyses, as they missing values bias the results through elimination of observations. None of these variables were the primary important predictors of hit rate in DA of the individual studies.

Since threat conspicuity and task duration could not be analyzed, as they are nested between the studies, they were kept constant and DA on the full model was conducted. Age was also kept constant, since it was not an important predictor in any of the studies yet could not be excluded in the natural world in a HAT context. The constants are greyed out in **Figure 39**.

Dominance Analysis Combined Studies

Full Model: Constrained DA

DA was conducted on the full model, such that:

$$winsorized\ Hit\ Rate \sim Performance + Gender + Agent\ Report\ Modality$$
$$+ Threat\ Conspicuity + Task\ Duration + Age$$

Herein, the constants were threat conspicuity, task duration, and age.

Complete Dominance

Table 40 shows that when combining the studies, and keeping threat conspicuity, task duration, and age constant, agent report modality (Task Composition) the most important predictor of hit rate. It completely dominated all other factors. Complete dominance could not be established for the NASA-TLX performance subscale (Task Perception) and gender (Human/Agent Qualities).

Table 40

Complete dominance results in Combined Studies.

Variable	Agent Report Modality	Performance	Gender
Agent Report Modality	0.5	1	1
Performance	0	0.5	0.5
Gender	0	0.5	0.5

Note. A dominance value of 1 indicates dominance of the row variable of the column variable; 0 indicates dominance of the column variable over the row variable; 0.5 indicates that dominance could not be determined.

Conditional Dominance

The conditional dominance figure (**Figure 40**) elucidates that dominance could not be established between performance and gender. Both predictors have a similar low contribution to hit rate.

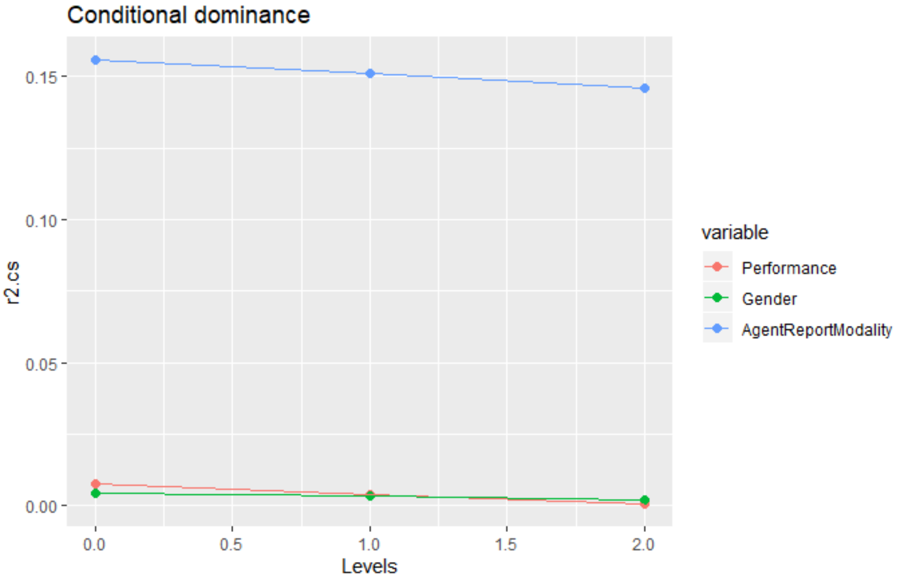


Figure 40. Conditional dominance results in Combined Studies.

Note. The plot shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model.

General Dominance

As shown in **Figure 41**, in terms of general dominance performance dominated gender.

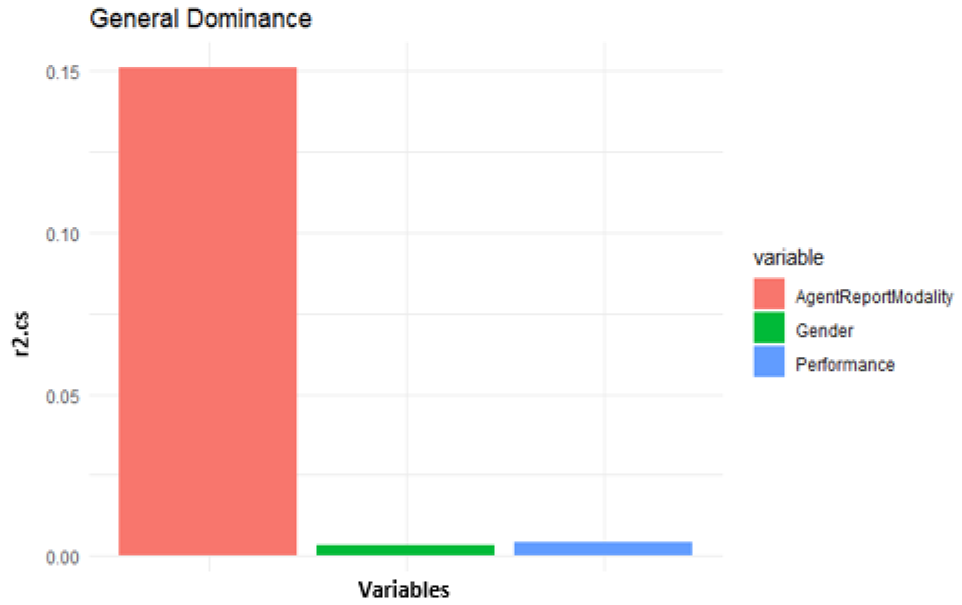


Figure 41. General dominance results in Combined Studies.

Note. The general dominance bar graph shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) averaged over all possible subset model sizes.

Bootstrap

The results of S = 1000 bootstrap samples indicated that the confidence that agent report modality would dominate performance and gender was 100% (**Table 41**). Dominance between performance and gender was undetermined for the complete and conditional dominance levels, which was replicated in 72.3% - 78.3% of the bootstrap samples. Dominance tended toward performance, as seen by the higher mean (resp. 0.609 and 0.639). Indeed, dominance of performance over gender was confirmed on the general dominance level in 100% of the bootstrap samples.

Table 41

Bootstrap results in Combined Studies.

Variable *i*	Variable *j*	D_{ij}	$\overline{D}_{\mathrm{ij}}$	SE(D_{ij})	P_{ij}	P_{ji}	P_{noij}	Reproducibility
Complete Dominance								
Performance	Gender	0.5	0.609	0.206	0.217	0.000	**0.783**	**0.783**
Performance	Agent Report Modality	0	0.000	0.000	0.000	**1.000**	0.000	1.000
Gender	Agent Report Modality	0	0.000	0.000	0.000	**1.000**	0.000	1.000
Conditional Dominance								
Performance	Gender	0.5	0.639	0.224	0.277	0.000	**0.723**	**0.723**
Performance	Agent Report Modality	0	0.000	0.000	0.000	**1.000**	0.000	1.000
Gender	Agent Report Modality	0	0.000	0.000	0.000	**1.000**	0.000	1.000
General Dominance								
Performance	Gender	1	1.000	0.000	**1.000**	0.000	0.000	**1.000**
Performance	Agent Report Modality	0	0.000	0.000	0.000	**1.000**	0.000	**1.000**
Gender	Agent Report Modality	0	0.000	0.000	0.000	**1.000**	0.000	**1.000**

Note. D_{ij} is the dominance value of the original analyses, wherein $D_{ij} = 1 - D_{ji}$. Although each pair has two possible orders (*ij* and *ji*), only one order is shown to reduce redundancy

The $P_{..}$ values indicate the proportion of the S = 1000 bootstrap sample that replicated D_{ij}, such that $P_{ij} = Pr(D_{ij} = 1)$, $P_{ji} = Pr(D_{ij} = 0)$, $P_{noij} = Pr(D_{ij} = 0.5)$. The reproducibility value refers to the proportion of the bootstrap sample that replicated D_{ij}.

The bold values imply a reference to the dominance value from the sample (D_{ij}).

Model Fit Evaluation

The most important predictors were combined into a hierarchical beta regression model, such that:

$$winsorized\ Hit\ Rate \sim Agent\ Report\ Modality + Performance + Gender + Age + Threat\ Conspicuity + Task\ Duration$$

The beta coefficients and significance testing are presented in **Table 42**. The pseudo R^2 of the model was 0.520, which was a considerable improvement compared to the fits of the full models in the individual studies. The model was significantly better at predicting hit rate than the null model, $\chi^2(10)$ =1079.80, p < 0.001. Furthermore, significance testing of the beta coefficients indicated that agent report modality was significant, while performance and gender were not significant. Additionally, threat conspicuity and task duration were indeed significant in predicting hit rate, which signified the importance of taking the variables into account.

Table 42

Results beta regression on model in Combined Studies.

	Beta Coefficient	SE	z-value	Pr(>|z|)
Intercept	-1.52	0.27	-5.66	**< 0.001**
Agent Report Modality (Dual)	-3.74	0.35	-10.74	**< 0.001**
Agent Report Modality (Single Adaptive)	-3.65	0.40	-9.15	**< 0.001**
Agent Report Modality (Visual)	0.23	0.09	2.57	**0.010**
Performance	-0.00	0.00	-0.91	0.361
Gender	0.10	0.06	1.78	0.076
Age	0.01	0.01	0.76	0.446
Threat Conspicuity	3.42	0.19	17.68	**0.000**
Task Duration	0.13	0.01	9.62	**0.000**

Note. Significant values are in bold.

Lastly, the residuals and predicted vs. observed values plots indicate that there was some grouping around the errors (Error! Reference source not found., Appendix F). This was most likely related to the differences in hit rate between studies A, B and C on the one hand, and study D on the other hand. This notion was also suggested in the predicted vs. observed values plot, wherein two groups of observations were present.

Hypothesis

The null Hypothesis, Task Composition, Perception of Task, and Human/Agent Qualities are equally important to task performance, was rejected. When combining the four studies with threat conspicuity (**Table 39**) and non-important human variables kept constant, Task Composition (agent report modality) was most important to hit rate. Both Task Perception (performance), and Human/Agent Qualities (gender) were of little importance to task performance.

Based on the analyses, the Core model was best represented as shown in **Figure 42**Error! Reference source not found.. Task Composition factors were the most important contributors to task performance, with little contribution of Task Perception and Human(/Agent) Factors.

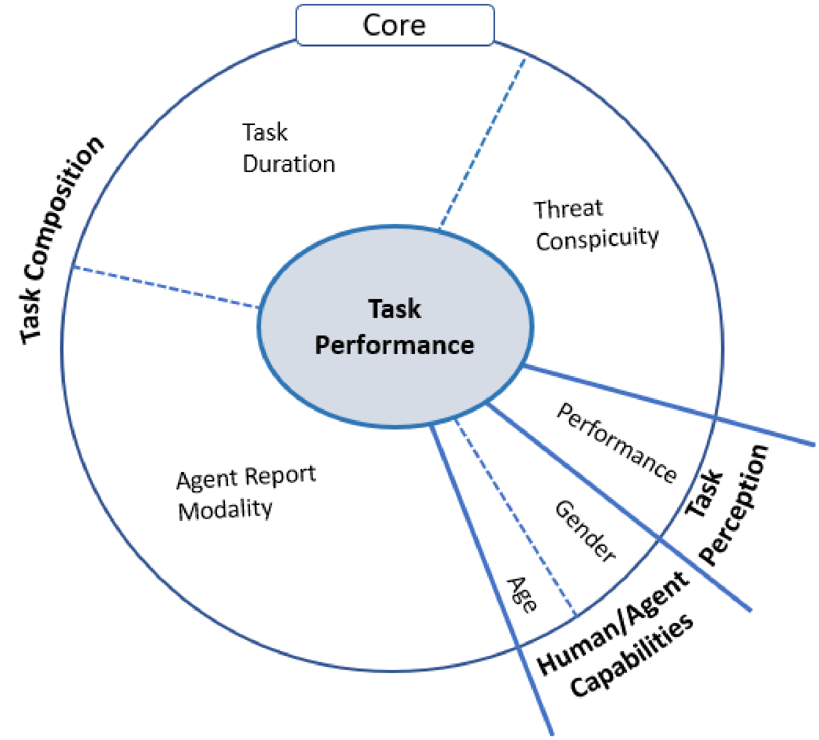


Figure 42. Updated Core model based on the results of the Combined Studies.

Note. The size of the sections of the pie represent the relative importance of the factor to task performance.

CHAPTER FIVE: DISCUSSION

The main objectives of the current effort were to (1) develop a model of military HAT performance and (2) to develop an approach to validate the model and apply this method to test the proposed model against empirical data. The experimental data was borrowed from studies conducted for the RCTA program (Childers et al., 2016), reported by Abich et al. (2017), Barber et al. (2018), Barber et al. (2019), Bendell et al. (2020), and Kopinsky (2017).

Objective 1: Model of Simulated Military Human-Agent Teaming

To develop the model, important constructs in relation to HAT performance were identified and integrated into a comprehensive model centered around task performance (**Error! Reference source not found.**). The proposed model consists of three layers. The outer Layer has the least direct impact on Task Performance: The Environmental Layer. This Layer consists of environmental variables, such as the scenario in which the mission takes place, environmental conditions, and overall awareness of the task, relationship, environment, and performance (situation awareness). The Relationship Layer focuses on the relationship between the human and agent teammate(s), with constructs as mutual trust, mental models, and transparency. The Core model directly impacts Task Performance and consists of Task Components, Task Perception, and the Qualities the Human/Agent bring to the team. The Core model was validated in this effort. Variables within each category, and between the three categories, were hypothesized to be of equal importance to task performance.

Task Composition refers to elements of the task and is known to affect task performance (Green, 1993; Lu et al., 2013; See et al., 1995; Szalma et al., 2008). Some of the most common analyzed components of task composition in relation to HAT performance are event rate, signal probability, and modality (Teo et al., 2018). However, other task components may be manipulated as well, as indicated by the data analyzed in the present effort.

The way in which individuals perceive the task (Task Perception) also affects task performance. Task Perception relates to the individual's compensatory strategies, or self-regulation, to modulate performance (Hancock & Warm, 1989; Hockey, 1997; Matthews, Winter, et al., 2019). Through perception of increased demand and potential drops in performance, individuals make a strategic decision in terms of up- or downregulating their information processing resources or effort toward the task (Hockey, 1997). In this manuscript, Task Perception was operationalized as the score on the NASA-TLX subscales, which are reflective of perception of cost incurred by the task, or perceived workload (Hart & Staveland, 1988).

Lastly, Human/Agent Qualities refer to the qualities that each entity brings to the team. Human qualities include differences in personality, experience, age, and gender. Agent qualities pertain to characteristics such as morphology, level of automation, and reliability. In the borrowed experimental studies, the agent was simulated to be 100% reliable, fully autonomous and capable of performing its task.

Objective 2: Model Validation Approach

Data from four simulated military HAT studies were taken from previous efforts under the RCTA (reported by Abich et al., 2017; Barber et al., 2018; Barber et al., 2019; Bendell et al., 2020; Kopinsky, 2017) to validate the Core model. Herein, participants performed a continuous threat detection task, while an autonomous agent conducted its own task out-of-sight and reported intermittently to the human team member. Task performance was operationalized in terms of accuracy of the primary mission, which was threat detection. Threat detection was performed by the human teammate. Performance was measured as hit rate, i.e., the ratio of correctly identified threats to number of total threats available. As a proportional variable, the data followed a beta-distribution (Ferrari & Cribari-Neto, 2004).

To test this relative-importance based model against beta-distributed empirical data, a validation approach was proposed:

1. Apply dominance analysis (DA) on beta regression models to determine the most important contributors to the outcome variable. DA compares the unique additional contribution of each predictor to the outcome variable in all regression subset model sizes (Azen & Budescu, 2003; Budescu, 1993).

2. Establish the robustness and generalizability of the dominance results by bootstrapping the dominance values (Azen & Budescu, 2003; Efron, 1981).

3. Combine the most important predictors into a hierarchical beta regression model and evaluate the fit of the model (Ferrari & Cribari-Neto, 2004).

As part of the development of the validation approach, different pseudo R^2 were as goodness-of-fit estimators of dominance analysis based on beta regression models. Cox and Snell's pseudo R^2 (Cox & Snell, 2018) was the most appropriate statistics and was integrated in a dominanceanalysis package that is now available in R for public use (Bustos & Countinho, 2019).

In the following sections, the results of the analyses for each study are discussed in terms of hypothesized and reported results. Subsequently, the overarching implications are discussed in relation to the proposed model and modeling approach.

Study A

In Study A, agent morphology type (animal-like vs. machine-like; category Human/Agent Quality) and visual complexity (basic vs. enhanced visual cues; category Task Composition) were manipulated. The threat detection task was conducted under low event rate with high threat probability. Average hit rate (*M* = 0.97, *SD* = 0.05) was within the military performance standard (Naval Education and Training Command, 2009). No DA was conducted on Task Composition, since one Task Composition predictor was manipulated: visual complexity. The results are summarized in **Table 43**.

Table 43

Summarized result of dominance analyses in Study A.

Null Hypothesis	Importance	Result
1. All Human/Agent Qualities contribute equally to hit rate	Agent morphology type > age > military experience. Video gaming experience and gender suppressors	Null hypothesis 1 not rejected
2. All Task Perception variables contribute equally to hit rate	Performance. Effort, mental demand, and temporal demand suppressors	Null hypothesis 2 not rejected
3. Task Composition, Perception of Task, and Human/Agent Qualities are equally important to hit rate	Performance (TP)[a] > visual complexity (TC)[a] > agent report modality (TC)[a] > age (H/A)[a]	Null hypothesis 3 not rejected

Note. In the column Importance the > symbol signifies the dominance of the variable over the others. For instance, agent morphology type dominated age and military experience, while age also dominated military experience.

[a] TP = Task Perception, TC = Task Composition, H/A = Human/Agent Qualities.

None of the null hypotheses were rejected since statistical significance was not established for the differences in unique additional contribution between the predictors. For the Human/Agent Qualities, DA indicated that agent morphology was the most important contributor to hit rate, followed by age, and military experience. However, this difference was not statistically significant in the subsequent analyses. Two potential suppressors were identified, video gaming experience and gender. Suppressor variables gain importance over different model subset sizes through collinearity with other predictors in the model, rather than through their direct association with the outcome variable (Azen & Budescu, 2003; Smith et al., 1992). Thus, video gaming experience and gender were not yielding an important unique contribution to hit rate in Study A.

DA of the NASA-TLX subscales, reflective of Task Perception, indicated that in this study the performance subscale was qualitatively the most important predictor of hit rate. The effort, mental demand, and temporal demand subscales were suppressor variables. However, since subsequent statistical analyses did not establish significance for these differences, the null Hypothesis 2, all Task Perception variables are equally important to hit rate, was not rejected.

The qualitatively most important predictors were combined into a full statistical model. Herein, non-dominant human/agent variables were held constant to account for their explained variance without analyzing their dominance effects (Azen & Budescu, 2003). DA on the full model, holding military experience, video gaming experience, and gender constant, indicated that the NASA-TLX performance subscale (Task Perception) was the most important predictor of hit rate. Visual complexity (Task Composition) was the second most important variable, followed by agent type and age (Human/Agent Qualities). Under the parameters set by this study, e.g., low event rate and high signal probability, human or agent variables contributed little to performance. However, subsequent analyses were not significant, thus, the null hypothesis that Task Composition, Task Perception and Human/Agent Qualities contributed equally to task performance was not rejected.

The dominance pattern was not robust based on the bootstrap results. The strongest level of dominance results (complete dominance) were replicated in 58.5% of the bootstraps at most, which indicated that the confidence that this result will be replicated in the natural world was low. The lack of robust generalizability to the population was most likely explained by the poor model fit, as indicated by the low pseudo R^2 (0.038) and lack of significance when this full model was compared to the null model.

Study B

In study B, event rate and agent task type (pull vs. receive agent report) were manipulated. The threat detection task was conducted under low event rate, medium event rate, or high event rate. Task duration was either five or ten minutes, depending on the scenario. This study collected data at a university and at a military base, to understand the effects of differences in military experience to performance. The average hit rate was 0.95 (*SD* = 0.07) and within bounds of the military standard (Naval Education and Training Command, 2009). The hypotheses and results are summarized in **Table 44**.

Table 44

Summarized result of dominance analyses in Study B.

Null Hypothesis	Importance	Result
1. All Human/Agent Qualities contribute equally to hit rate	Military experience > Age & gender	Null hypothesis 1 rejected
2. All Task Perception variables contribute equally to hit rate	Performance. Effort and temporal demand suppressors	Null hypothesis 2 rejected
3. All Task Composition variables contribute equally hit rate	Event rate > Task duration > Task type	Null hypothesis 3 rejected
4. Task Composition, Perception of Task, and Human/Agent Qualities are equally important to hit rate	Event rate (TC) [a] > Performance (TP) [a] > Military experience (H/A) [a]	Null hypothesis 4 rejected

Note. In the column Importance the > symbol signifies the dominance of the variable over the others. For instance, event rate dominated task duration and task type, while task duration also dominated task type.

[a] TP = Task Perception, TC = Task Composition, H/A = Human/Agent Qualities.

DA on the Human/Agent Qualities, i.e., age, gender, and military experience, indicated that military experience was the most important predictor of hit rate. Null Hypothesis 1, all Human/Agent Qualities are equally important to hit rate, was rejected.

Of the Task Perception variables, the NASA-TLX performance subscale was the most important predictor of hit rate. The effort and temporal demand subscales were suppressor variables. Mental and physical demand were the least important predictors of hit rate. Null Hypothesis 2, all Task Perception variables are equally important to hit rate, was rejected.

DA of the Task Composition variables indicated that event rate completely dominated task type (push vs. pull reports) and task duration. Task duration and task type were not important to hit rate.

DA on the full model using the most important predictors, and the human variables as constants, indicated that event rate (Task Composition) was the most important predictor of hit rate. The NASA-TLX performance subscale (Task Perception) was the second most important variable. Human Qualities were the least important contributors to hit rate. The null hypothesis that Task Composition, Task Perception and Human/Agent Qualities were equally important to hit rate was rejected.

Of the full model, the complete dominance results were replicated in 81.0 to 99.3% of the bootstraps, which indicated that the confidence that this result will be replicated in the natural world was high. The bootstrapped general dominance values emphasized the importance of event rate and the performance subscale. The fit of the full model was poor (pseudo $R^2 = 0.107$) but significant compared to the null model. This significance was driven by event rate, emphasizing the importance of Task Composition on task performance.

Study C

In study C, event rate (low vs. high) and agent report modality (single-adaptive vs. dual) were manipulated. However, event rate was not evaluated for importance since experimental blocks could not be combined. Some blocks that were expected to be identical resulted in significantly different results (Barber et al., 2019). The average hit rate was again high, 0.95 (*SD* = 0.07), within the bounds of the military performance standard (Naval Education and Training Command, 2009). Since agent report modality was the only manipulated Task Composition variable, no DA was conducted on Task Composition alone. The hypotheses and results are summarized in **Table 45**.

Table 45

Summarized result of dominance analyses in Study C.

Null Hypothesis	Importance	Result
1. All Human/Agent Qualities contribute equally to hit rate	Video gaming experience > gender > age Military experience suppressor	Null hypothesis 1 rejected
2. All Task Perception variables contribute equally to hit rate	Temporal demand	Null hypothesis 2 rejected
3. Task Composition, Perception of Task, and Human/Agent Qualities are equally important to hit rate	Temporal demand (TP) [a] > Video gaming experience (H/A) [a] > Agent report modality (TC) [a]	Null hypothesis 3 rejected

Note. In the column Importance the > symbol signifies the dominance of the variable over the others. For instance, video gaming experience dominated gender and age, while gender also dominated age.

[a] TP = Task Perception, TC = Task Composition, H/A = Human/Agent Qualities.

DA of the Human/Agent Qualities indicated that video gaming experience was the most important predictor of hit rate. Military experience was identified as a potential suppressor variable that gained importance through collinearity with other predictors in the model. The predictor of lowest importance to hit rate was age. Null Hypothesis 1, all Human/Agent Qualities are equally important to hit rate, was rejected.

DA of the NASA-TLX subscales, reflective of Task Perception, indicated that in this study the temporal demand subscale was the most important contributor to hit rate. Thus, the amount of experienced time pressure due to rate or pace of the task (Hart & Staveland, 1988) was an important predictor of hit rate. This result is unsurprising given the long duration of the three scenarios, each 32 minutes, all participants were exposed to. Null Hypothesis 2, all Task Perception variables are equally important to hit rate, was rejected.

DA on the full model using the most important predictors, and the human variables as constants, indicated that temporal demand (Task Perception) was the most important predictor of hit rate, followed by video gaming experience (Human/Agent Qualities) and agent report modality (Task Composition). Null Hypothesis 3, Task Composition, Perception of Task, and Human/Agent Qualities are equally important to task performance, was rejected.

The generalizability of the results of the full model was fairly robust. The complete dominance of temporal demand was replicated in 81.3 to 92.7% of the bootstraps, which indicated that the confidence that this result will be replicated in the natural world was high. However, the complete dominance of video gaming experience was less robust (43.2% - 71.5% reproducibility). Nonetheless, in the significance testing of the full hierarchical model, video gaming experience was significant. Moreover, the fit of the full model, although poor (pseudo R^2 = 0.189), was significant compared to the null model.

Study D

Study D manipulated the delivery frequency of agent reports (immediate vs. interval) and agent report modality (auditory vs. visual). The threat detection task occurred at a constant high event rate with a low signal probability. Average hit rate was considerably lower (*M* = 0.67, *SD* = 0.11) than Study A, B, and C and well below the military standard of performance (Naval Education and Training Command, 2009). Data regarding video gaming were not available and all participants were non-military (i.e., students). The hypotheses and results are summarized in **Table 46**.

.

Table 46

Summarized result of dominance analyses in Study D.

Null Hypothesis	Importance	Result
Hypothesis 1. All Human/Agent Qualities contribute equally to hit rate	Gender > Age	Null hypothesis 1 rejected
Hypothesis 2. All Task Perception variables contribute equally to hit rate	Mental demand. Temporal demand and effort suppressors	Null hypothesis 2 rejected
Hypothesis 3. All Task Composition variables contribute equally hit rate	Agent report modality > Agent report delivery frequency	Null hypothesis 3 rejected
Hypothesis 4. Task Composition, Perception of Task, and Human/Agent Qualities are equally important to hit rate	Gender (H/A) [a] > Mental demand (TP) [a] > Agent report modality (TC) [a]	Null hypothesis 4 rejected

Note. In the column Importance the > symbol signifies the dominance of the variable over the others. For instance, gender dominated mental demand and agent report modality, while mental demand also dominated agent report modality.

[a] TP = Task Perception, TC = Task Composition, H/A = Human/Agent Qualities.

DA of the Human/Agent Qualities variables, i.e., age and gender, indicated that gender was the most important predictor of hit rate. Null Hypothesis 1, all Human/Agent Qualities are equally important to hit rate, was rejected.

DA of the NASA-TLX subscales, reflective of Task Perception, indicated that in this study the mental demand subscale was the most important predictor of hit rate, followed by the effort subscale. The temporal demand subscale was identified as a potential suppressor variable. Null Hypothesis 2, all Task Perception variables are equally important to hit rate, was rejected.

DA of the Task Composition variables, agent report modality and report delivery frequency, indicated that agent report modality was the most important predictor of hit rate.

DA on the full Core model using the most important predictors, and age as a constant, indicated that gender (Human/Agent Qualities) was the most important predictor of hit rate, followed by mental demand (Task Perception) and agent report modality (Task Composition). Null Hypothesis 4, Task Composition, Perception of Task, and Human/Agent Qualities are equally important to task performance, was rejected.

The generalizability of the results of the full model was robust. The complete dominance results of gender were replicated in 73.2 to 99.1% of the (*S* = 1000) bootstraps, which indicated that the confidence that this result will be replicated in the natural world was high

The fit of the full model, based on the hierarchy of importance, was poor (pseudo R^2 = *0.243)* yet significant compared to the null model. The significance of the coefficients confirmed that gender, mental demand, and agent report modality were important predictors of hit rate.

Summary Results

The pattern of dominance was different between studies, potentially due to the different independent variables that were manipulated within each study. However, the studies also differed in the content of the agent reports, threat criterion, and design of the humanoid character models (see Appendix A through D). These factors could not be accounted for in the present research effort, as they were either fully nested between the studies or unidentified (in case of agent report content for Study D). Other differences between the studies were in terms of event rate and threat probability, two factors that influence task difficulty (Wickens & Hollands, 2000). These latter two predictors were collapsed into a new variable: threat conspicuity (**Table 39**).

Threat conspicuity refers to the ease of perceiving a threat under conditions of event rate and signal probability. An exploratory DA was conducted on the combined studies, keeping this task composition factor constant.

Combined Studies

DA was conducted on the full model, with task difficulty parameters (threat conspicuity and task duration) and age held constant. The hypothesis that Task Composition, Task Perception, and Human/Agent Qualities were equally important to hit rate was rejected. The analysis indicated that agent report modality (Task Composition) was the most important contributor to hit rate, followed by the NASA-TLX performance subscale (Task Perception) and gender (Human/Agent Qualities). The results were very robust. Complete dominance of agent report modality was dominated in 100% of the (S = 1000) bootstrap samples. General dominance of performance over gender was replicated in 100% of the bootstrap samples as well.

The fit of the full model was considerably better than the fit of the models of the individual studies (pseudo $R^2 = 0.520$ compared to pseudo $R^2 \leq 0.243$). Moreover, the beta regression model was significant compared to the null model. The significance of the coefficients revealed that not only agent report modality was indeed a significant predictor of hit rate, so were threat conspicuity and task duration.

Overarching Implications

Model

The validation results of the Core model, based on four studies, unveils a number of implications. First, the analyses of each of the studies showed that the factors within and between each section of the Core model, i.e., Task Composition, Task Perception, and Human/Agent Qualities, were not equal contributors to task performance (**Figure 43**). All studies falsified the model in this sense. However, since the results were drastically different between the studies, which factors are most important predictors of hit rate remained unknown based on the available experimental data. This is where the second implication needs to be discussed.

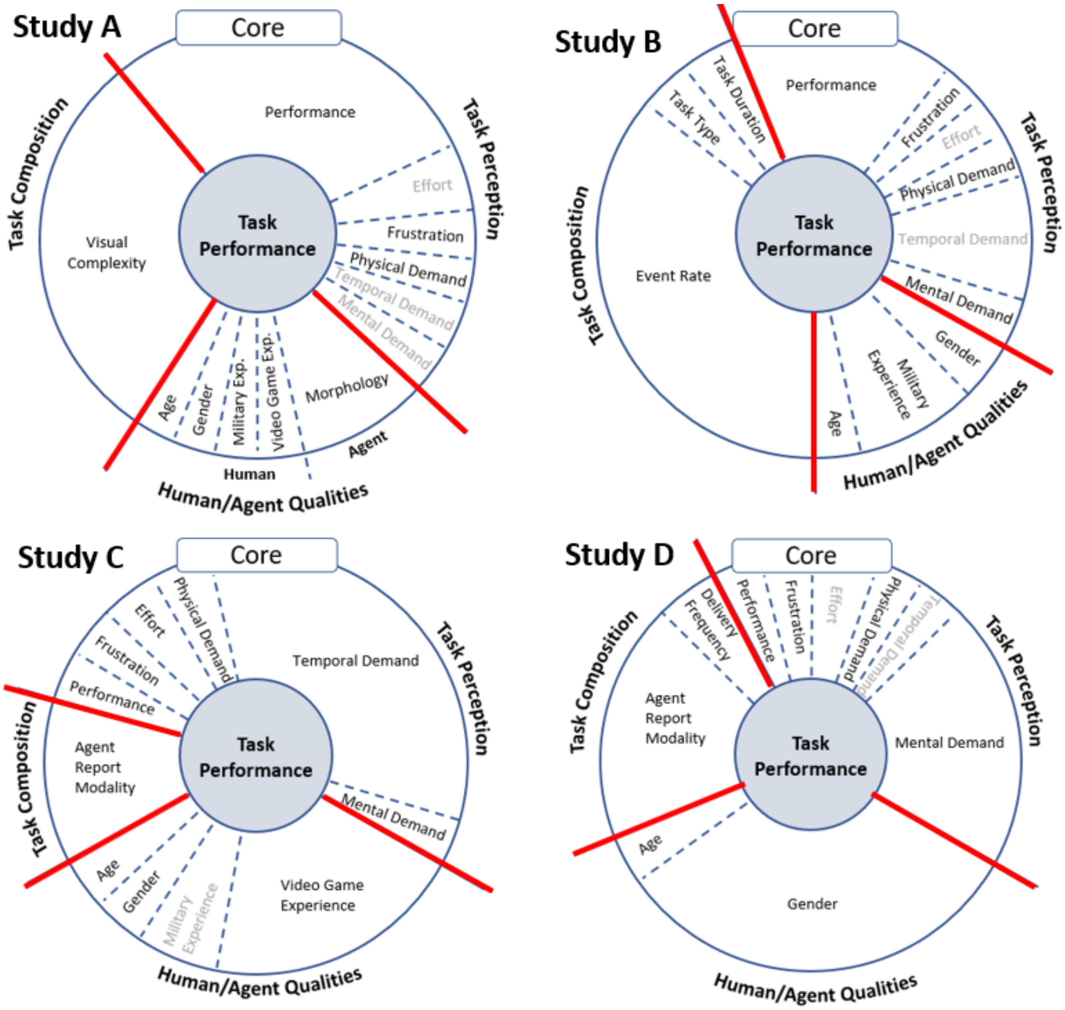


Figure 43. Summary of validation results per study.

Note. The validation results of the Core model are visually summarized in this figure. The size of the sections of the pie charts represent the relative size of importance of components. In Study A, Task Perception factors are most important to task performance, followed by Task Composition, and last Human/Agent Qualities. In contrast, Task Composition factors were most important in Study B, followed by Task Perception, and Human/Agent Qualities last. Study C identified Task Perception components as most important contributors to hit rate, followed by Human/Agent Qualities, and Task Composition factors last. Task Composition factors were also of lesser importance in Study D, where Human/Agent Qualities were most important, followed by Task Perception.

The second implication of the analyses is that task difficulty factors should be taken into account when analyzing the relative importance of factors to task performance. Task difficulty factors are task-specific elements that are manipulated to vary the difficulty of the task (Wickens & Hollands, 2000). In the experimental data here, the studies differed in event rate (number of characters on screen per minute; Wickens & Hollands, 2000), signal probability (the likelihood that one of these characters was a threat; Warm & Jerison, 1980), and task duration. The results in **Figure 43** reflect dominance patterns when these task difficulty factors are not taken into account and suggests that importance varies considerably between studies.

However, when the studies were combined, to allow for consideration of task difficulty factors, i.e., kept constant in the dominance analysis to take their explained variance into account (Azen & Budescu, 2003), the results showed that Task Composition factors matter most (**Figure 44**). Moreover, significance testing of the beta coefficients in the full model, wherein constants are evaluated, revealed that these task difficulty factors were also significant in predicting hit rate. The way in which participants rate their perceived workload related to the task (Task Perception) and the qualities team members bring to the table (Human/Agent Qualities) did not bare importance. Thus, the Core model was consistently falsified, indicating that Task Composition, Task Perception, and Human/Agent Qualities are not equally important to military HAT accuracy performance.

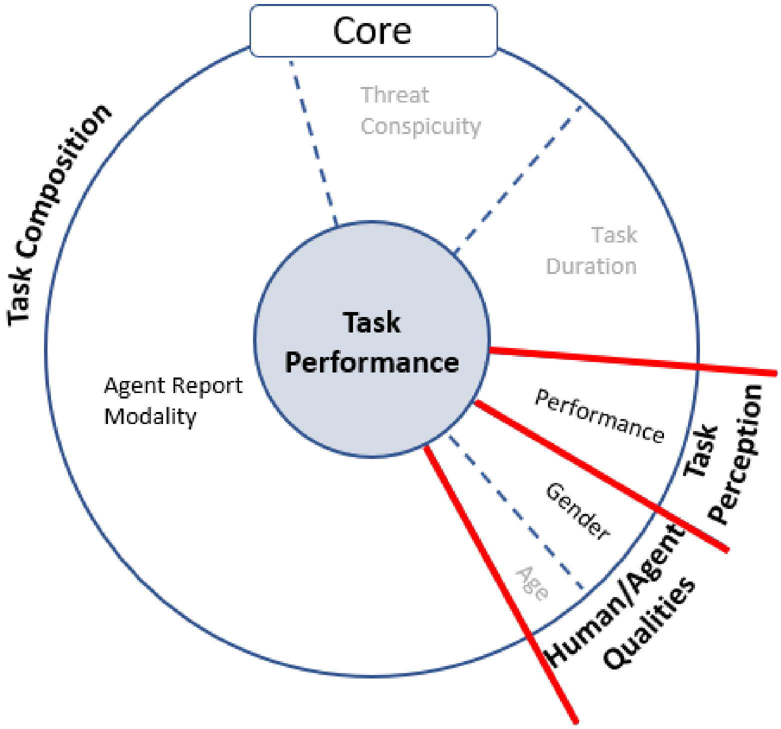


Figure 44. Summarized dominance analysis results in combined studies with task difficulty parameters kept constant.

Note. This figure summarizes the validation results of the Core model for the four studies combined, with task difficulty parameters (task duration and threat conspicuity, see **Table 39**) and age held constant. The size of the sections of the pie charts represent the relative size of importance of components. When task difficulty factors are taken into account, Task Composition factors are most important to task performance.

Conclusion Model

Based on the results, the assumption of the Core model, Task Components, Task Perception, and Human/Agent Qualities are equally important to hit rate, was falsified. The results between the studies were too different to reliably establish the most important contributor to hit rate, which may in part have been due to differences in task difficulty between the studies. However, when task difficulty parameters are taken into account, Task Composition factors were identified as most important to performance. These analyses also unveiled the need to take task difficulty parameters into account when examining dominance patterns.

Modeling Approach

A modeling approach was developed to validate importance-based models with a proportional outcome variable. The validation method consisted of the following steps:

1. Apply DA on beta regression models to determine the most important contributors to the outcome variable (Azen & Budescu, 2003; Budescu, 1993).

2. Establish the robustness and generalizability of the dominance results by bootstrapping the dominance values (Azen & Budescu, 2003; Efron, 1981).

3. Combine the most important predictors into a hierarchical beta regression model and evaluate the fit of the model (Ferrari & Cribari-Neto, 2004).

To conduct DA on beta regression models, four different pseudo R^2 statistics were tested. The most appropriate pseudo R^2 (Cox & Snell, 2018) was integrated with beta regression models in the dominanceanalysis package in R (Bustos & Countinho, 2019). This combined method has shown to be capable of establishing complete, conditional, and general dominance of predictors in beta-distributed data. This allows researchers to understand which predictors are most important to performance. The pseudo R^2 was also useful as a goodness-of-fit estimator, confirming previous studies (Shou & Smithson, 2015).

Conducting bootstrap procedures on the dominance values allowed a more robust evaluation of the dominance values. Dominance analysis alone is a qualitative relative weight analysis, which traditionally has not yielded statistical significance or confidence estimations (Budescu, 1993). Applying Azen and Budescu's (2003) bootstrap procedure in this effort yielded a confidence percentage indicative of generalization to the actual population. Moreover, bootstrap procedures confirmed the hierarchy of dominance as set forth by Budescu (1993). He suggested that complete dominance is a higher level of dominance than conditional, and lastly general dominance. If a predictor is completely dominant over another predictor, it is by definition also conditionally and generally dominant over said predictor (Budescu, 1993). Indeed, the bootstrap analyses indicated that generalizability of complete dominance is more difficult to establish, i.e., the confidence percentage of generalizability tended to be lower, than conditional and general dominance.

The last step of the validation approach is to combine the most important predictors, as identified by DA and bootstrapping, into a full hierarchical beta regression model (Ferrari & Cribari-Neto, 2004). This step added statistical significance testing to the traditional dominance analysis. As such, the significance of the dominant predictors was established and fortified the most dominant contributors of task performance. Evaluation of the full model, in terms of the fit, also yields a comparative goodness-of-fit approach, along with significance testing of the model against the null model (Ferrari & Cribari-Neto, 2004).

Conclusion Validation Approach

The developed model validation approach identified the most important contributors to hit rate per study, relative to all other predictors present in the model. The added bootstrap and model fit evaluation procedures allowed for significance testing of the dominance findings, a step that was previously lacking in DA. This approach has filled a gap in science; now importance-based models, with proportion-based outcome variables, can be validated with an R package that fluidly integrates beta regression into DA: https://rdrr.io/cran/dominanceanalysis/ (Bustos & Countinho, 2019).

Limitations

The first goal of the present effort was to develop a model of simulated military HAT to fill the gap in science. The proposed model is limited in a number of ways. First, the model is a step toward a conceptual model, rather than a true conceptual model that elucidates the interrelations between all concepts (Imenda, 2014). The proximity of the layers (Core, Relationship layer, and Environmental layer) to the center of task performance represents the hypothesized direct impact of these grouped variables to performance. However, the proposed model lacks directionality between and within the layers. The present research effort was a first step in modeling simulated military HAT. Future research should capitalize and continue this work to provide the finalized conceptual model.

In the present effort, model testing was limited to the Core model, as a first step in validation of this model. However, the Core could only be tested against available empirical data. This meant that components of Task Perception, i.e., perceived stress, and Human/Agent Qualities, i.e., personality differences, were lacking. Moreover, in the studies, the agent was simulated to be fully autonomous and 100% reliable; therefore, these Agent Qualities were not tested within the Core model. Furthermore, task performance, the focal point of this simulated military HAT model, was operationalized in terms of human performance in terms of threat detection performance. Here, this was an appropriate metric of HAT performance, as in dismounted military operations the Soldier is still recommended to make threat/no threat (i.e., life or death) decisions, rather than the agent (Singer, 2009). Moreover, the performance was conducted within the proposed HAT paradigm wherein the human and agent both contributed to the mission. The agent scouted the outer cordon and reported its findings to the human teammate. However, the extent to which the results from the present effort generalize to studies wherein task performance is operationalized in terms of agent and/or mission performance (e.g., time of completion) is unknown. Moreover, this outcome variable did not enhance our understanding of the global performance, which included both accuracy and response time. Typically, accurate responses in terms of decision-making come at a cost of prolonged response time and fast responses come at a cost of accuracy (Pachella & Pew, 1968).

Another limitation of the present effort relates to the design of the studies from which the experimental data was used. Three of the four studies showed a ceiling effect on the performance outcome (within military standards), while one study had an average outcome well below the military performance standard. These differences in results may have affected the dominance analysis findings in a non-controllable way. This is a limitation of the developed validation approach: it cannot transcend data limitations. An attempt was made to account for task difficulty factors between the studies and reevaluate the dominance pattern of predictors. However, in these attempts, most of the Task Composition factors were excluded from the analysis. Values could not be imputed as the results between the studies were different and the variables were manipulated factors. Given these limitations, no conclusions could be made regarding the most important predictors of hit rate.

Lastly, the finding that Task Composition was most important for hit rate may not be extended to other simulated military HAT studies. Even though the generalizability of the results were very robust, they may only pertain to studies with similar data. Moreover, the results may not yet extend to military HAT in the natural world either. Simulation is an ecologically valid approach to understanding phenomena in the natural world. However, the psychological conditions are very different between simulated military studies and the military battlefield. Similarly, collaborating with a simulated intelligent agent may not be a correct approximation of working with an agent face-to-face in the field either. Therefore, the implications of the results of the present research effort is limited to the described scope of simulated military HAT.

Future Research

In the current effort, a model of military HAT was developed that integrates important HAT-constructs and the Core model was validated against available empirical data. While the results falsified the assumption of the Core model that Task Composition, Task Perception, and

Human/Agent Qualities are equally important to performance, it is still unknown which factors contribute most to task performance. The experimental data used to test the model (a) prevented inclusion of all proposed components of the Core model, (b) differed in task difficulty parameters and outcome variables, and (c) were nested in terms of task difficulty. Future research should focus on testing the Core model with all of the proposed variables included.

Moreover, task difficulty parameters need further examination in simulated military HAT studies. When event rate and signal probability follow vigilance research (See et al., 1995), hit rate is high (> 0.90) and within the bounds of the military standard of performance (Naval Education and Training Command, 2009). However, when these task difficulty variables are changed such that event rate is constant and high (60 characters/minute) with a lower signal probability (0.09-0.10), average hit rate plummets and falls well below the military standard. Future research should focus on deepening the understanding the factors that affect performance under distinct task difficulty levels in simulated military HAT.

Additionally, another area of interest that future research should pursue is the beforementioned speed-accuracy tradeoff in these dynamic threat detection tasks. In a threat detection task, participants decide whether or not they think a character is a threat by clicking or not clicking on a character (Pachella & Pew, 1968). The speed-accuracy tradeoff can be examined using a response signal procedure that requires a response immediately after a signal appears, or using a deadline procedure, wherein a response should be given within a certain time limit (Dambacher & Hübner, 2013). In the current borrowed studies, the deadline procedure is in better alignment with the methodology than the response signal procedure, as participants were to click on a simulated threat before it would walk off the screen. However, a future study also should meet other design requirements to be able to calculate the speed-accuracy tradeoff, such as controlled/designed time-on-screen (deadline) for the characters and instructions to detect threats as accurately and as rapidly as possible under various task difficulty levels (Dambacher & Hübner, 2013; Wickelgren, 1977). The borrowed studies used here were not designed in this manner and the data thus cannot be evaluated in terms of a speed-accuracy tradeoff. Moreover, traditionally, speed-accuracy tradeoffs with deadline procedures are not used in dual-task paradigms (Dambacher & Hübner, 2013). Therefore, future studies looking to examine the speed-accuracy tradeoff in dynamic threat detection tasks may need to remove the interrupting agent reporting tasks in order to adequately understand the tradeoff.

Lastly, as mentioned in the limitations, the proposed model is a first step into providing a complete conceptual model that elucidates interrelations between concepts, both within and between the layers of the model. The relationships between the concepts within the Core model should be further clarified, following the suggested guidelines in this section. Then, the interrelations between the constructs within the Relationship Layer and the Environmental Layer need to be further developed, validated, and mapped within the conceptual model. As a last step, the transactional interactions between the three sections of the model (Core, Relationship Layer, Environmental Layer) should be tested.

Funding

Parts of the research reported in this document were performed in connection with Contract Number W911NF-10-2-0016 with the U.S. Army Research Laboratory. The views and conclusions contained in this document are those of the author and should not be interpreted as presenting the official policies or position, either expressed or implied, of the U.S. Army Research Laboratory, or the U.S. Government unless so designated by other authorized documents. Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

APPENDIX A: DESCRIPTION BORROWED STUDY A

APPENDIX A: DESCRIPTION BORROWED STUDY A

Participants were recruited from the University of Central Florida's undergraduate psychology pool in exchange for course credit. Two participants classified as military based on their extensive self-reported military experience.

The study was ran on a Human-Robot Interaction testbed that was built in an Unreal Games Engine environment (Epic Games, Inc., 2019). A virtual reality system was used to create an immersed, simulated experience. The HTC Vive (D'Orazio, 2015) system was used for this study. This system consists of two SteamVR base stations, a head-mounted display, with a camera near the bottom, and two wireless handheld controllers, allowing participants to interactively move in 3D space. The base stations create a 360 degree virtual space up and emit infrared pulses at 60 pulses per second, allowing the Vive system to track the participant's physical location (D'Orazio, 2015; Steamworks, 2019). The headset and controllers both have infrared sensors that interact with the base stations, allowing the system to track the accessories in 3D space ("HTC Vive," 2019). The headset refreshes at 90 Hz and has a 110 degree field of view (FOV), although an entire 360 degree FOV is available due to the physical affordances of the system (D'Orazio, 2015; VIVE, 2019). In the display, two OLED panels are available, one for each eye, with a combined display resolution of 2160 x 1200 pixels ("HTC Vive," 2019).

With the controller, participants could point to characters on screen. Pointing the controller created a simulated laser beam, which allowed participants to aim. With the controller's trigger, they could identify a threat. A trigger-click temporarily highlighted the character, as a feedback of response mechanism.

Experimental Design

This study was a mixed design, with visual complexity (of the signal detection display and icons) as a between-subjects variable (two levels: low vs. high) and agent type as a within-subjects variable (two levels: legged (study A.1) vs. wheeled (study A.2)). Neither of these factors was represented sufficiently in the other included studies to generate statistical power for the present effort. Thus, these independent variables were not coded in the dataset for the present study. The order of presentation was coded, which agent was presented first was counterbalanced and randomized in Study A. Each task duration was approximately 10 minutes.

Participant data was collected in accordance with the approved IRB. Video gaming experience was rated as shown in **Table 47.**

Table 47

Rating scale for video gaming frequency.

1	2	3	4	5	6
Never	Rarely	Once every few months	Monthly	Weekly	Daily

Threat Detection Task

The ongoing task was a threat detection task, wherein participants were to identify threats among the characters walking across the screen. The event rate for the characters was set at 15 per minute. The characters were of three types: friendly soldiers, friendly civilians, and enemy civilians (insurgents). **Figure 45** shows the friendly civilians and soldiers. **Figure 46** displays the range of enemy civilians. Participants identified a threat by clicking on them with the HTC Vive controller.

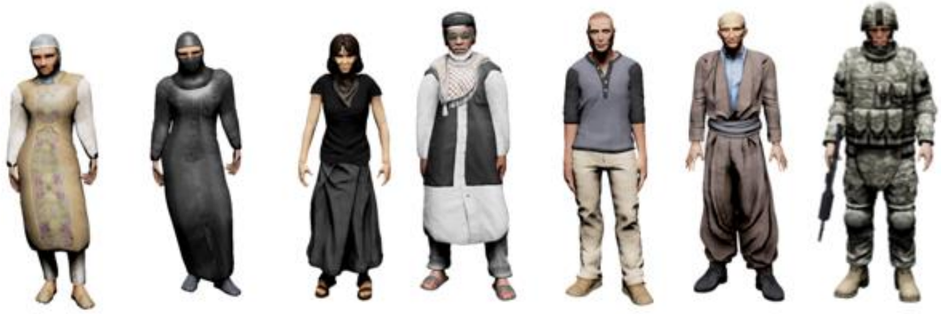


Figure 45. Friendly civilians and soldiers (non-threats) in Study A.

Note. Copyright permission in APPENDIX K.

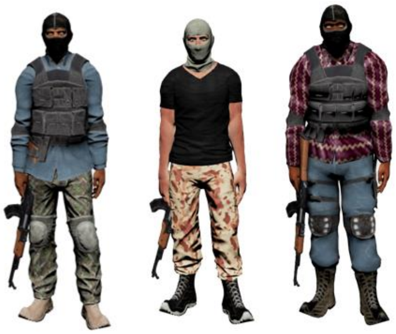


Figure 46. Enemy civilians (threats) in Study A.

Note. Copyright permission in APPENDIX K.

Agent Reporting Task

As part of the cordon-and-search mission, the autonomous robot teammate would conduct their search of a designated cordon and report back. These reports were presented visually in the interface as a text report with additional visual informational elements, which were manipulated between the low and high visual complexity conditions. The reports contained information regarding what was found (e.g., three IEDs, two insurgents, five weapon crates) and where it was found (e.g., East side of the building, North side of the building, first floor of the building). These reports were created based on Subject Matter Expert interviews with a former U.S. Army Staff Sergeant and the U.S. Army Handbook (Headquarters Department of the Army, 2006). A report was sent regularly, that is, every 15 to 18 seconds and the text updates lasted 10 seconds each.

APPENDIX B: DESCRIPTION BORROWED STUDY B

APPENDIX B: DESCRIPTION BORROWED STUDY B

Two samples were utilized in study B. One sample used undergraduate students from the University of Central Florida that were recruited through the Psychology resource pool for course credit. Three participants from this sample had multiple years of military experience (two in the National Guard and one Air Force Reservist). As such, these participants were classified as military rather than student. The other sample were Soldiers from Ft. Benning's officer school. Soldiers volunteered and did not receive compensation for their participation.

Equipment

This study was ran on a desktop-based version of a custom simulation that was developed in the Unreal 4 Game Engine (Epic Games, Inc., 2019). The task was viewed on a 30" monitor with a resolution of 2560 x 1600. In the top center of the screen, a simulated multimodal interface (MMI) would become available when visual agent reports were sent to the participant. The MMI matches the size of a Toughpad FZ-M1 tablet with a resolution of 602 x 377 pixels.

Experimental Design

Study B consists of two conditions. Participants actively pulled agent reports, but under constant (B.1) or changing (B.2) event rate. The order of presentation was counterbalanced.

In conditions B.1 and B.2 event rate was manipulated as a within-subjects variable. In B.1 the ongoing threat detection task had a constant number of characters on screen per minute, which was set at 30 characters/minute. In B.2, the event rate changed halfway during the scenario. Half of the scenario ran in a low event rate, with 15 characters/minute, while the remainder ran in a high event rate, with 60 characters/minute. The order of the event rate shift, either from low-to-high or high-to-low, was counterbalanced within the design.

Threat Detection Task

In each condition, the ongoing task was a simulated military Cordon-and-Search operation (Sutherland et al., 2010), in which participants were tasked with capturing photos of threats to help build the agent teammate’s database with examples of threats. With each click, they heard a camera snapshot sound as feedback of response. There were four types of characters: friendly Soldiers and friendly civilians (**Figure 47**), and enemy Soldiers and armed civilians (**Figure 48**). Both enemy Soldiers and insurgents were threats and required a picture being taken by the participant.

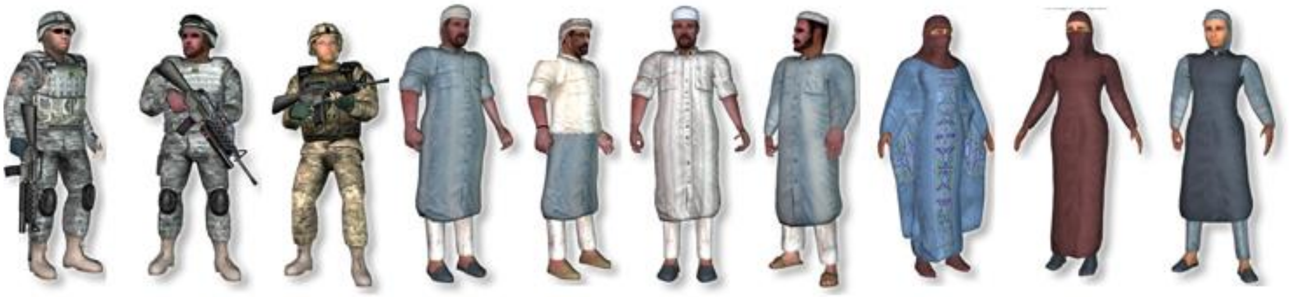


Figure 47. Friendly Soldiers and civilians (non-threats) in Study B.

Note. Copyright permission in APPENDIX K.

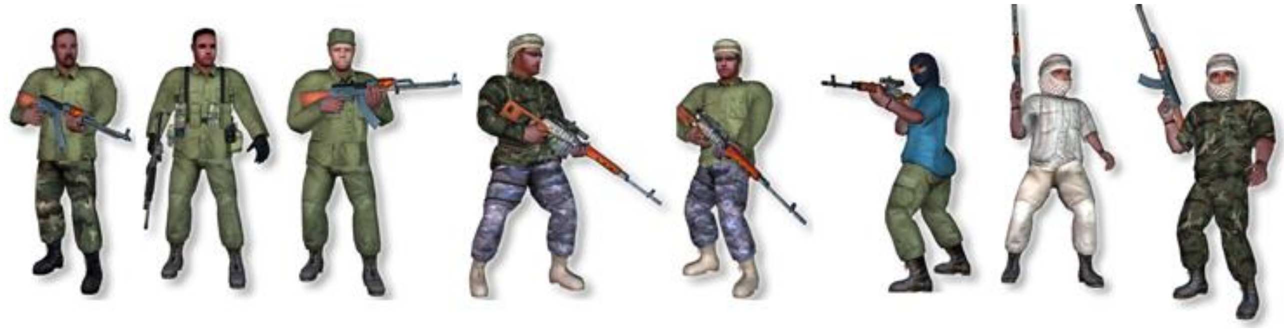


Figure 48. Enemy soldiers and armed civilians (threats) in Study B.

Note. Copyright permission in APPENDIX K.

Agent Reporting Task

Participants could request a report from the agent teammate regarding the number of threats (critical, non-critical, and non-targets) if they wanted to. The multimodal interface could be brought up and a report was requested by clicking on text or image. These reports provided participants situation awareness to respond to commander queries (SA probes).

The information displayed in either report was identical. In the image report, boxes were shown around threats and critical threats, while the text report showed the number of threats, critical threats, and non-threats (not needed for probes). Participants also had the freedom to pull text and image reports sequentially.

APPENDIX C: DESCRIPTION BORROWED STUDY C

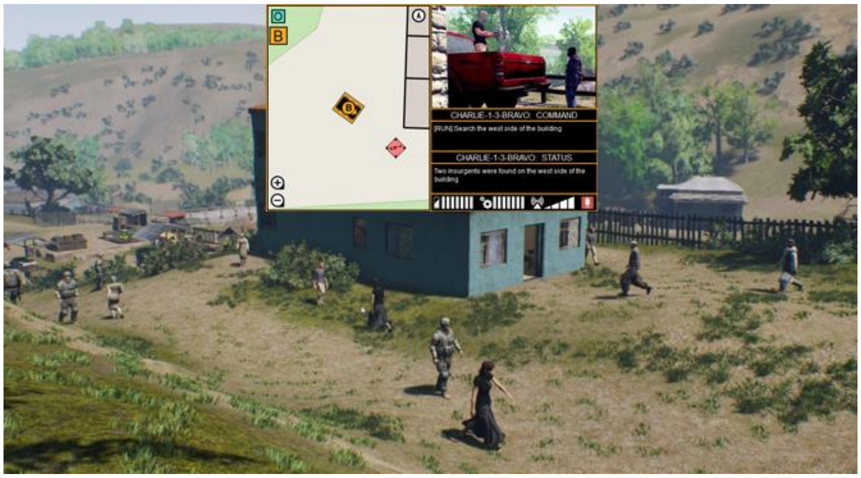
APPENDIX C: DESCRIPTION BORROWED STUDY C

Participants were recruited from the University of Central Florida's undergraduate psychology pool in exchange for course credit.

Equipment

This study was ran on a desktop-based version of a custom simulation that was developed in the Unreal 4 Game Engine (Epic Games, Inc., 2019). The task was viewed on a 30" monitor with a resolution of 2560 x 1600. The simulated environment was a typical Middle Eastern urban environment (**Figure 49**), in which characters walked across the screen. In the top center of the screen, an MMI would become available when visual agent reports were sent to the participant. The MMI matches the size of a Toughpad FZ-M1 tablet used in Barber et al. (2015), with a resolution of 602 x 377 pixels. Auditory reports were delivered through text-to-speech generated with Microsoft's speech platform Software Development Kit version 11 (Microsoft, 2019), based on Window's 10 default male voice.

The MMI has three sections (Error! Reference source not found.). The left section provides an aerial map of the environment with the location of the reporting robot and a military symbol of what was found. The right section of the MMI consists of an image of what was found (top right) and of a visual text of the complete report (bottom right). The auditory report mimicked the visual text.



CHARLIE-1-3-BRAVO: COMMAND

[RUN] Search the west side of the building

CHARLIE-1-3-BRAVO: STATUS

Two insurgents were found on the west side of the building

Figure 49. Simulated environment in Study C.

Note. The simulated environment in Study C shows threats, non-threats, and the multimodal interface used for agent-to-human communications. Copyright permission is found in APPENDIX K: COPYRIGHT.

Experimental Design

Two within-subjects factors were manipulated in this study, in which participants conducted an ongoing threat detection task and a concurrent agent reporting task that simulated a military cordon-and-search operation (Sutherland et al., 2010). These factors were manipulated over three scenarios, that each lasted approximately 32 minutes (**Figure 10**).

In all three conditions, event rate, operationalized as the number of characters on screen per minute, was varied. It changed every eight minutes from low (15 characters/minute) to high (60 characters/minute) and high to low. An exception in this design, are the first and last blocks; these only lasted four minutes.

Condition C.1 and C.2 varied the modality in which agent reports were delivered every eight minutes, with the exception of the first and last block. The only difference between the two conditions is the modality in the starting block. In Condition C.3 the reports were sent in two modalities simultaneously.

Threat Detection Task

Participants performed the role of a squad leader in an outer cordon area. During the task, three types of characters walked around a building and surrounding area. Non-threats were friendly soldiers, dressed in full camouflage and armor with a weapon, and friendly civilians, characterized by civilian clothing and absence of a weapon (**Figure 50**). Threats were enemy civilians recognizable by casual clothing or clothing mixed with camouflage, a weapon, and a mask (**Figure 51**). Participants identified threats by clicking on them with a mouse. This action highlighted the character briefly as feedback of response.

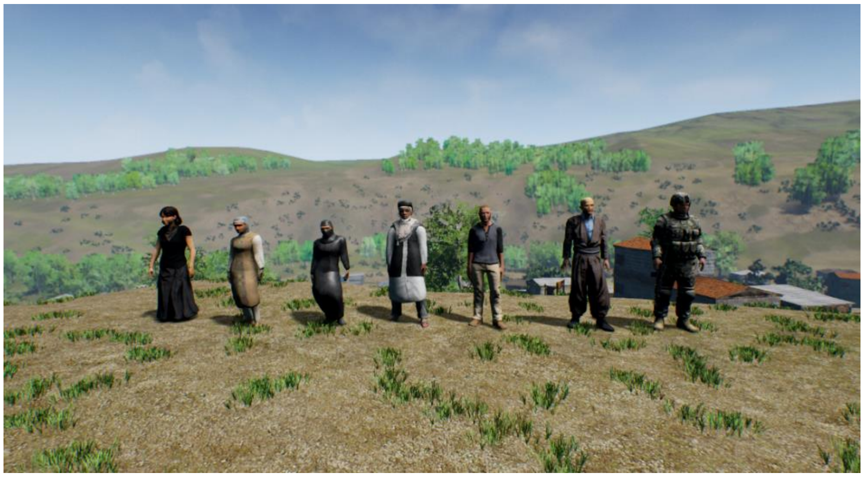


Figure 50. Friendly soldiers and friendly civilians, all non-threats, in Study C.

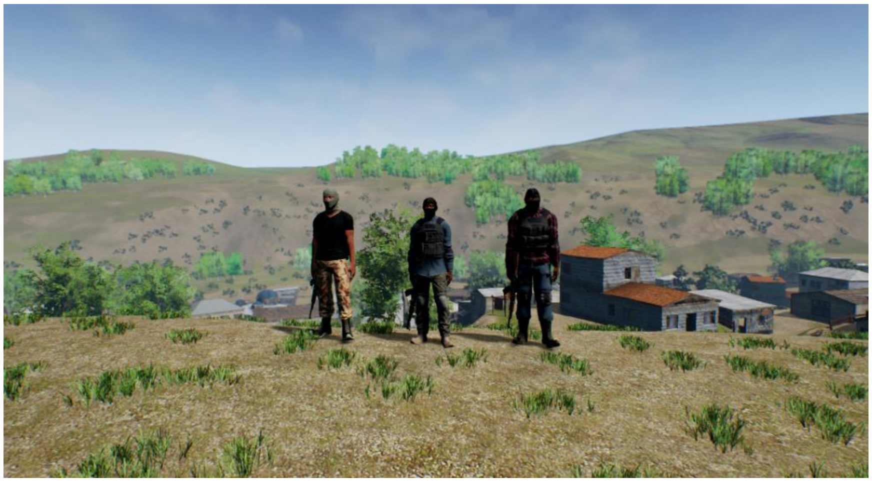


Figure 51. Enemy civilians (threats) in Study C.

Agent Reporting Task

As part of the cordon and search mission, two out-of-sight agents scouted the inner cordon and reported their findings back to the squad leader. These reports included information regarding identification (money bags, IEDs, weapon crates, or insurgents) and location (inside the building on the first or second floor, and outside the building based on four cardinal directions). These reports were created based on Subject Matter Expert interviews with a former U.S. Army Staff Sergeant and the U.S. Army Handbook (Headquarters Department of the Army, 2006).

The agent teammates sent these reports auditorily and/or visually, depending on the experimental condition. A report was sent regularly, that is, every 15 to 18 seconds. The information conveyed in each condition was identical. Visual reports, either in a single-modality condition or dual-modality condition, automatically prompted the appearance of the MMI. The visual display was generated by the system rather than having the participant initiate display of the visual report, to ensure equal time was spent in both the auditory and visual modality. Over a four-minute block, nine agent reports were delivered. Approximately every 18 seconds a report was delivered. Thus, within a four-minute block, nine reports were delivered, resulting in 72 reports in the eight four-minute blocks.

APPENDIX D: DESCRIPTION BORROWED STUDY D

APPENDIX D: DESCRIPTION BORROWED STUDY D

Participants were recruited via the Psychology undergraduate student resource pool at the University of Central Florida. No military experience was reported by any of the participants.

Equipment

The simulation ran in a custom-built platform (FIRE; Vasquez, Bendell, Talone, & Jentsch, 2018) in the Unreal 4 Game Engine (Epic Games, Inc., 2019). The HTC Vive virtual reality system was used to create an immersive and interactive 3D experience (VIVE, 2019). A MMI was rendered inside the FIRE, modeled after a military-implemented Toughpad (Barber et al., 2015). Participants could pull the MMI up with the HTC VIVE controller. The MMI displayed an image of what the agent is looking at, command text, as well as sections that relay the current status of the agent teammate including battery levels, mechanical health, and Wi-Fi connectivity. The controller was used to open and close the MMI, to increase the size of transmitted images to full-screen, and to reply to input requests.

Experimental Design

This study employed a mixed design, wherein two two-level factors were manipulated. Each participant experienced two sensory modalities of agent report delivery (visual text vs. auditory speech) in two separate scenarios, each lasting approximately 16 minutes. The between-subjects variable was the timing of agent report delivery. Reports could be delivered regularly every minute (Condition D.1) or immediately, which was irregular (Condition D.2). This created four different orders for the scenarios, which were randomized and counterbalanced. For the purpose of the current effort, timing of report delivery was encoded into the variable Agent Report Event Rate.

Threat Detection Task

Participants performed a simulated military cordon-and-search operation (Sutherland et al., 2010), wherein they teamed with an agent teammate. The ongoing task was a threat detection task. As characters walked across the screen, in a Middle Eastern urban environment, participants were asked to identify threats by clicking on them with the HTC Vive controller. The controller emitted a laser-like beam in the environment, allowing participants to aim precisely. A click on any character would briefly highlight the character, generating feedback of response to the participant. Six characters were employed (**Figure 52**), each carrying an object (**Figure 53**). Threats were characters carrying a small handgun.

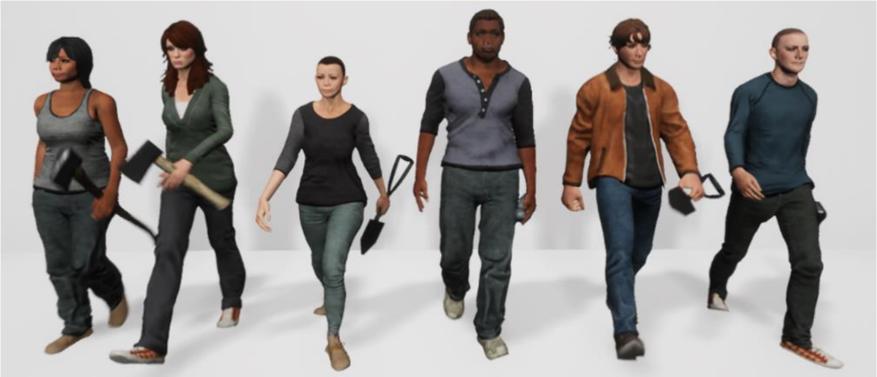


Figure 52. Character models employed in Study D.

Note. The characters that carry a handgun are threats. Copyright permission in APPENDIX K.

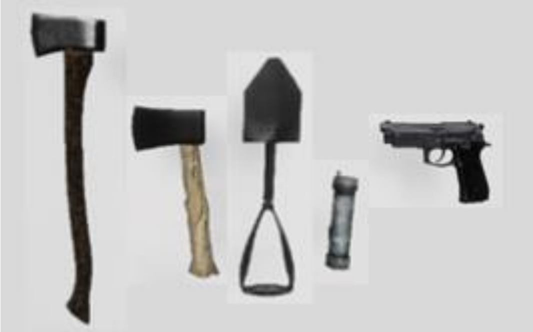


Figure 53. Threat identifier in Study D.

Note. The figure shows the objects that characters could carry, wherein the handgun was an identifier for threats. Copyright permission in APPENDIX K.

Agent Reporting Task

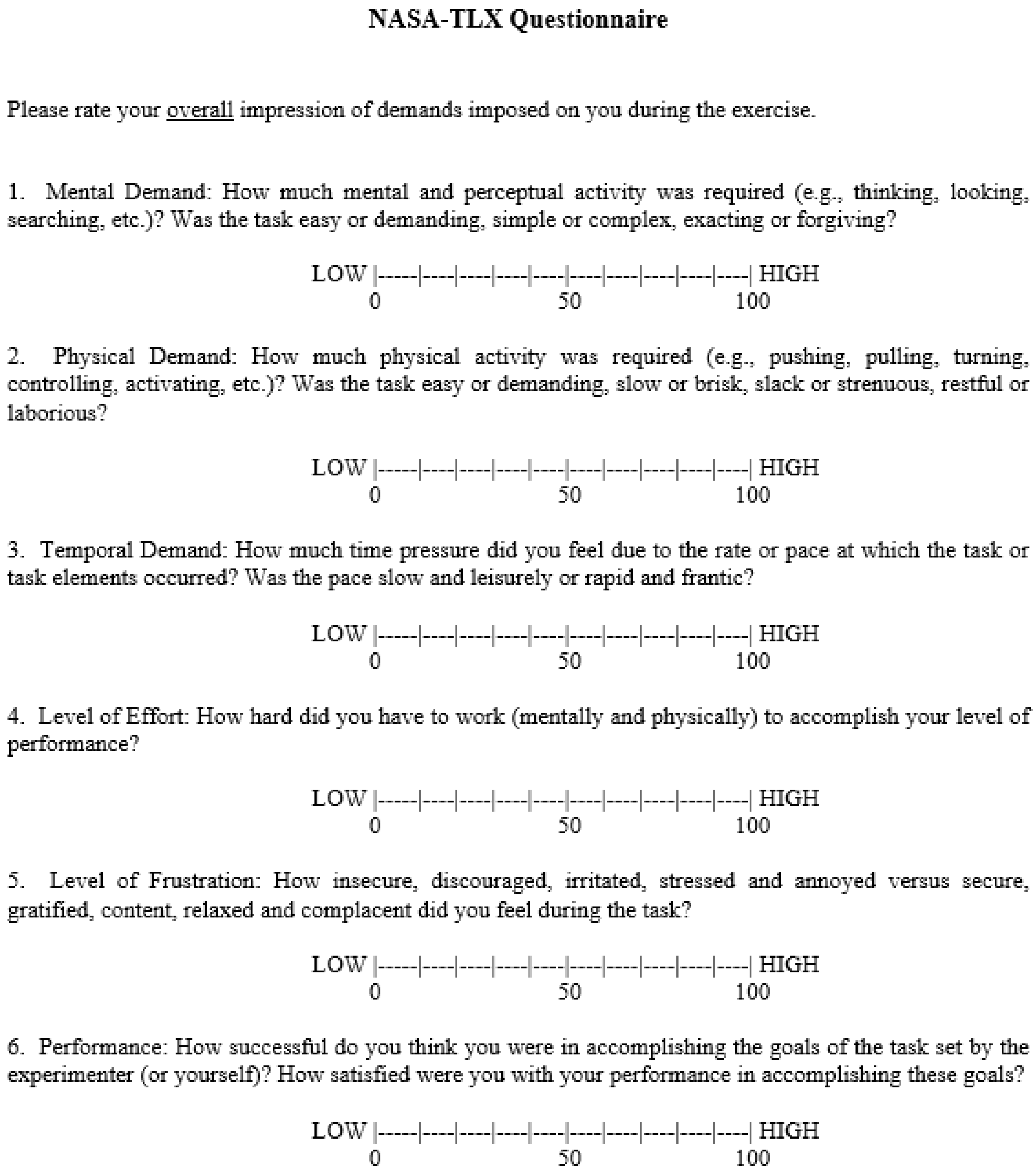
During the threat detection task, the agent teammate scouted the inner cordon simultaneously. The agent searched the environment for IEDs and took pictures, thereby producing reports that it sent to the participant.

During each scenario, a total of 34 reports were presented to participants; timing of delivery was manipulated as between-subjects variable. To ensure all reports were attended to, an auditory tone alerted participants one second prior to release of each report. There were 30 non-critical reports that contained information pertaining to the route, such as obstacles encountered. Four reports were critical and included an IED image review request. Report review was possible by clicking a button on the controller to pull up the MMI. They could raise the controller to bring the MMI up or keep the controller down to look down at the simulated MMI. Participants had 15 seconds to review the report. Once the image was reviewed, participants needed to determine if a hazard (IED) was present or the area was clear with another button click.

The modality through which reports were delivered was auditory or visual. All participants conducted each scenario. In the auditory condition, all non-critical reports were sent through speech alone. Critical IED review requests were still sent visually, as these required visual inspection. Contrary, in the visual report condition, all reports were solely transmitted through the MMI.

APPENDIX E: NASA-TLX

APPENDIX E NASA-TLX



NASA-TLX Questionnaire

Please rate your <u>overall</u> impression of demands imposed on you during the exercise.

1. Mental Demand: How much mental and perceptual activity was required (e.g., thinking, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?

LOW |-----|----|----|----|----|----|----|----|----|----| HIGH
0 50 100

2. Physical Demand: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?

LOW |-----|----|----|----|----|----|----|----|----|----| HIGH
0 50 100

3. Temporal Demand: How much time pressure did you feel due to the rate or pace at which the task or task elements occurred? Was the pace slow and leisurely or rapid and frantic?

LOW |-----|----|----|----|----|----|----|----|----|----| HIGH
0 50 100

4. Level of Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?

LOW |-----|----|----|----|----|----|----|----|----|----| HIGH
0 50 100

5. Level of Frustration: How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

LOW |-----|----|----|----|----|----|----|----|----|----| HIGH
0 50 100

6. Performance: How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

LOW |-----|----|----|----|----|----|----|----|----|----| HIGH
0 50 100

APPENDIX F: RESULTS

APPENDIX F RESULTS

Study A

Linearity Check

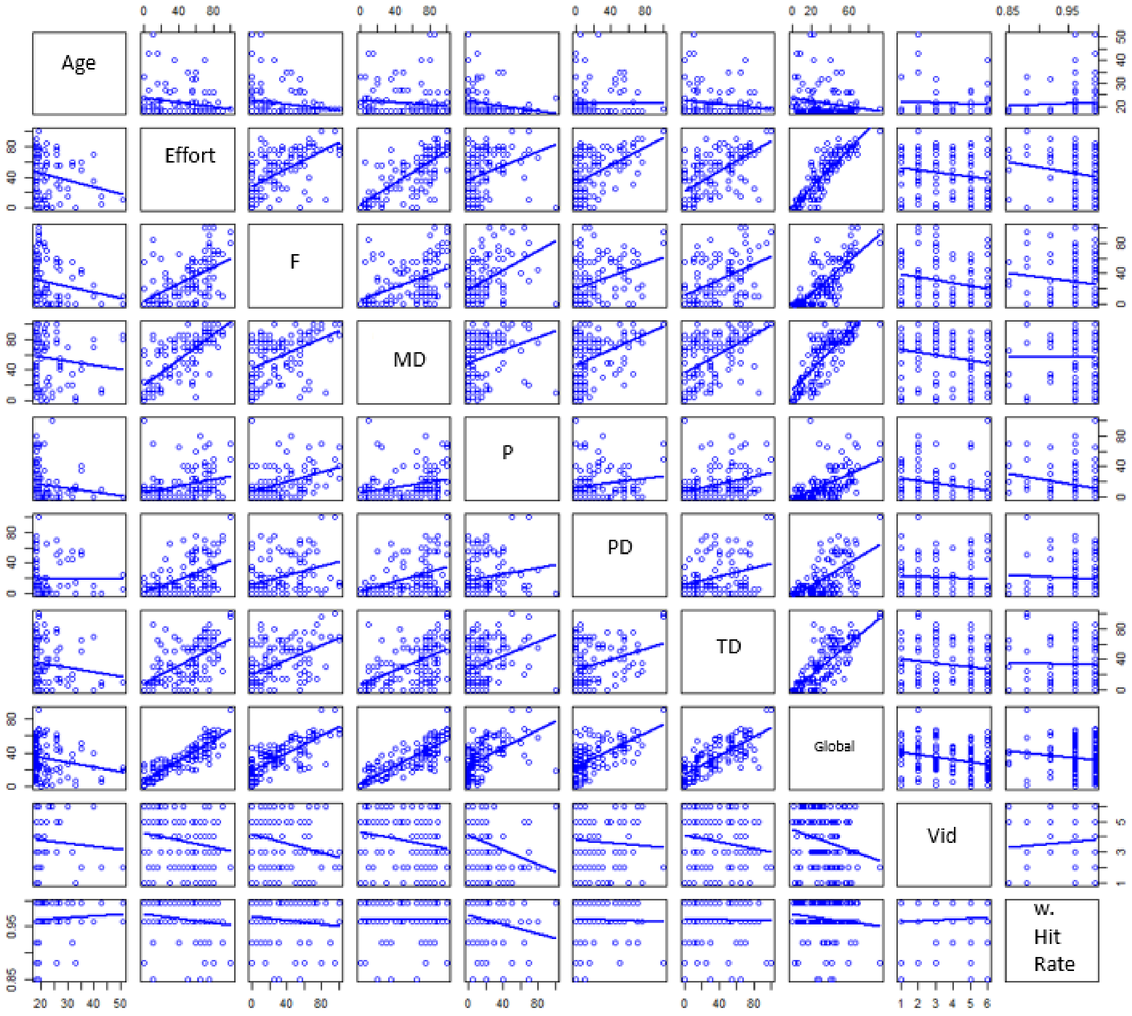


Figure 54. Scatterplot matrix continuous variables Study A.

Note. Spearman’s correlation was used. The abbreviations represent: F = Frustration subscale on NASA-TLX, MD = Mental Demand subscale on NASA-TLX, P = Performance subscale on NASA-TLX, PD = Physical Demand subscale on NASA-TLX, TD = Temporal Demand subscale on NASA-TLX, Global = average score on NASA-TLX, Vid = Video Gaming Experience, w. Hit Rate = winsorized hit rate.

Dominance Analysis

Human/Agent Qualities

Table 48.

Raw dominance analysis results Human/Agent Qualities in Study A.

	Additional contribution (pseudo R^2) of				
Subset model ***X***	Age	Gender	Video Gaming Experience	Military Experience	Agent Type
k = 0	0.004	0.000	0.000	0.001	0.004
Age		0.000	0.000	0.000	0.004
Gender	0.004		0.000	0.002	0.004
Video Gaming Experience	0.004	0.000		0.001	0.004
Military Experience	0.002	0.000	0.000		0.004
Agent Type	0.003	0.000	0.000	0.002	
Conditional dominance *k* = 1	0.003	0.000	0.000	0.001	0.004
Age + Gender			0.001	0.000	0.004
Age + Video Gaming Experience		0.000		0.000	0.004
Age + Military Experience		0.000	0.000		0.004
Age + Agent Type		0.000	0.000	0.000	
Gender + Video Gaming Experience	0.004			0.002	0.004
Gender + Military Experience	0.002		0.001		0.004
Gender + Agent Type	0.003		0.000	0.002	
Video Gaming Experience + Military Experience	0.003	0.000			0.004
Video Gaming Experience + Agent Type	0.004	0.000		0.001	
Military Experience + Agent Type	0.002	0.000	0.000		
Conditional dominance *k* = 2	0.003	0.000	0.000	0.001	0.004
Age + Gender + Video Gaming Experience				0.000	0.004
Age + Gender + Military Experience			0.001		0.004

	Additional contribution (pseudo R^2) of				
Subset model ***X***	Age	Gender	Video Gaming Experience	Military Experience	Agent Type
Age + Gender + Agent Type			0.001	0.000	
Age + Video Gaming Experience + Military Experience		0.000			0.004
Age + Video Gaming Experience + Agent Type		0.001		0.000	
Age + Military Experience + Agent Type		0.000	0.000		
Gender + Video Gaming Experience + Military Experience	0.003				0.004
Gender + Video Gaming Experience + Agent Type	0.004			0.002	
Gender + Military Experience + Agent Type	0.002		0.001		
Video Gaming Experience + Military Experience + Agent Type	0.002	0.001			
Conditional dominance *k* = 3	0.003	0.000	0.001	0.001	0.004
Age + Gender + Video Gaming Experience + Military Experience					0.004
Age + Gender + Video Gaming Experience + Agent Type				0.000	
Age + Gender + Military Experience + Agent Type			0.001		
Age + Video Gaming Experience + Military Experience + Agent Type		0.001			
Gender + Video Gaming Experience + Military Experience + Agent Type	0.002				
Conditional dominance *k* = 4	0.002	0.001	0.001	0.000	0.004
Age + Gender + Video Gaming Experience + Military Experience + Agent Type					
Overall average	0.003	0.000	0.000	0.001	0.004

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, wherein *k* = 0 indicates that no other variables are in the model aside of the predictor under evaluation. Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

Task Perception Variables

Table 49

Raw dominance analysis results Task Perception in Study A

	Additional contribution (pseudo R^2) of					
Subset model ***X***	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
k = 0	0.021	0.000	0.001	0.000	0.015	0.013
Performance		0.000	0.001	0.001	0.007	0.004
Mental Demand	0.021		0.001	0.000	0.028	0.015
Physical Demand	0.021	0.000		0.000	0.014	0.012
Temporal Demand	0.022	0.000	0.001		0.025	0.016
Effort	0.014	0.014	0.001	0.010		0.003
Frustration	0.012	0.002	0.000	0.003	0.004	
Conditional dominance *k* = 1	0.018	0.003	0.001	0.003	0.016	0.010
Performance + Mental Demand			0.002	0.000	0.024	0.007
Performance + Physical Demand		0.001		0.001	0.007	0.003
Performance + Temporal Demand		0.000	0.002		0.020	0.008
Performance + Effort		0.017	0.000	0.014		0.000
Performance + Frustration		0.003	0.000	0.004	0.004	
Mental Demand + Physical Demand	0.022			0.000	0.029	0.014
Mental Demand + Temporal Demand	0.022		0.001		0.036	0.016
Mental Demand + Effort	0.017		0.001	0.008		0.003
Mental Demand + Frustration	0.013		0.000	0.002	0.017	
Physical Demand + Temporal Demand	0.022	0.000			0.026	0.015
Physical Demand + Effort	0.013	0.014		0.011		0.003
Physical Demand + Frustration	0.012	0.002		0.003	0.005	
Temporal Demand + Effort	0.017	0.011	0.002			0.005
Temporal Demand + Frustration	0.013	0.000	0.000		0.014	
Effort + Frustration	0.011	0.014	0.001	0.012		
Conditional dominance *k* = 2	0.016	0.006	0.001	0.006	0.018	0.008
Performance + Mental Demand + Physical Demand				0.001	0.023	0.006
Performance + Mental Demand + Temporal Demand			0.002		0.034	0.008
Performance + Mental Demand + Effort			0.001	0.011		0.000
Performance + Mental Demand + Frustration			0.001	0.002	0.018	
Performance + Physical Demand + Temporal Demand		0.000			0.020	0.007
Performance + Physical Demand + Effort		0.017		0.014		0.000
Performance + Physical Demand + Frustration		0.003		0.005	0.004	
Performance + Temporal Demand + Effort		0.014	0.001			0.001
Performance + Temporal Demand + Frustration		0.001	0.000		0.014	

Performance + Effort + Frustration		0.017	0.000	0.015		

	Additional contribution (pseudo R^2) of					
Subset model ***X***	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
Mental Demand + Physical Demand + Temporal Demand	0.022				0.037	0.015
Mental Demand + Physical Demand + Effort	0.016			0.009		0.003
Mental Demand + Physical Demand + Frustration	0.013			0.002	0.018	
Mental Demand + Temporal Demand + Effort	0.020		0.002			0.005
Mental Demand + Temporal Demand + Frustration	0.014		0.000		0.025	
Mental Demand + Effort + Frustration	0.014		0.001	0.010		
Physical Demand + Temporal Demand + Effort	0.016	0.012				0.006
Physical Demand + Temporal Demand + Frustration	0.014	0.000			0.016	
Physical Demand + Effort + Frustration	0.010	0.014		0.014		
Temporal Demand + Effort + Frustration	0.013	0.011	0.003			
Conditional dominance k = 3	0.015	0.009	0.001	0.008	0.021	0.005
Performance + Mental Demand + Physical Demand + Temporal Demand					0.034	0.007
Performance + Mental Demand + Physical Demand + Effort				0.011		0.000
Performance + Mental Demand + Physical Demand + Frustration				0.002	0.018	
Performance + Mental Demand + Temporal Demand + Effort			0.001			0.001
Performance + Mental Demand + Temporal Demand + Frustration			0.001		0.027	
Performance + Mental Demand + Effort + Frustration			0.001	0.011		
Performance + Physical Demand + Temporal Demand + Effort		0.014				0.002

	Additional contribution (pseudo R^2) of					
Subset model ***X***	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
Performance + Physical Demand + Effort + Frustration		0.017		0.016		
Performance + Temporal Demand + Effort + Frustration		0.014	0.001			
Mental Demand + Physical Demand + Temporal Demand + Effort	0.019					0.006
Mental Demand + Physical Demand + Temporal Demand + Frustration	0.014				0.027	
Mental Demand + Physical Demand + Effort + Frustration	0.013			0.011		
Mental Demand + Temporal Demand + Effort + Frustration	0.016		0.003			
Physical Demand + Temporal Demand + Effort + Frustration	0.012	0.011				
Conditional dominance *k* = 4	0.015	0.011	0.001	0.010	0.024	0.003
Performance + Mental Demand + Physical Demand + Temporal Demand + Effort						0.001
Performance + Mental Demand + Physical Demand + Temporal Demand + Frustration					0.028	
Performance + Mental Demand + Physical Demand + Effort + Frustration				0.012		
Performance + Mental Demand + Temporal Demand + Effort + Frustration			0.001			
Performance + Physical Demand + Temporal Demand + Effort + Frustration		0.014				
Mental Demand + Physical Demand + Temporal Demand + Effort + Frustration	0.015					
Conditional dominance *k* = 5	0.015	0.014	0.001	0.012	0.028	0.001
Performance + Mental Demand + Physical Demand + Temporal Demand + Effort + Frustration						
Overall average	0.017	0.007	0.001	0.007	0.020	0.007

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, wherein *k* = 0 indicates that no other variables are in the model aside of the predictor under evaluation. Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

Full Model

Table 50

Raw dominance analysis results Full Model in Study A.

	Additional contribution (pseudo R^2) of			
Subset Model ***X***	Age	Task Type	Visual Complexity	Performance
Gender + Video Gaming Experience + Military Experience	0.003	0.004	0.017	0.020
Gender + Video Gaming Experience + Military Experience + Age		0.004	0.015	0.019
Gender + Video Gaming Experience + Military Experience + Task Type	0.002		0.016	0.018
Gender + Video Gaming Experience + Military Experience + Visual Complexity	0.000	0.004		0.011
Gender + Video Gaming Experience + Military Experience + Performance	0.002	0.002	0.008	
Conditional dominance *k* = 4	0.002	0.003	0.014	0.017
Gender + Video Gaming Experience + Military Experience + Age + Task Type			0.014	0.017
Gender + Video Gaming Experience + Military Experience + Age + Visual Complexity		0.004		0.011
Gender + Video Gaming Experience + Military Experience + Age + Performance		0.002	0.007	
Gender + Video Gaming Experience + Military Experience + Task Type + Visual Complexity	0.000			0.010
Gender + Video Gaming Experience + Military Experience + Task Type + Performance	0.001		0.008	
Gender + Video Gaming Experience + Military Experience + Visual Complexity + Performance	0.000	0.002		
Conditional dominance *k* = 5	0.000	0.002	0.010	0.013
Gender + Video Gaming Experience + Military Experience + Age + Task Type + Visual Complexity				0.010
Gender + Video Gaming Experience + Military Experience + Age + Task Type + Performance			0.007	

	Additional contribution (pseudo R^2) of			
Subset Model ***X***	Age	Task Type	Visual Complexity	Performance
Gender + Video Gaming Experience + Military Experience + Age + Visual Complexity + Performance		0.002		
Gender + Video Gaming Experience + Military Experience + Task Type + Visual Complexity + Performance	0.000			
Conditional dominance *k* = 6	0.000	0.002	0.007	0.010
Gender + Video Gaming Experience + Military Experience + Age + Task Type + Visual Complexity + Performance				
Overall average	0.001	0.003	0.010	0.013

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, keeping video gaming experience, military experience, and gender constant (Azen & Budescu, 2003). Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

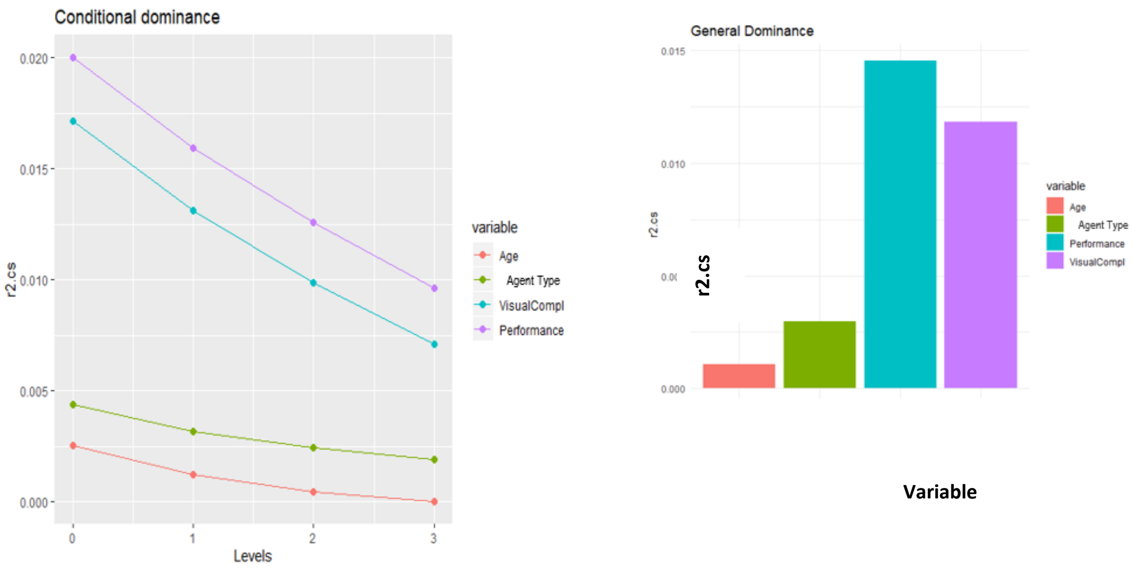


Figure 55. Conditional and general dominance results full model in Study A.

Note. The conditional and general dominance patterns conformed to the complete dominance patterns for the full model in Study A. The performance subscale of the NASA-TLX dominated all other predictors, followed by visual xomplexity, agent morphology type, and age.

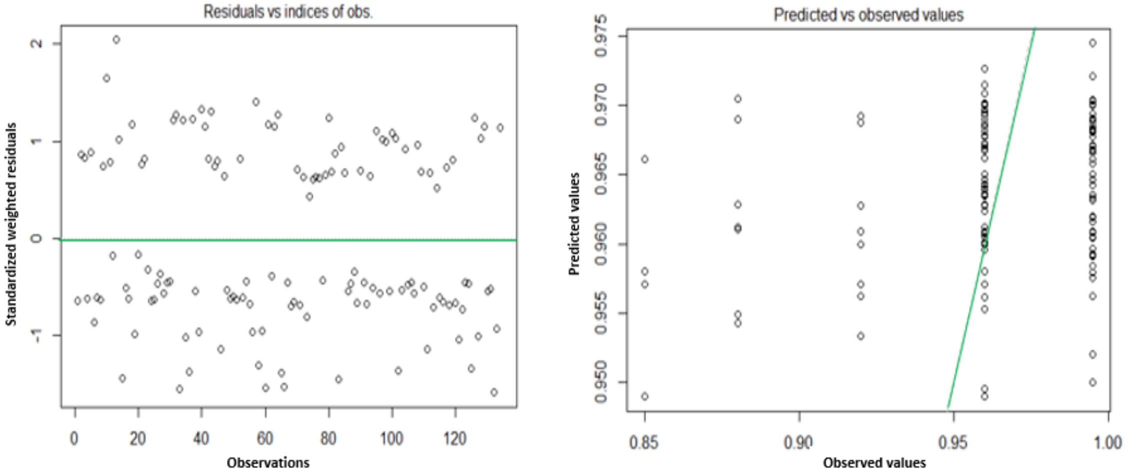


Figure 56. Study A full model evaluation plots

Note. The residual plot is on the left, the predicted vs. observed values on the right, and a fitted line based on maximum likelihood.

Study B

Linearity Check

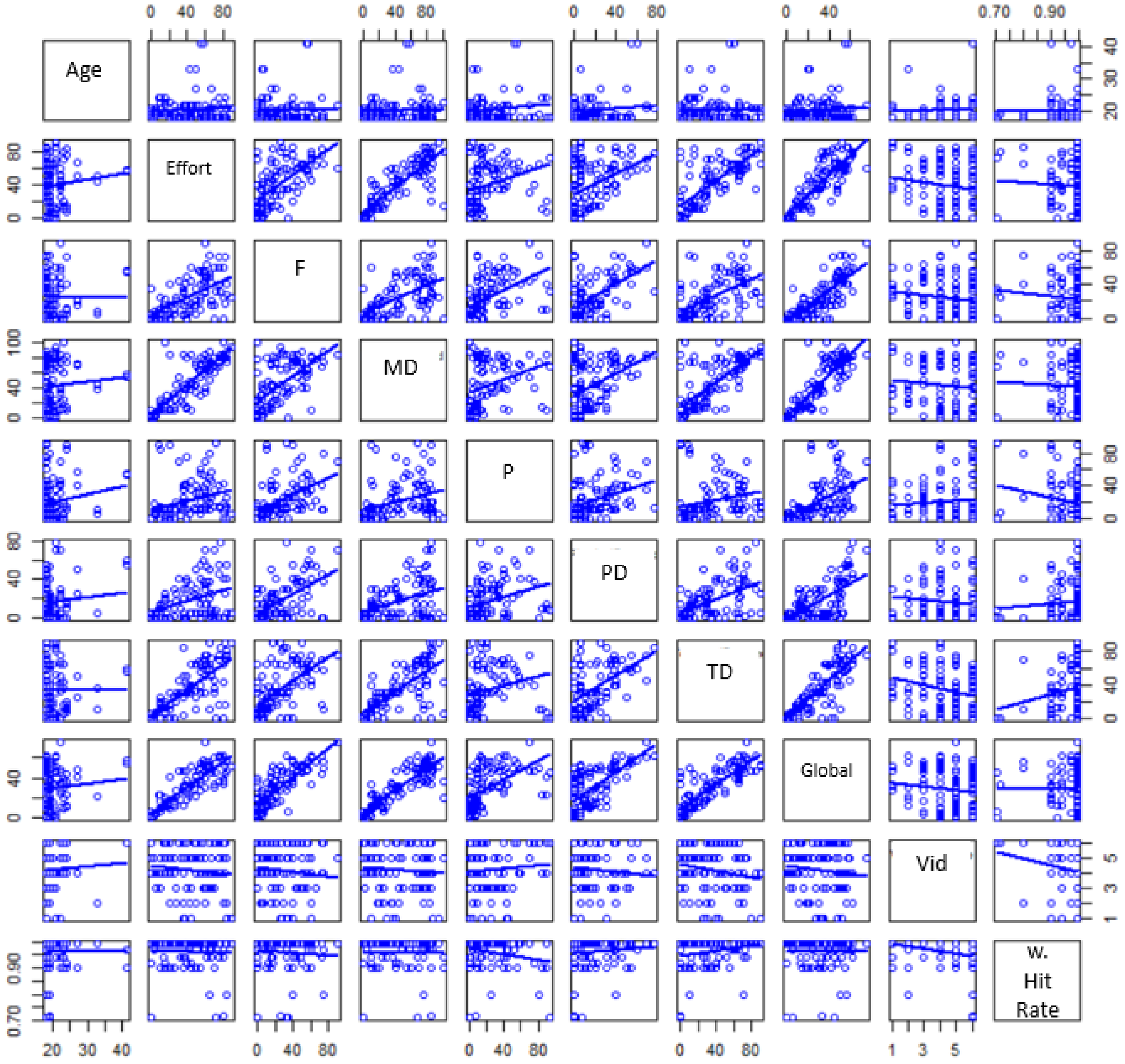


Figure 57. Scatterplot matrix continuous variables in Study B.

Note. Spearman's correlation was used. The abbreviations represent: F = Frustration subscale on NASA-TLX, MD = Mental Demand subscale on NASA-TLX, P = Performance subscale on NASA-TLX, PD = Physical Demand subscale on NASA-TLX, TD = Temporal Demand subscale on NASA-TLX, Global = average score on NASA-TLX, Vid = Video Gaming Experience, w. Hit Rate = winsorized hit rate.

Dominance Analysis

Human/Agent Qualities

Table 51

Raw dominance analysis results Human/Agent Qualities in Study B.

	Additional contribution (pseudo R^2) of		
Subset model ***X***	Age	Gender	Military Experience
k = 0	0.008	0.005	0.014
Age		0.006	0.006
Gender	0.008		0.012
Military Experience	0.000	0.004	
Conditional dominance *k* = 1	0.004	0.005	0.009
Age + Gender			0.005
Age + Military Experience		0.004	
Gender + Military Experience	0.001		
Conditional dominance *k* = 2	0.001	0.004	0.005
Age + Gender + Military Experience			
Overall average	0.004	0.005	0.009

Note. This table presents the raw output of the dominance analyses, wherein video gaming experience was excluded due to a large number of missing values. The unique additional contribution of each predictor is shown over all possible subset model sizes, wherein *k* = 0 indicates that no other variables are in the model aside of the predictor under evaluation. Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

Task Perception

Table 52

Raw dominance analysis results Task Perception in Study B.

	Additional contribution (pseudo R^2) of					
Subset model ***X***	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
k = 0	0.015	0.001	0.001	0.002	0.001	0.005
Performance		0.000	0.000	0.007	0.000	0.000
Mental Demand	0.014		0.000	0.012	0.000	0.005
Physical Demand	0.015	0.000		0.004	0.001	0.004
Temporal Demand	0.020	0.010	0.003		0.012	0.017
Effort	0.014	0.000	0.000	0.013		0.004
Frustration	0.011	0.001	0.000	0.015	0.000	
Conditional dominance *k* = 1	0.015	0.002	0.001	0.010	0.003	0.006
Performance + Mental Demand			0.000	0.014	0.002	0.000
Performance + Physical Demand		0.000		0.008	0.000	0.000
Performance + Temporal Demand		0.008	0.001		0.014	0.006
Performance + Effort		0.002	0.000	0.021		0.000
Performance + Frustration		0.000	0.000	0.013	0.000	
Mental Demand + Physical Demand	0.014			0.013	0.000	0.005
Mental Demand + Temporal Demand	0.017		0.002		0.003	0.010
Mental Demand + Effort	0.016		0.000	0.014		0.005
Mental Demand + Frustration	0.010		0.000	0.017	0.000	
Physical Demand + Temporal Demand	0.018	0.009			0.011	0.015
Physical Demand + Effort	0.014	0.000		0.015		0.004
Physical Demand + Frustration	0.010	0.001		0.015	0.000	
Temporal Demand + Effort	0.022	0.001	0.002			0.011
Temporal Demand + Frustration	0.009	0.003	0.000		0.006	
Effort + Frustration	0.010	0.001	0.000	0.020		
Conditional dominance *k* = 2	0.014	0.003	0.001	0.015	0.004	0.006
Performance + Mental Demand + Physical Demand				0.015	0.002	0.001
Performance + Mental Demand + Temporal Demand			0.000		0.006	0.003
Performance + Mental Demand + Effort			0.000	0.019		0.000
Performance + Mental Demand + Frustration			0.000	0.017	0.001	
Performance + Physical Demand + Temporal Demand		0.007			0.014	0.005
Performance + Physical Demand + Effort		0.001		0.021		0.000
Performance + Physical Demand + Frustration		0.000		0.013	0.000	

	Additional contribution (pseudo R^2) of					
Subset model ***X***	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
Performance + Temporal Demand + Effort		0.000	0.001			0.002
Performance + Temporal Demand + Frustration		0.004	0.000		0.010	
Performance + Effort + Frustration		0.002	0.000	0.023		
Mental Demand + Physical Demand + Temporal Demand	0.016				0.003	0.009
Mental Demand + Physical Demand + Effort	0.016			0.015		0.005
Mental Demand + Physical Demand + Frustration	0.010			0.017	0.000	
Mental Demand + Temporal Demand + Effort	0.021		0.002			0.010
Mental Demand + Temporal Demand + Frustration	0.009		0.000		0.002	
Mental Demand + Effort + Frustration	0.011		0.000	0.019		
Physical Demand + Temporal Demand + Effort	0.021	0.001				0.010
Physical Demand + Temporal Demand + Frustration	0.008	0.003			0.006	
Physical Demand + Effort + Frustration	0.010	0.001		0.020		
Temporal Demand + Effort + Frustration	0.013	0.000	0.000			
Conditional dominance *k* = 3	0.013	0.002	0.000	0.018	0.004	0.004
Performance + Mental Demand + Physical Demand + Temporal Demand					0.006	0.002
Performance + Mental Demand + Physical Demand + Effort				0.020		0.000
Performance + Mental Demand + Physical Demand + Frustration				0.016	0.001	
Performance + Mental Demand + Temporal Demand + Effort			0.000			0.002
Performance + Mental Demand + Temporal Demand + Frustration			0.000		0.006	
Performance + Mental Demand + Effort + Frustration			0.000	0.021		
Performance + Physical Demand + Temporal Demand + Effort		0.000				0.002
Performance + Physical Demand + Temporal Demand + Frustration		0.004			0.010	
Performance + Physical Demand + Effort + Frustration		0.002		0.022		
Performance + Temporal Demand + Effort + Frustration		0.000	0.000			

	Additional contribution (pseudo R^2) of					
Subset model ***X***	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
Mental Demand + Physical Demand + Temporal Demand + Effort	0.020					0.009
Mental Demand + Physical Demand + Temporal Demand + Frustration	0.009				0.002	
Mental Demand + Physical Demand + Effort + Frustration	0.011			0.019		
Mental Demand + Temporal Demand + Effort + Frustration	0.013		0.000			
Physical Demand + Temporal Demand + Effort + Frustration	0.013	0.000				
Conditional dominance *k* = 4	0.013	0.001	0.000	0.020	0.005	0.003
Performance + Mental Demand + Physical Demand + Temporal Demand + Effort						0.002
Performance + Mental Demand + Physical Demand + Temporal Demand + Frustration					0.006	
Performance + Mental Demand + Physical Demand + Effort + Frustration				0.021		
Performance + Mental Demand + Temporal Demand + Effort + Frustration			0.000			
Performance + Physical Demand + Temporal Demand + Effort + Frustration		0.000				
Mental Demand + Physical Demand + Temporal Demand + Effort + Frustration	0.013					
Conditional dominance *k* = 5	0.013	0.000	0.000	0.021	0.006	0.002
Performance + Mental Demand + Physical Demand + Temporal Demand + Effort + Frustration						
Overall average	0.014	0.001	0.000	0.014	0.004	0.004

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, wherein *k* = 0 indicates that no other variables are in the model aside of the predictor under evaluation. Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

Task Composition

Table 53

Raw dominance analysis results Task Composition in Study B.

	Additional contribution (pseudo R^2) of		
Subset model ***X***	Event Rate	Task Type	Task Duration
k = 0	0.055	0.000	0.004
Event Rate		0.000	0.001
Task Type	0.055		0.006
Task Duration	0.051	0.002	
Conditional dominance *k* = 1	0.053	0.001	0.004
Event Rate + Task Type			0.002
Event Rate + Task Duration		0.002	
Task Type + Task Duration	0.051		
Conditional dominance *k* = 2	0.051	0.002	0.002
Event Rate + Task Type + Task Duration			
Overall average	0.053	0.001	0.003

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, wherein *k* = 0 indicates that no other variables are in the model aside of the predictor under evaluation. Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

Full Model

Table 54

Raw dominance analysis results Full Model in Study B.

	Additional contribution (pseudo R^2) of		
Subset model ***X***	Military Experience	Event Rate	Performance
Age + Gender	0.000	0.048	0.017
Conditional dominance *k* = 2	0.000	0.048	0.017
Age + Gender + Military Experience		0.052	0.017
Age + Gender + Event Rate	0.005		0.015
Age + Gender + Performance	0.000	0.045	
Conditional dominance *k* = 3	0.002	0.049	0.016

Age + Gender + Military Experience + Event Rate			0.014
Age + Gender + Military Experience + Performance		0.050	
Age + Gender + Event Rate + Performance	0.004		
Conditional dominance *k* = 4	0.004	0.050	0.014
Age + Gender + Military Experience + Event Rate + Performance			
Overall average	0.002	0.049	0.016

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, keeping age and gender constant (Azen & Budescu, 2003). Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

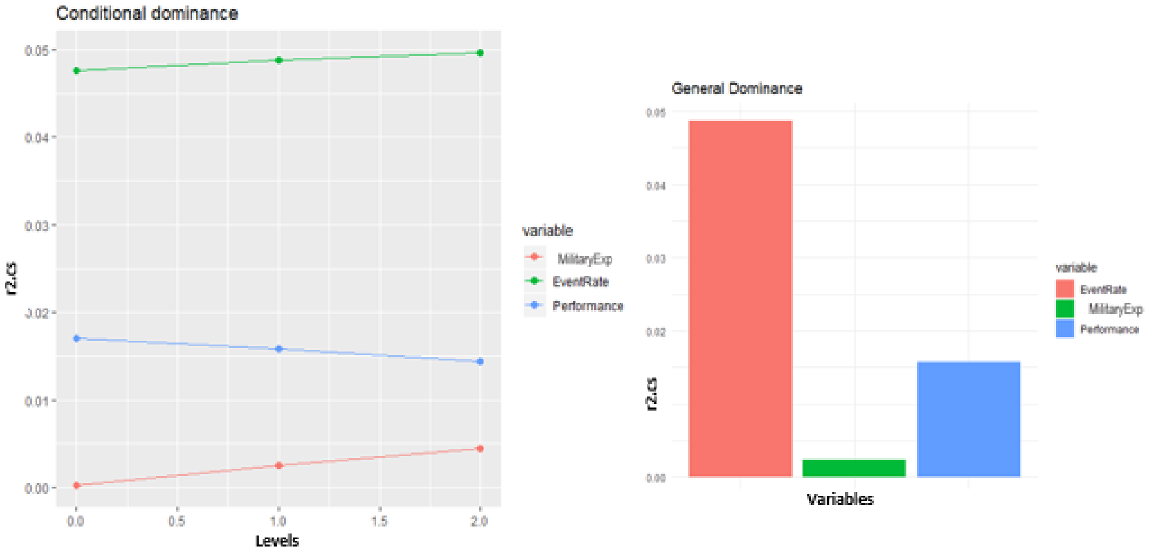


Figure 58. Conditional and general dominance results Full Model in Study B.

Note. The conditional dominance plot (left) shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model. The general dominance bar graph (right) shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) averaged over all possible subset model sizes.

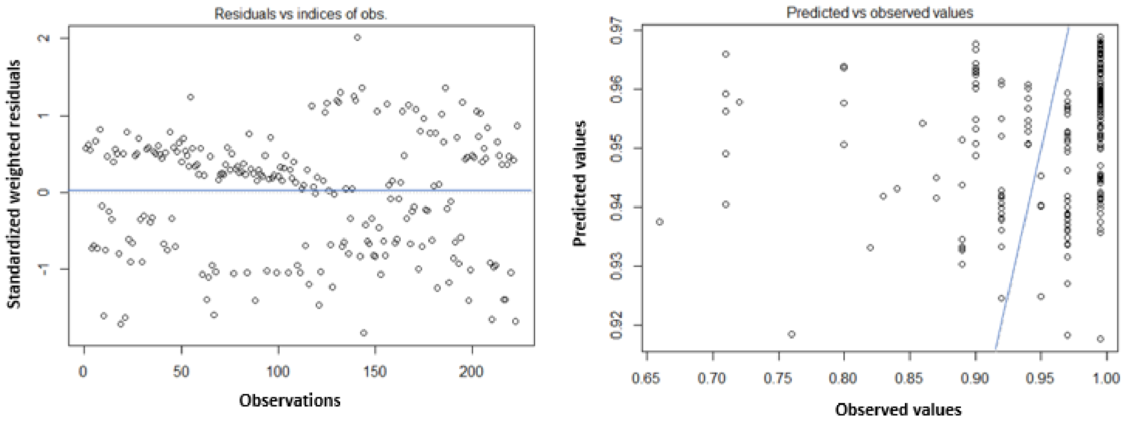


Figure 59. Study B full model evaluation plots.

Note. The residual plot is shown on the left and the predicted vs. observed values on the right, with a fitted line based on maximum likelihood.

Study C

Linearity Check

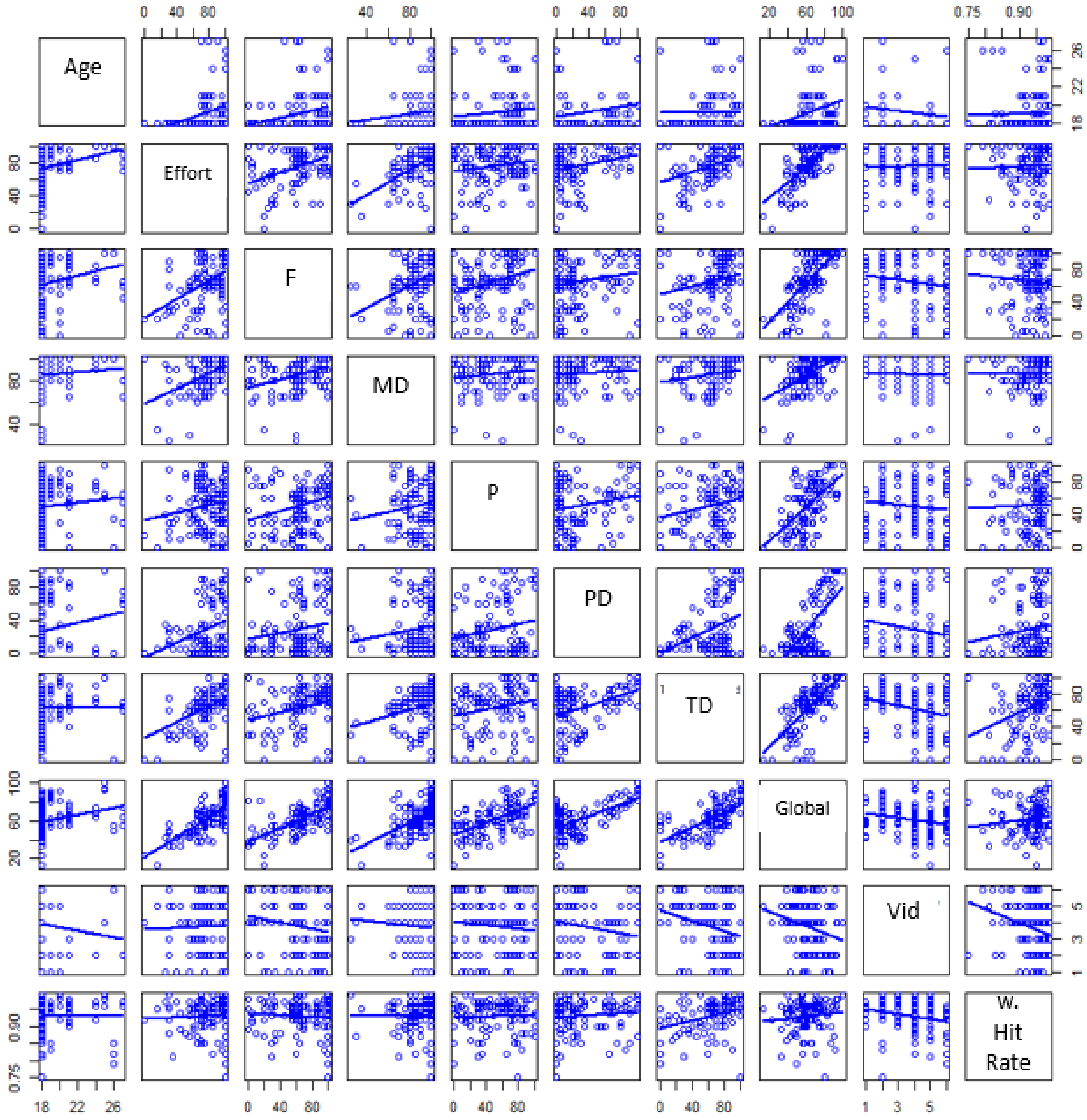


Figure 60. Scatterplot matrix between continuous variables in Study C.

Note. Spearman's correlation was used. The abbreviations represent: F = Frustration subscale on NASA-TLX, MD = Mental Demand subscale on NASA-TLX, P = Performance subscale on NASA-TLX, PD = Physical Demand subscale on NASA-TLX, TD = Temporal Demand subscale on NASA-TLX, Global = average score on NASA-TLX, Vid = Video Gaming Experience, w. Hit Rate = winsorized hit rate.

Dominance Analysis

Human/Agent Qualities

Table 55

Raw dominance analysis results Human/Agent Qualities in Study C.

	Additional contribution (pseudo R^2) of			
Subset model ***X***	Age	Gender	Military Experience	Video Gaming Experience
k = 0	0.010	0.027	0.013	0.048
Age		0.022	0.012	0.041
Gender	0.005		0.020	0.027
Military Experience	0.009	0.034		0.058
Video Gaming Experience	0.003	0.006	0.022	
Conditional dominance *k* = 1	0.005	0.021	0.018	0.042
Age + Gender			0.018	0.024
Age + Military Experience		0.028		0.050
Age + Video Gaming Experience		0.005	0.021	
Gender + Military Experience	0.003			0.032
Gender + Video Gaming Experience	0.002		0.025	
Military Experience + Video Gaming Experience	0.001	0.008		
Conditional dominance *k* = 2	0.002	0.014	0.021	0.035
Age + Gender + Military Experience				0.029
Age + Gender + Video Gaming Experience			0.023	
Age + Military Experience + Video Gaming Experience		0.007		
Gender + Military Experience + Video Gaming Experience	0.000			
Conditional dominance *k* = 3	0.000	0.007	0.023	0.029
Age + Gender + Military Experience + Video Gaming Experience				
Overall average	0.004	0.017	0.019	0.039

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, wherein *k* = 0 indicates that no other variables are in the model aside of the predictor under evaluation. Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

Task Perception

Table 56

Raw dominance analysis results Task Perception in Study C.

	Additional contribution (pseudo R^2) of					
Subset model ***X***	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
k = 0	0.004	0.000	0.031	0.117	0.006	0.003
Performance		0.000	0.028	0.113	0.004	0.006
Mental Demand	0.005		0.032	0.128	0.009	0.003
Physical Demand	0.002	0.001		0.088	0.000	0.006
Temporal Demand	0.000	0.011	0.002		0.008	0.019
Effort	0.003	0.003	0.025	0.120		0.007
Frustration	0.008	0.000	0.035	0.134	0.011	
Conditional dominance *k* = 1	0.003	0.003	0.025	0.117	0.007	0.008
Performance + Mental Demand			0.030	0.123	0.008	0.006
Performance + Physical Demand		0.002		0.086	0.000	0.010
Performance + Temporal Demand		0.011	0.002		0.009	0.021
Performance + Effort		0.004	0.024	0.117		0.011
Performance + Frustration		0.000	0.032	0.128	0.009	
Mental Demand + Physical Demand	0.002			0.098	0.002	0.005
Mental Demand + Temporal Demand	0.000		0.002		0.002	0.013
Mental Demand + Effort	0.003		0.025	0.121		0.006
Mental Demand + Frustration	0.008		0.035	0.138	0.013	
Physical Demand + Temporal Demand	0.000	0.011			0.010	0.020
Physical Demand + Effort	0.001	0.003		0.097		0.008
Physical Demand + Frustration	0.005	0.000		0.102	0.003	
Temporal Demand + Effort	0.000	0.004	0.003			0.012
Temporal Demand + Frustration	0.002	0.004	0.002		0.001	
Effort + Frustration	0.006	0.002	0.026	0.124		

	Additional contribution (pseudo R^2) of					
Subset model ***X***	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
Conditional dominance *k* = 2	0.003	0.004	0.018	0.113	0.006	0.011
Performance + Mental Demand + Physical Demand				0.095	0.002	0.008
Performance + Mental Demand + Temporal Demand			0.002		0.002	0.015
Performance + Mental Demand + Effort			0.024	0.118		0.010
Performance + Mental Demand + Frustration			0.032	0.132	0.012	
Performance + Physical Demand + Temporal Demand		0.011			0.010	0.021
Performance + Physical Demand + Effort		0.004		0.096		0.012
Performance + Physical Demand + Frustration		0.000		0.098	0.002	
Performance + Temporal Demand + Effort		0.004	0.003			0.014
Performance + Temporal Demand + Frustration		0.004	0.002		0.001	
Performance + Effort + Frustration		0.003	0.025	0.120		
Mental Demand + Physical Demand + Temporal Demand	0.000				0.003	0.013
Mental Demand + Physical Demand + Effort	0.002			0.098		0.007
Mental Demand + Physical Demand + Frustration	0.005			0.105	0.004	
Mental Demand + Temporal Demand + Effort	0.000		0.003			0.011
Mental Demand + Temporal Demand + Frustration	0.002		0.002		0.000	
Mental Demand + Effort + Frustration	0.007		0.026	0.125		
Physical Demand + Temporal Demand + Effort	0.000	0.004				0.012
Physical Demand + Temporal Demand + Frustration	0.002	0.004			0.002	
Physical Demand + Effort + Frustration	0.005	0.002		0.101		
Temporal Demand + Effort + Frustration	0.002	0.003	0.003			

	Additional contribution (pseudo R^2) of					
Subset model ***X***	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
Conditional dominance *k* = 3	0.003	0.004	0.012	0.109	0.004	0.012
Performance + Mental Demand + Physical Demand + Temporal Demand					0.003	0.015
Performance + Mental Demand + Physical Demand + Effort				0.096		0.010
Performance + Mental Demand + Physical Demand + Frustration				0.102	0.004	
Performance + Mental Demand + Temporal Demand + Effort			0.003			0.012
Performance + Mental Demand + Temporal Demand + Frustration			0.002		0.000	
Performance + Mental Demand + Effort + Frustration			0.024	0.120		
Performance + Physical Demand + Temporal Demand + Effort		0.004				0.013
Performance + Physical Demand + Temporal Demand + Frustration		0.004			0.002	
Performance + Physical Demand + Effort + Frustration		0.002		0.097		
Performance + Temporal Demand + Effort + Frustration		0.003	0.003			
Mental Demand + Physical Demand + Temporal Demand + Effort	0.000					0.010
Mental Demand + Physical Demand + Temporal Demand + Frustration	0.002				0.000	
Mental Demand + Physical Demand + Effort + Frustration	0.005			0.102		
Mental Demand + Temporal Demand + Effort + Frustration	0.002		0.003			
Physical Demand + Temporal Demand + Effort + Frustration	0.002	0.003				
Conditional dominance *k* = 4	0.002	0.003	0.007	0.103	0.002	0.012
Performance + Mental Demand + Physical Demand + Temporal Demand + Effort						0.012

	Additional contribution (pseudo R^2) of					
Subset model ***X***	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
Performance + Mental Demand + Physical Demand + Temporal Demand + Frustration					0.000	
Performance + Mental Demand + Physical Demand + Effort + Frustration				0.098		
Performance + Mental Demand + Temporal Demand + Effort + Frustration			0.002			
Performance + Physical Demand + Temporal Demand + Effort + Frustration		0.003				
Mental Demand + Physical Demand + Temporal Demand + Effort + Frustration	0.002					
Conditional dominance *k* = 5	0.002	0.003	0.002	0.098	0.000	0.012
Performance + Mental Demand + Physical Demand + Temporal Demand + Effort + Frustration						
Overall average	0.003	0.003	0.016	0.110	0.004	0.010

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, wherein *k* = 0 indicates that no other variables are in the model aside of the predictor under evaluation. Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

Full Model

Table 57

Raw dominance analysis results Full Model in Study C.

	Additional contribution (pseudo R^2) of		
Subset model *X*	Video Gaming Experience	Temporal Demand	Agent Report Modality
Age + Gender + Military Experience	0.039	0.091	0.019
Conditional dominance *k* = 3	0.039	0.091	0.019
Age + Gender + Military Experience + Video Gaming Experience		0.089	0.018
Age + Gender + Military Experience + Temporal Demand	0.037		0.026
Age + Gender + Military Experience + Agent Report Modality	0.038	0.098	
Conditional dominance *k* = 4	0.037	0.093	0.022
Age + Gender + Military Experience + Video Gaming Experience + Temporal Demand			0.025
Age + Gender + Military Experience + Video Gaming Experience + Agent Report Modality		0.096	
Age + Gender + Military Experience + Temporal + Agent Report Modality	0.036		
Conditional dominance *k* = 5	0.036	0.096	0.025
Age + Gender + Military Experience + Video Gaming Experience + Temporal + Agent Report Modality			
Overall average	0.037	0.093	0.022

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, keeping age and gender constant (Azen & Budescu, 2003). Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

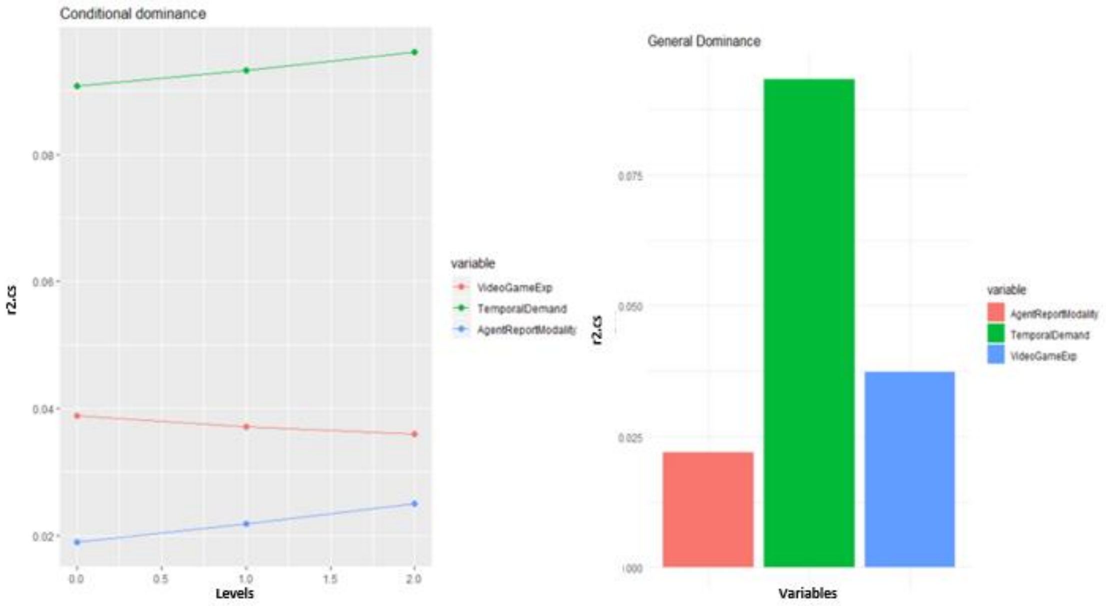


Figure 61. Conditional and general dominance results Full Model in Study C.

Note. The conditional dominance plot (left) shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model. The general dominance bar graph (right) shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) averaged over all possible subset model sizes.

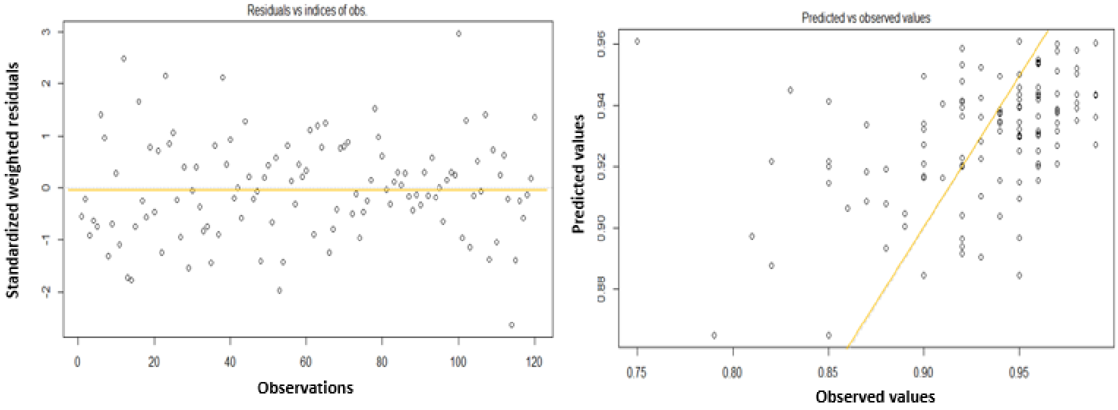


Figure 62. Study C full model evaluation plots.

Note. The residual plot is shown on the left and the predicted vs. observed values on the right, with a fitted line based on maximum likelihood.

Study D

Linearity

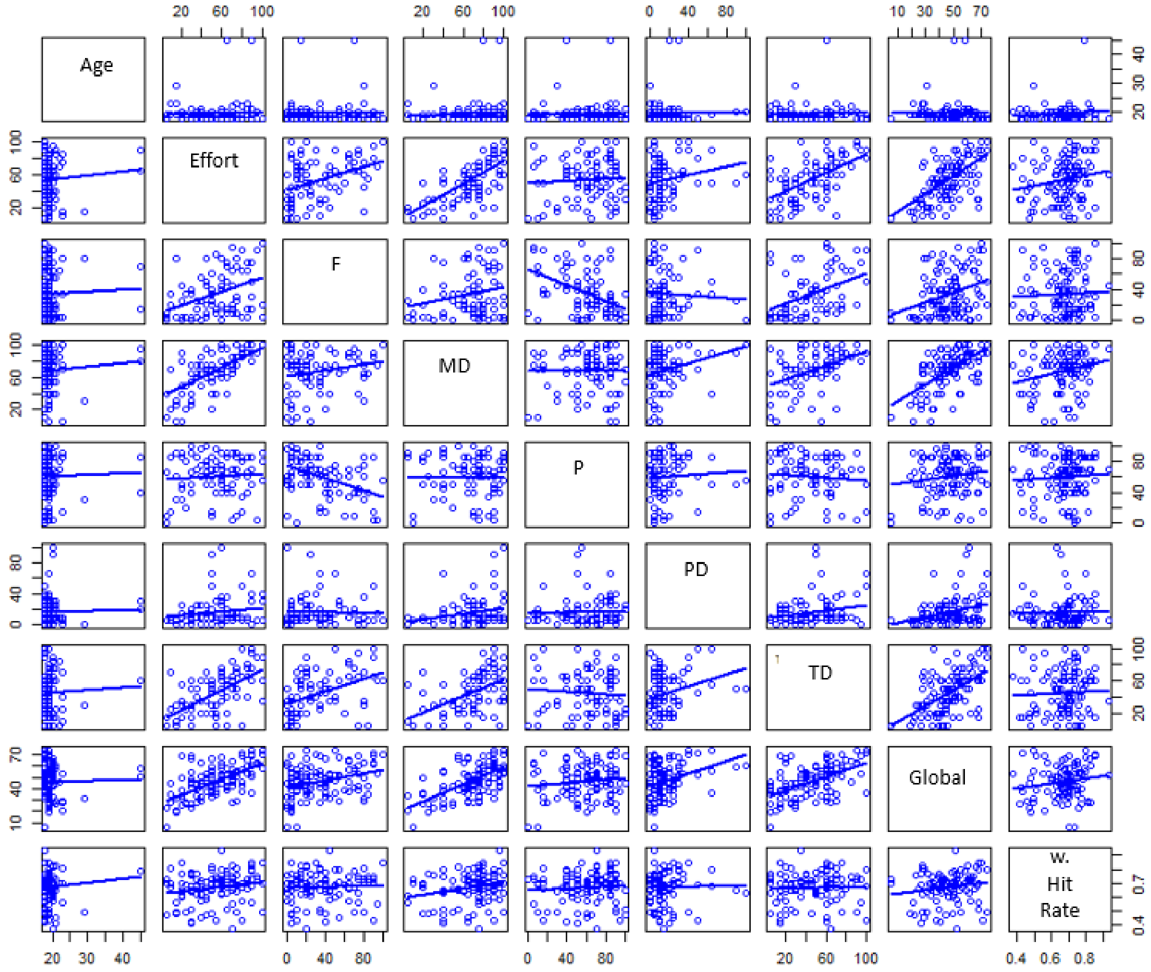


Figure 63. Scatterplot matrix between continuous variables in Study D.

Note. Spearman’s correlation was used. The abbreviations represent: F = Frustration subscale on NASA-TLX, MD = Mental Demand subscale on NASA-TLX, P = Performance subscale on NASA-TLX, PD = Physical Demand subscale on NASA-TLX, TD = Temporal Demand subscale on NASA-TLX, Global = average score on NASA-TLX, w. Hit Rate = winsorized hit rate.

Dominance Analysis

Human/Agent Qualities

Table 58

Raw dominance analysis results Human/Agent Qualities in Study D.

	Additional contribution (pseudo R^2) of	
Subset model *X*	Age	Gender
$k = 0$	0.011	0.125
Age		0.117
Gender	0.004	
Conditional dominance $k = 1$	0.004	0.117
Age + Gender		
Overall average	0.007	0.121

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, wherein $k = 0$ indicates that no other variables are in the model aside of the predictor under evaluation. Conditional dominance indicates the average unique contribution for that subset model size (k) for the predictor under evaluation. The overall average presents the average over all average k model sizes.

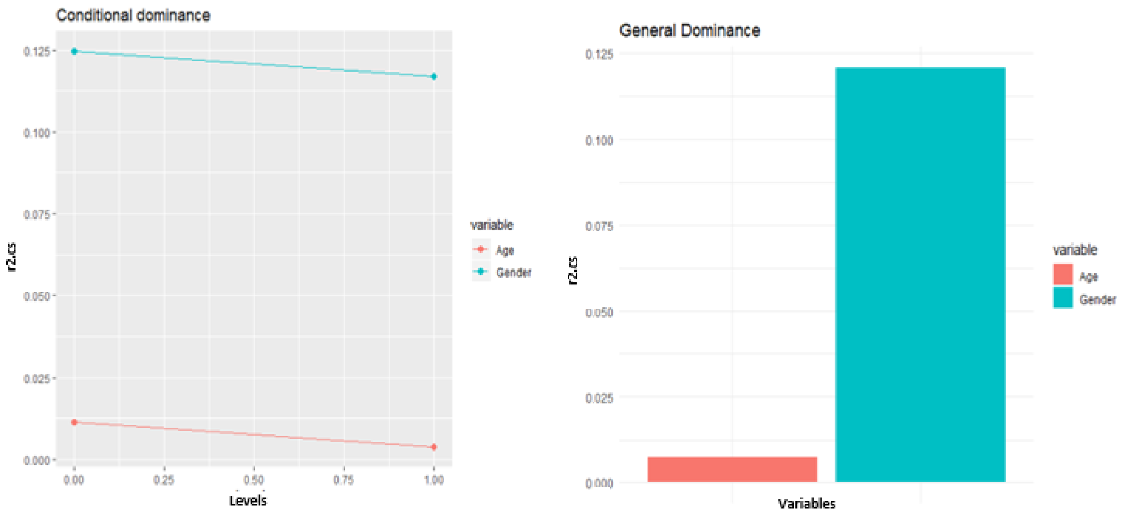


Figure 64. Conditional and general dominance results Human/Agent Qualities in Study D.

Note. The conditional dominance plot (left) shows the unique contribution (in Cox & Snell’s (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model. The general dominance bar graph (right) shows the unique contribution (in Cox & Snell’s (2018) pseudo R^2) averaged over all possible subset model sizes.

Task Perception

Table 59

Raw dominance analysis results Task Perception in Study D.

	Additional contribution (pseudo R^2) of					
Subset model *X*	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
k = 0	0.007	0.060	0.001	0.002	0.038	0.004
Performance		0.059	0.001	0.003	0.036	0.013
Mental Demand	0.006		0.000	0.005	0.003	0.000
Physical Demand	0.006	0.059		0.002	0.037	0.005
Temporal Demand	0.007	0.062	0.001		0.043	0.003
Effort	0.004	0.024	0.000	0.007		0.000
Frustration	0.015	0.055	0.002	0.001	0.034	
Conditional dominance *k* = 1	0.008	0.052	0.001	0.003	0.030	0.004
Performance + Mental Demand			0.001	0.004	0.002	0.003
Performance + Physical Demand		0.059		0.002	0.035	0.013
Performance + Temporal Demand		0.060	0.000		0.039	0.010
Performance + Effort		0.025	0.000	0.006		0.001
Performance + Frustration		0.049	0.001	0.000	0.024	
Mental Demand + Physical Demand	0.006			0.005	0.003	0.000
Mental Demand + Temporal Demand	0.005		0.000		0.008	0.002
Mental Demand + Effort	0.005		0.000	0.011		0.000
Mental Demand + Frustration	0.008		0.000	0.007	0.003	
Physical Demand + Temporal Demand	0.007	0.062			0.043	0.003
Physical Demand + Effort	0.004	0.025		0.007		0.000
Physical Demand + Frustration	0.014	0.054		0.000	0.032	
Temporal Demand + Effort	0.003	0.028	0.000			0.000
Temporal Demand + Frustration	0.014	0.061	0.001		0.040	
Effort + Frustration	0.005	0.024	0.000	0.007		
Conditional dominance *k = 2*	0.007	0.045	0.000	0.005	0.023	0.003
Performance + Mental Demand + Physical Demand				0.004	0.002	0.002
Performance + Mental Demand + Temporal Demand			0.000		0.007	0.007
Performance + Mental Demand + Effort			0.001	0.009		0.001
Performance + Mental Demand + Frustration			0.000	0.008	0.001	
Performance + Physical Demand + Temporal Demand		0.060			0.039	0.011
Performance + Physical Demand + Effort		0.026		0.006		0.001

	Additional contribution (pseudo R^2) of					
Subset model *X*	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
Performance + Physical Demand + Frustration		0.048		0.000	0.023	
Performance + Temporal Demand + Effort		0.028	0.000			0.002
Performance + Temporal Demand + Frustration		0.057	0.001		0.031	
Performance + Effort + Frustration		0.025	0.000	0.007		
Mental Demand + Physical Demand + Temporal Demand	0.005				0.008	0.002
Mental Demand + Physical Demand + Effort	0.005			0.010		0.000
Mental Demand + Physical Demand + Frustration	0.008			0.006	0.003	
Mental Demand + Temporal Demand + Effort	0.004		0.000			0.000
Mental Demand + Temporal Demand + Frustration	0.010		0.000		0.007	
Mental Demand + Effort + Frustration	0.006		0.000	0.011		
Physical Demand + Temporal Demand + Effort	0.003	0.028				0.000
Physical Demand + Temporal Demand + Frustration	0.014	0.060			0.039	
Physical Demand + Effort + Frustration	0.005	0.024		0.007		
Temporal Demand + Effort + Frustration	0.005	0.028	0.000			
Conditional dominance *k = 3*	0.007	0.039	0.000	0.007	0.016	0.003
Performance + Mental Demand + Physical Demand + Temporal Demand					0.007	0.006
Performance + Mental Demand + Physical Demand + Effort				0.009		0.001
Performance + Mental Demand + Physical Demand + Frustration				0.008	0.001	
Performance + Mental Demand + Temporal Demand + Effort			0.000			0.003
Performance + Mental Demand + Temporal Demand + Frustration			0.000		0.004	
Performance + Mental Demand + Effort + Frustration			0.001	0.011		
Performance + Physical Demand + Temporal Demand + Effort		0.028				0.002
Performance + Physical Demand + Temporal Demand + Frustration		0.056			0.030	

	Additional contribution (pseudo R^2) of					
Subset model *X*	Performance	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration
Performance + Physical Demand + Effort + Frustration		0.026		0.007		
Performance + Temporal Demand + Effort + Frustration		0.030	0.000			
Mental Demand + Physical Demand + Temporal Demand + Effort	0.004					0.000
Mental Demand + Physical Demand + Temporal Demand + Frustration	0.010				0.007	
Mental Demand + Physical Demand + Effort + Frustration	0.006			0.011		
Mental Demand + Temporal Demand + Effort + Frustration	0.006		0.000			
Physical Demand + Temporal Demand + Effort + Frustration	0.005	0.028				
Conditional dominance *k = 4*	0.006	0.034	0.000	0.009	0.010	0.003
Performance + Mental Demand + Physical Demand + Temporal Demand + Effort						0.003
Performance + Mental Demand + Physical Demand + Temporal Demand + Frustration					0.004	
Performance + Mental Demand + Physical Demand + Effort + Frustration				0.011		
Performance + Mental Demand + Temporal Demand + Effort + Frustration			0.000			
Performance + Physical Demand + Temporal Demand + Effort + Frustration		0.030				
Mental Demand + Physical Demand + Temporal Demand + Effort + Frustration	0.006					
Conditional dominance *k = 5*	0.006	0.030	0.000	0.011	0.004	0.003
Performance + Mental Demand + Physical Demand + Temporal Demand + Effort + Frustration						
Overall average	0.007	0.043	0.000	0.006	0.020	0.003

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, wherein k = 0 indicates that no other variables are in the model aside of the predictor under evaluation. Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

Task Composition

Table 60

Raw dominance analysis results Task Composition in Study D.

	Additional contribution (pseudo R^2) of	
Subset model *X*	Delivery Frequency	Agent Report Modality
k = 0	0.007	0.028
Delivery Frequency		0.028
Agent Report Modality	0.007	
Conditional dominance *k* = 1	0.007	0.028
Delivery Frequency + Agent Report Modality		
Overall average	0.007	0.028

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, wherein *k* = 0 indicates that no other variables are in the model aside of the predictor under evaluation. Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

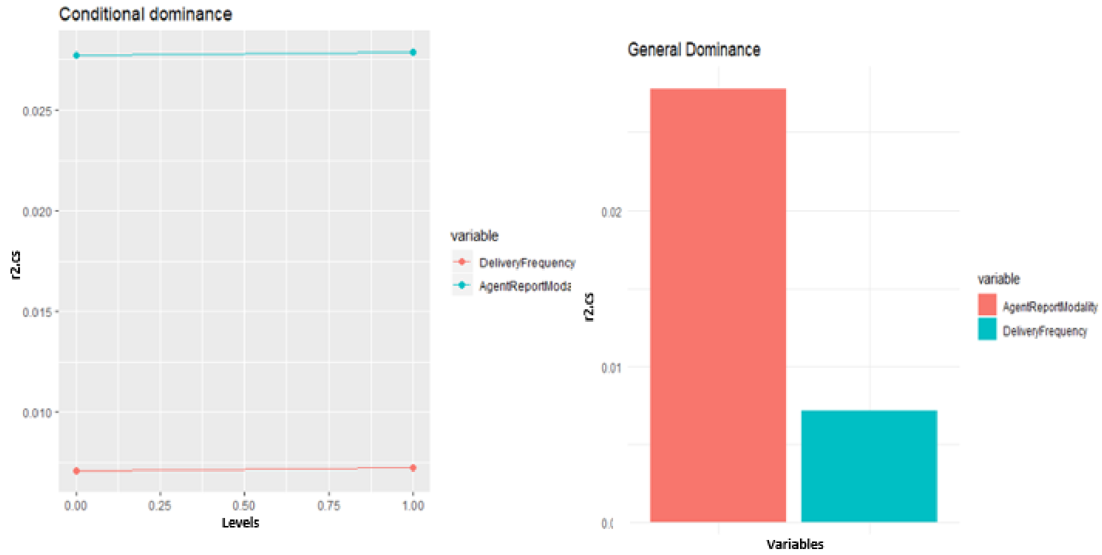


Figure 65. Conditional and general dominance results Task Composition in Study D.

Note. The conditional dominance plot (left) shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model. The general dominance bar graph (right) shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) averaged over all possible subset model sizes.

Full Model

Table 61

Raw dominance analysis results Full Model in Study D.

	Additional contribution (pseudo R^2) of		
Subset model ***X***	Gender	Mental Demand	Agent Report Modality
Age	0.117	0.055	0.025
Conditional dominance *k* = 1	0.117	0.055	0.025
Age + Gender		0.072	0.026
Age + Mental Demand	0.134		0.036
Age + Agent Report Modality	0.118	0.067	
Conditional dominance *k* = 2	0.126	0.069	0.031
Age + Gender + Mental Demand			0.041
Age + Gender + Agent Report Modality		0.087	
Age + Mental Demand + Agent Report Modality	0.138		
Conditional dominance *k* = 3	0.138	0.087	0.041
Age + Gender + Mental Demand + Agent Report Modality			
Overall average	0.127	0.070	0.032

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, keeping age constant (Azen & Budescu, 2003). Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

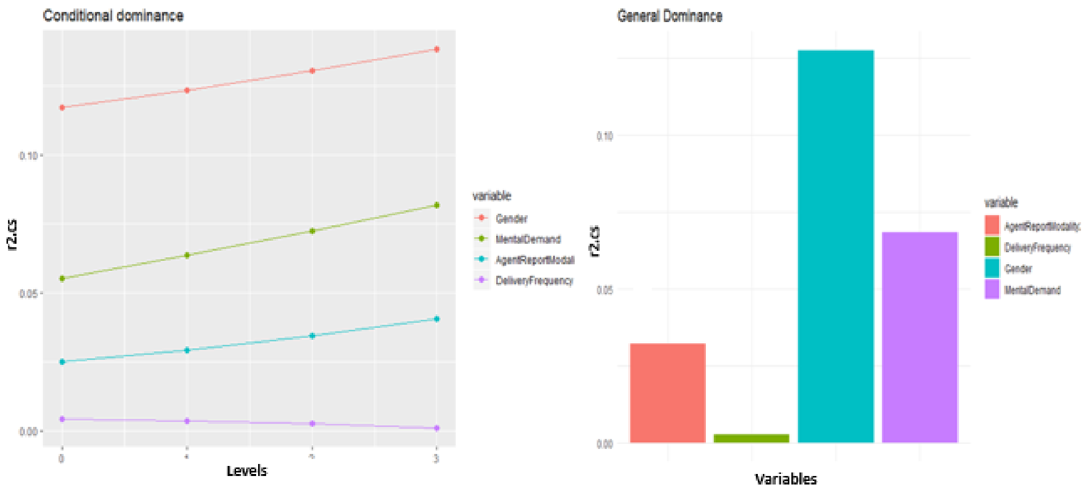


Figure 66. Conditional and general dominance results Full Model in Study D.

Note. The conditional dominance plot (left) shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) over different subset model sizes or levels. For example, a level of 1 indicates that one additional predictor is in the regression model. The general dominance bar graph (right) shows the unique contribution (in Cox & Snell's (2018) pseudo R^2) averaged over all possible subset model sizes.

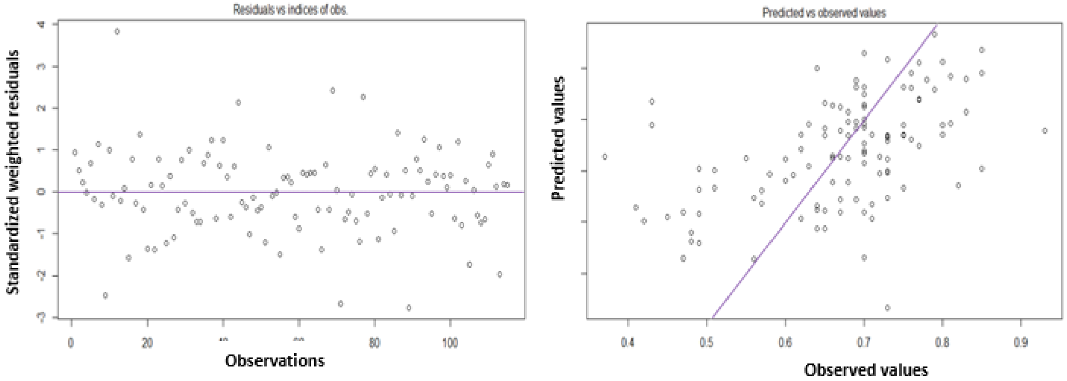


Figure 67. Study D full model evaluation plots.

Note. The residual plot is shown on the left and the predicted vs. observed values on the right, with a fitted line based on maximum likelihood.

Combined Studies

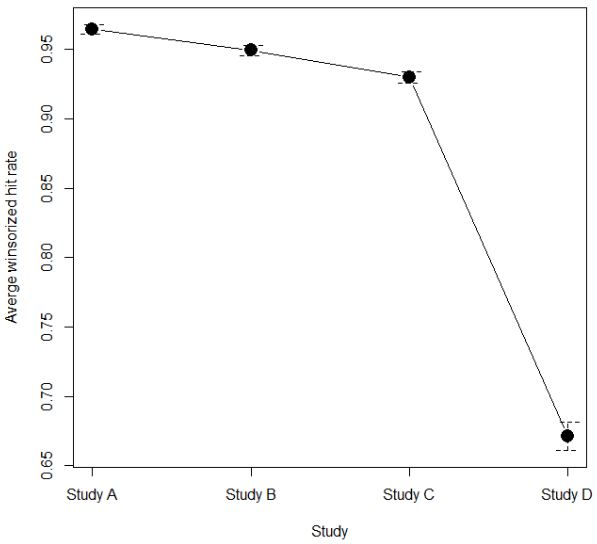


Figure 68. Plot of average winsorized hit rate by study; error bars represent standard error.

Note. Study D was significantly lower in winsorized hit rate than study A, B, and C, Welch' *F*(3, 302.56) = 264.45, *p* < .001.

Dominance Analysis Full Model

Full Model with Suppressors

Threat conspicuity, task duration, and age were held constant, while all potential suppressors and human variables were included.

Table 62

Raw dominance analysis results Full Model with suppressors in Combined Studies.

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration	0.005	0.014	0.001	0.001	0.004	0.008	0.000
k = 3	0.005	0.014	0.001	0.001	0.004	0.008	0.000
Age + Threat Conspicuity + Task Duration + Performance		0.020	0.000	0.002	0.003	0.010	0.000
Age + Threat Conspicuity + Task Duration + Temporal Demand	0.010		0.013	0.000	0.005	0.006	0.000
Age + Threat Conspicuity + Task Duration + Mental Demand	0.004	0.026		0.002	0.003	0.010	0.000
Age + Threat Conspicuity + Task Duration + Gender	0.005	0.013	0.001		0.004	0.007	0.000
Age + Threat Conspicuity + Task Duration + Agent Report Modality	0.004	0.015	0.000	0.002		0.008	0.000
Age + Threat Conspicuity + Task Duration + Video Gaming Experience	0.006	0.011	0.002	0.000	0.004		0.000
Age + Threat Conspicuity + Task Duration + Military Experience	0.005	0.014	0.001	0.002	0.004	0.009	
Conditional dominance k = 4	0.006	0.016	0.003	0.001	0.004	0.008	0.000
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand			0.011	0.001	0.003	0.006	0.000
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand		0.030		0.002	0.003	0.011	0.000
Age + Threat Conspicuity + Task Duration + Performance + Gender		0.018	0.000		0.003	0.008	0.000

	Additional contribution (pseudo R^2) of					
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Performance + Agent Report Modality		0.020	0.000	0.002		0.010	0.000
Age + Threat Conspicuity + Task Duration + Performance + Video Gaming Experience		0.016	0.001	0.000	0.003		0.000
Age + Threat Conspicuity + Task Duration + Performance + Military Experience		0.020	0.000	0.002	0.003	0.010	
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand	0.008			0.000	0.003	0.008	0.001
Age + Threat Conspicuity + Task Duration + Temporal Demand + Gender	0.011		0.013		0.005	0.006	0.000
Age + Threat Conspicuity + Task Duration + Temporal Demand + Agent Report Modality	0.009		0.011	0.000		0.006	0.000
Age + Threat Conspicuity + Task Duration + Temporal Demand + Video Gaming Experience	0.011		0.015	0.001	0.005		0.000
Age + Threat Conspicuity + Task Duration + Temporal Demand + Military Experience	0.011		0.014	0.000	0.005	0.006	
Age + Threat Conspicuity + Task Duration + Mental Demand + Gender	0.004	0.025			0.003	0.008	0.000
Age + Threat Conspicuity + Task Duration + Mental Demand + Agent Report Modality	0.003	0.026		0.002		0.009	0.000

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Mental Demand + Video Gaming Experience	0.005	0.024		0.000	0.003		0.000
Age + Threat Conspicuity + Task Duration + Mental Demand + Military Experience	0.004	0.027		0.002	0.003	0.010	
Age + Threat Conspicuity + Task Duration + Gender + Agent Report Modality	0.004	0.014	0.001			0.007	0.000
Age + Threat Conspicuity + Task Duration + Gender + Video Gaming Experience	0.006	0.011	0.002		0.004		0.000
Age + Threat Conspicuity + Task Duration + Gender + Military Experience	0.005	0.013	0.001		0.004	0.007	
Age + Threat Conspicuity + Task Duration + Agent Report Modality + Video Gaming Experience	0.005	0.012	0.001	0.000			0.000
Age + Threat Conspicuity + Task Duration + Agent Report Modality + Military Experience	0.004	0.015	0.000	0.002		0.009	
Age + Threat Conspicuity + Task Duration + Video Gaming Experience + Military Experience	0.006	0.011	0.002	0.000	0.004		
Conditional dominance k = 5	0.006	0.019	0.005	0.001	0.004	0.008	0.000
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand				0.001	0.002	0.008	0.001

	Additional contribution (pseudo R^2) of					
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Gender			0.011		0.003	0.006	0.000
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Agent Report Modality			0.009	0.001		0.006	0.000
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Video Gaming Experience			0.012	0.000	0.003		0.001
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Military Experience			0.011	0.001	0.003	0.007	
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Gender		0.029			0.003	0.008	0.000
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Agent Report Modality		0.030		0.002		0.010	0.000
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Video Gaming Experience		0.028		0.000	0.002		0.000
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Military Experience		0.031		0.002	0.003	0.011	

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Performance + Gender + Agent Report Modality		0.018	0.000			0.007	0.000
Age + Threat Conspicuity + Task Duration + Performance + Gender + Video Gaming Experience		0.017	0.001		0.003		0.000
Age + Threat Conspicuity + Task Duration + Performance + Gender + Military Experience		0.018	0.000		0.003	0.008	
Age + Threat Conspicuity + Task Duration + Performance + Agent Report Modality + Video Gaming Experience		0.017	0.001	0.000			0.000
Age + Threat Conspicuity + Task Duration + Performance + Agent Report Modality + Military Experience		0.021	0.000	0.002		0.010	
Age + Threat Conspicuity + Task Duration + Performance + Video Gaming Experience + Military Experience		0.017	0.001	0.000	0.003		
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Gender	0.008				0.003	0.008	0.001
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Agent Report Modality	0.007			0.000		0.007	0.001

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Video Gaming Experience	0.009			0.001	0.003		0.001
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Military Experience	0.008			0.000	0.003	0.008	
Age + Threat Conspicuity + Task Duration + Temporal Demand + Gender + Agent Report Modality	0.009		0.011			0.006	0.000
Age + Threat Conspicuity + Task Duration + Temporal Demand + Gender + Video Gaming Experience	0.011		0.015		0.005		0.000
Age + Threat Conspicuity + Task Duration + Temporal Demand + Gender + Military Experience	0.011		0.014		0.005	0.006	
Age + Threat Conspicuity + Task Duration + Temporal Demand + Agent Report Modality + Video Gaming Experience	0.010		0.013	0.000			0.001
Age + Threat Conspicuity + Task Duration + Temporal Demand + Agent Report Modality + Military Experience	0.009		0.012	0.000		0.006	
Age + Threat Conspicuity + Task Duration + Temporal Demand + Video Gaming Experience + Military Experience	0.011		0.016	0.000	0.005		

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Mental Demand + Gender + Agent Report Modality	0.004	0.024				0.008	0.000
Age + Threat Conspicuity + Task Duration + Mental Demand + Gender + Video Gaming Experience	0.005	0.025			0.003		0.000
Age + Threat Conspicuity + Task Duration + Mental Demand + Gender + Military Experience	0.004	0.026			0.004	0.008	
Age + Threat Conspicuity + Task Duration + Mental Demand + Agent Report Modality + Video Gaming Experience	0.004	0.024		0.000			0.000
Age + Threat Conspicuity + Task Duration + Mental Demand + Agent Report Modality + Military Experience	0.003	0.027		0.002		0.010	
Age + Threat Conspicuity + Task Duration + Mental Demand + Video Gaming Experience + Military Experience	0.005	0.025		0.000	0.003		
Age + Threat Conspicuity + Task Duration + Gender + Agent Report Modality + Video Gaming Experience	0.005	0.012	0.001				0.000
Age + Threat Conspicuity + Task Duration + Gender + Agent Report Modality + Military Experience	0.004	0.014	0.001			0.007	

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Gender + Video Gaming Experience + Military Experience	0.006	0.012	0.002		0.004		
Age + Threat Conspicuity + Task Duration + Agent Report Modality + Video Gaming Experience + Military Experience	0.005	0.012	0.001	0.000			
Conditional dominance k = 6	0.007	0.021	0.007	0.001	0.003	0.008	0.000
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Gender					0.002	0.008	0.001
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Agent Report Modality				0.001		0.008	0.001
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Video Gaming Experience				0.001	0.002		0.002
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Military Experience				0.001	0.002	0.009	
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Gender + Agent Report Modality			0.009			0.006	0.001

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Gender + Video Gaming Experience			0.013		0.003		0.001
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Gender + Military Experience			0.012		0.004	0.006	
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Agent Report Modality + Video Gaming Experience			0.011	0.000			0.001
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Agent Report Modality + Military Experience			0.010	0.001		0.007	
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Video Gaming Experience + Military Experience			0.014	0.000	0.003		
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Gender + Agent Report Modality		0.028				0.008	0.000
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Gender + Video Gaming Experience		0.029			0.002		0.000

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Gender + Military Experience		0.030			0.003	0.008	
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Agent Report Modality + Video Gaming Experience		0.027		0.000			0.000
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Agent Report Modality + Military Experience		0.031		0.003		0.011	
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Video Gaming Experience + Military Experience		0.029		0.000	0.003		
Age + Threat Conspicuity + Task Duration + Performance + Gender + Agent Report Modality + Video Gaming Experience		0.017	0.001				0.000
Age + Threat Conspicuity + Task Duration + Performance + Gender + Agent Report Modality + Military Experience		0.019	0.000			0.008	
Age + Threat Conspicuity + Task Duration + Performance + Gender + Video Gaming Experience + Military Experience		0.017	0.001		0.003		

	Additional contribution (pseudo R^2) of					
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Performance + Agent Report Modality + Video Gaming Experience + Military Experience		0.017	0.001	0.000			
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Gender + Agent Report Modality	0.007					0.008	0.001
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Gender + Video Gaming Experience	0.008				0.003		0.001
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Gender + Military Experience	0.009				0.003	0.008	
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Agent Report Modality + Video Gaming Experience	0.008			0.001			0.002
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Agent Report Modality + Military Experience	0.007			0.001		0.008	
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Video Gaming Experience + Military Experience	0.009			0.001	0.003		

	Additional contribution (pseudo R^2) of					
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Temporal Demand + Gender + Agent Report Modality + Video Gaming Experience	0.010		0.014				0.001
Age + Threat Conspicuity + Task Duration + Temporal Demand + Gender + Agent Report Modality + Military Experience	0.009		0.012			0.006	
Age + Threat Conspicuity + Task Duration + Temporal Demand + Gender + Video Gaming Experience + Military Experience	0.011		0.016		0.005		
Age + Threat Conspicuity + Task Duration + Temporal Demand + Agent Report Modality + Video Gaming Experience + Military Experience	0.010		0.014	0.000			
Age + Threat Conspicuity + Task Duration + Mental Demand + Gender + Agent Report Modality + Video Gaming Experience	0.004	0.025					0.000
Age + Threat Conspicuity + Task Duration + Mental Demand + Gender + Agent Report Modality + Military Experience	0.004	0.025				0.008	
Age + Threat Conspicuity + Task Duration + Mental Demand + Gender + Video Gaming Experience + Military Experience	0.005	0.026			0.003		

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Mental Demand + Agent Report Modality + Video Gaming Experience + Military Experience	0.004	0.025		0.000			
Age + Threat Conspicuity + Task Duration + Gender + Agent Report Modality + Video Gaming Experience + Military Experience	0.005	0.013	0.001				
Conditional dominance k = 7	0.007	0.024	0.009	0.001	0.003	0.008	0.001
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Gender + Agent Report Modality						0.008	0.001
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Gender + Video Gaming Experience					0.002		0.002
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Gender + Military Experience					0.002	0.009	
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Agent Report Modality + Video Gaming Experience				0.001			0.002

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Agent Report Modality + Military Experience				0.001		0.009	
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Video Gaming Experience + Military Experience				0.000	0.002		
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Gender + Agent Report Modality + Video Gaming Experience			0.011				0.001
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Gender + Agent Report Modality + Military Experience			0.010			0.006	
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Gender + Video Gaming Experience + Military Experience			0.014		0.003		
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Agent Report Modality + Video Gaming Experience + Military Experience			0.012	0.000			

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Gender + Agent Report Modality + Video Gaming Experience		0.028					0.000
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Gender + Agent Report Modality + Military Experience		0.029				0.008	
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Gender + Video Gaming Experience + Military Experience		0.030			0.003		
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Agent Report Modality + Video Gaming Experience + Military Experience		0.029		0.000			
Age + Threat Conspicuity + Task Duration + Performance + Gender + Agent Report Modality + Video Gaming Experience + Military Experience		0.017	0.001				
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Gender + Agent Report Modality + Video Gaming Experience	0.008						0.001

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Gender + Agent Report Modality + Military Experience	0.008					0.008	
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Gender + Video Gaming Experience + Military Experience	0.009				0.003		
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Agent Report Modality + Video Gaming Experience + Military Experience	0.008			0.001			
Age + Threat Conspicuity + Task Duration + Temporal Demand + Gender + Agent Report Modality + Video Gaming Experience + Military Experience	0.010		0.015				
Age + Threat Conspicuity + Task Duration + Mental Demand + Gender + Agent Report Modality + Video Gaming Experience + Military Experience	0.004	0.026					
Conditional dominance k = 8	0.008	0.026	0.011	0.000	0.003	0.008	0.001
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Gender + Agent Report Modality + Video Gaming Experience							0.002

	Additional contribution (pseudo R^2) of					
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Gender + Agent Report Modality + Military Experience						0.008	
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Gender + Video Gaming Experience + Military Experience					0.002		
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Agent Report Modality + Video Gaming Experience + Military Experience				0.000			
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Gender + Agent Report Modality + Video Gaming Experience + Military Experience			0.012				
Age + Threat Conspicuity + Task Duration + Performance + Mental Demand + Gender + Agent Report Modality + Video Gaming Experience + Military Experience		0.029					

	Additional contribution (pseudo R^2) of						
Subset model X	Performance	Temporal Demand	Mental Demand	Gender	Agent Report Modality	Video Gaming Experience	Military Experience
Age + Threat Conspicuity + Task Duration + Temporal Demand + Mental Demand + Gender + Agent Report Modality + Video Gaming Experience + Military Experience	0.008						
Conditional dominance k = 9	0.008	0.029	0.012	0.000	0.002	0.008	0.002
Age + Threat Conspicuity + Task Duration + Performance + Temporal Demand + Mental Demand + Gender + Agent Report Modality + Video Gaming Experience + Military Experience							
Overall average	0.007	0.021	0.007	0.001	0.003	0.008	0.001

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, keeping threat conspicuity, task duration, and age were held constant (Azen & Budescu, 2003). All potential suppressors and human variables were included. Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

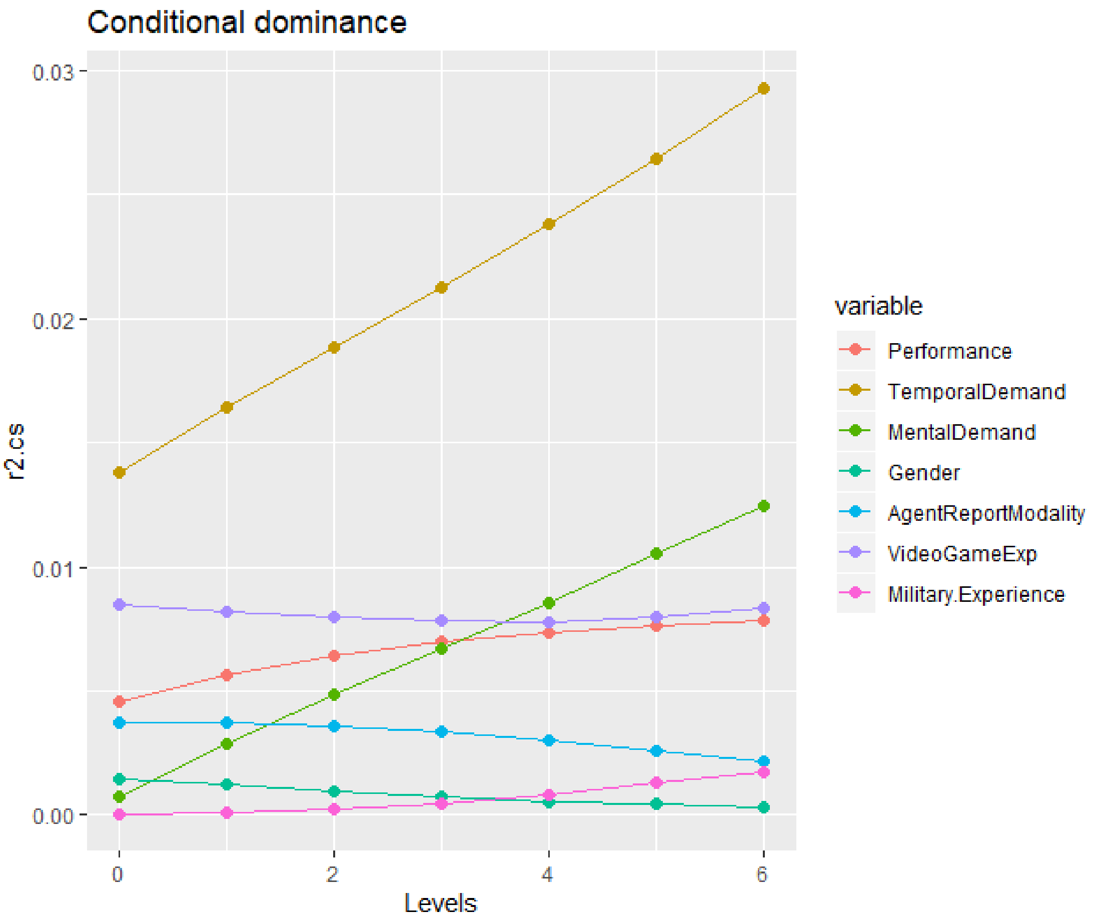


Figure 69. Conditional dominance of full model with suppressors for the combined studies.

Note. The conditional dominance plot identified temporal demand, mental demand, and military experience as suppressors (Azen & Budescu, 2003) in the combined studies. These were dropped from subsequent analyses. Video gaming experience was missing as a variable study D and was therefore also removed from further analyses.

Full Model without Suppressors

Table 63

Raw dominance analysis results of the full model without suppressors in combined Studies.

	Additional contribution (pseudo R^2) of		
Subset model ***X***	Performance	Gender	Agent Report Modality
Age + Threat Conspicuity + Task Duration	0.008	0.005	0.156
Conditional dominance *k* = 3	0.008	0.005	0.156
Age + Threat Conspicuity + Task Duration + Performance		0.005	0.149
Age + Threat Conspicuity + Task Duration + Gender	0.008		0.153
Age + Threat Conspicuity + Task Duration + Agent Report Modality	0.000	0.002	
Conditional dominance *k* = 4	0.004	0.003	0.151
Age + Threat Conspicuity + Task Duration + Performance + Gender			0.146
Age + Threat Conspicuity + Task Duration + Performance + Agent Report Modality		0.002	
Age + Threat Conspicuity + Task Duration + Gender + Agent Report Modality	0.001		
Conditional dominance *k* = 5	0.001	0.002	0.146
Age + Threat Conspicuity + Task Duration + Performance + Gender + Agent Report Modality			
Overall average	0.004	0.003	0.151

Note. This table presents the raw output of the dominance analyses. The unique additional contribution of each predictor is shown over all possible subset model sizes, keeping age, threat conspicuity, and task duration constant (Azen & Budescu, 2003). Conditional dominance indicates the average unique contribution for that subset model size (*k*) for the predictor under evaluation. The overall average presents the average over all average *k* model sizes.

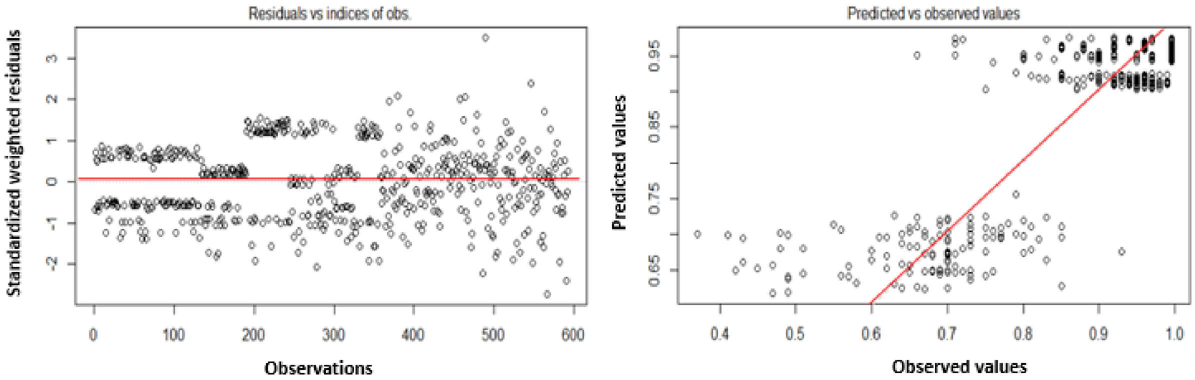


Figure 70. Combined Studies full model evaluation plots

Note. The residual plot is shown on the left and the predicted vs. observed values on the right, with a fitted line based on maximum likelihood

APPENDIX G: IRB FOR BORROWED EXPERIMENTAL STUDY A

APPENDIX G: IRB FOR BORROWED EXPERIMENTAL STUDY A

- **If you believe your activity may not meet the definition of “Human Research” subject to IRB oversight, contact the IRB Office prior to developing your protocol**
- **Be sure that all study materials are correct and consistent with the information in this protocol.**
- **The italicized bullet points below serve as general guidance to investigators on the kinds of information that may be applicable to include in each section. Please DELETE the italicized text in your protocol.**
- **Note that, depending on the nature of your research, some sections below will not be applicable. Indicate this as “N/A.”**
- **For any items described in the sponsor’s protocol or other documents submitted with the application, investigators may simply reference the page numbers of these documents.**

1) Protocol Title

A Novel Mixed Reality Interface For Effective and Efficient Human Robot Interaction with Unique Mobility Platforms

2) Principal Investigator

- Principal Investigator: Daniel J. Barber
- Co-PI: Florian Jentsch

1) Research Assistants: Andrew Watson, Jonathan Harris, Alexis San Javier, Thomas Pring, Christopher Miller, Austin Miller, Austin Carter, Nicholas Wyatt, Sasha Willis, Andrew Talone

3) Objectives

The goal for this experiment as currently defined is to understand how robot type and visual complexity of a mixed reality interface affects cooperative human-robot teaming in dismounted military applications. In order to accomplish we need to:

1. Measure how robot type (wheeled vs. legged) impacts users' expectations of robot capability and performance.
2. Measure how visual complexity (low vs. high) of a mixed reality interface affects primary task performance and situation awareness/working memory recall.

These objectives will be measured by:

- Collecting user feedback regarding the platform type (legged vs. wheeled) and presentation of information conveyed from a robot teammate through reports at different levels of mixed reality interface visual complexity (low vs. high).
- Collecting information regarding a user's ability to interpret robot communication data from multimodal reports.
- Collecting information regarding a user's ability to recall and recognize information during exchanges within a human-robot team.
- Collecting information regarding a user's workload while interacting with different platform types and levels of mixed reality interface visual complexity within a human-robot team.

- Collecting information regarding impact on a user’s situation awareness while interacting with different platform types and levels of mixed reality interface visual complexity within a human-robot team.
- Collecting information regarding a user’s usability preferences while interacting with different platform types and levels of mixed reality interface visual complexity within a human-robot team.
- Assessing the performance costs associated with different platform types and levels of mixed reality interface visual complexity within a human-robot team.

4) Background

Mixed Reality

Extensive research is required to develop a viable mixed reality visual display for human-robot collaboration, particularly with a focus on grounding, situation awareness, common and shared reference frames and spatial referencing [1]. This is especially true for dismounted military applications.

- Prior research for dismounted military applications has focused on a multimodal interface (MMI) running on a mobile device (e.g. a tablet) [2, 3]. Furthermore, research focused on head-mounted displays mostly focused on 2D augmentation [4]. Few studies have focused on 3D augmentation (also referred to as mixed reality) interfaces for dismounted military.

Visual Complexity

- Extensive research has been done on information complexity for visual displays for Air Traffic Controllers (ATCs) [5, 6, 7]. The guidelines, metrics and questionnaires for the ATC domain will be adapted for human-robot interaction in dismounted military applications.

- Extensive research has been done on display clutter for Heads-Up Displays (HUDs) for airplane pilots. As seen with the Air Traffic Controller research, displays with enhanced information provide pilots with information previously unavailable with traditional flight instrumentation; however, the display of additional information may result in display clutter and therefore inhibiting the processes and tasks they are designed to support. Furthermore, it was found moderate levels of clutter may be acceptable if the information is relevant to the task at hand [8].

- Moacdieh et al. studied the performance and attentional costs with Primary Flight Display (PFD) clutter. Using a flight simulator, the authors created low-, medium- and high-clutter PFDs for which pilots flew a simulated flight scenario containing intervals of high and low workload. The pilots were required to detect visual alerts and notifications that appeared on the PFD. Using eye tracking, performance and subjective measures, it was concluded that clutter significantly increased response time to alerts and a high workload resulted in more alerts being missed [9]. Our research will build upon this research and apply it to the domain of human-robot interaction in dismounted military applications.

- Ling et al. argue that visual complexity is found to be negatively correlated with usability and positively correlated with mental workload [10].

Robot Type

- Robots still lack the capabilities to dynamically interact with human team members. Abich et al. developed a simulation to overcome current limitations of robot platforms and focused on the development and assessment of communication functionality [11]. Legged robots currently lack the intelligence and capabilities to be a part of a dynamic human-robot team. An experimental environment is needed to understand the communication and interface requirements for humans interacting with unique mobility platforms.

- Research has shown that legged robots are anthropomorphized much more than wheeled robots. However, few studies have focused on how anthropomorphism can be utilized to create affective robot behavior needed for collaboration with humans in complex environments [12].

References

1. S. A. Green, M. Billinghurst, X. Chen, and J. G. Chase, ”Human-Robot Collaboration: A Literature Review and Augmented Reality Approach in Design," in *International Journal of Advanced Robotic Systems*, Vol. 5, No. 1, 2008.
2. J. Abich, D. J. Barber, and L. R. Elliot, ”An Initial Investigation of Exogenous Orienting Visual Display Cues for Dismounted Human-Robot Communication," in *Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems*, Florida, USA, July 27-31, 2016.
3. D. J. Barber, T. M. Howard, and Mr. R. Walter, ”A Multimodal Interface for Real-Time Soldier-Robot Teaming," *SPIE Proceedings Vo. 9837, Unmanned Systems Technology, XVIII,* May 13, 2016.
4. R. Kopinsky, A. Sharma, N. Gupta, C. Ordonez, E. G. Collins Jr., and D. Barber, "Human Guidance of Mobile Robots in Complex 3d Environments Using Smart Glasses." *Proc. SPIE 9837, Unmanned Systems Technology XVIII, 983709.*, May 2016
5. X. Manning, ”Complexity and Automation Displays of Air Traffic Control: Literature Review and Analysis," *Technical Report for US DOT and US FAA,* April, 2005.
6. X. Manning, ”Designing Questionnaires for Controlling and Managing Information Complexity in Visual Displays," *Technical Report for US DOT and US FAA*, August, 2008.
7. J. Xing, ” Information Complexity in Air Traffic Control Displays," *Technical Report for US FAA*, September, 2007.
8. A. Alexander, E. Stelzer, S. Kim, D. Kaber, ”Bottom-up and Top-down Contributors to Pilot Perceptions of Display Clutter in Advanced Flight Deck Technologies.” *Proceedings of the Human Factors And Ergonomics Society – 52nd Annual Meeting*, 2008
9. N. Moacdieh, J. Prinet, N. Sarter, “Effects of Modern Primary Flight Display Clutter: Evidence from Performance and Eye Tracking Data,” *Proceedings of the Human Factors and Ergonomics Society – 57th Annual Meeting*, 2013
10. C. Ling, M. Lopez, and R. Shehab, ”Complexity Questionnaires of Visual Displays: A Validation Study of Two Information Complexity Questionnaires of Visual Displays," *in Human Factors and Ergonomics in Manufacturing & Service Industries*, Vol. 25, No. 5, pp. 391-411, 2013.
11. J. Abich, D. J. Barber, and L. Reinerman-Jones, ”Experimental Environments for Dismounted Human-Robot Multimodal Communications," *Proceedings of 7th International Conference, VAMR 2015, Held as Part of HCI International 2015*, pp., Los Angeles, CA, USA, August 2-7, 2015.

12. V. K. Sims et al, ”Anthropomorphism of Robotic Forms: A Response to Affordances," *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*, 2015.

Hypotheses for this study include:

- H_1: Participants will exhibit higher recall accuracy for high visual complexity (HVC) scenarios (more information displayed on screen).
- H_2: Participants will exhibit shorter recall response times for HVC scenarios.
- H_3: Participants will report differences in workload (e.g. NASA-TLX, HRV) between HVC and low visual complexity (LVC).
- H_4: Participants will perform better (i.e. accuracy, response time) on primary signal detection task (SDT) for HVC scenarios (less information to memorize, better focus on SDT).
- H_5: Participants will perform better (i.e. accuracy, response time) on recall of robot reports in HVC scenarios.
- H_6: Participants will report differences in robot expectations and trust (REPI and trust score) between wheeled and legged robot type.

The results of this research will help us to develop guidelines on how to identify the appropriate visual complexity for a mixed reality interface for dismounted military applications. Furthermore, it will help us understand how robot platform type affects human-robot team collaborations.

5) Setting of the Human Research

Experimentation will be conducted at UCF and will use the UCF population. This experiment will involve participants performing tasks in a simulated environment and answering questionnaires. The experiment will be conducted at the Institute for Simulation and Training’s Partnershiup II building room 112.

6) Resources available to conduct the Human Research

- This project is funded by the RCTA FY2016 Task H7: HRI of Unique Mobility Platforms. Research staffing, testing equipment and testbeds are provided by the University of Central Florida's Institute for Simulation and Training.

- A power analysis was conducted a priori and determined that an appropriate sample size of 90 would be adequate to detect moderate effects at $\alpha = .05$, $\beta = .05$ for a mixed repeated-measures design.
- We estimate that the time period for this study will be six to nine months. This includes data collection and coding.
- Each of the research staff who will be interacting with participants has research experience that includes data collection, facilitating studies, recruiting participants, and analyzing data.
- All of the current staff has received CITI training.
- We anticipate that all measures and stimuli can be collected either on-line via Sona Systems/ UCF Qualtrics, or in the laboratory setting.

7) Study Design

NOTE: Researchers developing multi-faceted protocols (e.g., multiple phases, study groups, research components, etc.) may want to develop separate "Study Design" sections for each component of their research rather than trying to combine disparate components into a single section.

a) Recruitment Methods

i) Participants will be recruited from the general psychology and IST research pool using Sona Systems. Participants will receive course credit for their participation that can be used for a qualifying undergraduate psychology course.

ii) Researchers will not specifically identify or contact potential research participants. Rather, the study will be listed as available to be participated in, via UCF's SONA Systems. Our study will only be visible via SONA systems to potential participants who identify themselves to SONA Systems as being at least 18 years of age. Potential participants who meet this qualification will then be able to view our study as an available option for them to participate.

iii) If students are unable to participate in our study for reasons such as age, or if they do not wish to take part in our study for other personal reasons, the students will have the opportunity to arrange with their course professors an alternate assignment that will allow them to acquire the necessary course credit needed.

iv) No advertisements or other materials will be used to recruit study participants.

v) We anticipate needing approximately 90 participants to complete this study.

b) Participant Compensation

1. Participants will be offered course credit for their participation.
2. This research will conform to UCF Psychology Department's and IST's policy for granting course credit in return for research participation.The policy specifically states:
   All face-to-face studies are worth twice as much as online studies. Face-to-face studies must be credited at the rate of 0.5 credits per 30 minutes (rounded up) and online studies must be credited at the rate of 0.25 credits per 30 minutes (rounded up). Thus, if your face-to-face study takes approximately 20 minutes to complete, your study should be set up to award 0.5 points to each participant. If the face-to-face study takes 40 minutes to complete, the study should be set up to award participants 1 point. Likewise, a 20 minute online study would be worth 0.25 points and a 40 minute online study would be worth 0.50 points
3. If students are unable to participate in our study for reasons such as age, or if they do not wish to take part in our study for other personal reasons, the students will have the opportunity to arrange with their course professors an alternate assignment that will allow them to acquire the necessary course credit needed.

c) Inclusion and Exclusion Criteria

Participants involved in this study will be students who are enrolled in an undergraduate and graduate classes at the University of Central Florida and are over the age of 18. Participants will have to demonstrate eligibility (class registration) by signing up for Sona Systems and completing a pre-screening measure provided by Sona Systems (age). This pre-screening measure will screen students for age such that only students who are 18 years old and above, have normal or corrected to normal vision, and an ability to stand/walk without assistance will be able to sign up to participate in our study. Participants with a previous history of seizures will be excluded. This will be screened for as part of the pre-screening provided by SONA, and asked directly while providing consent to participate. Researchers will not attempt to recruit persons identified as being part of a vulnerable population (e.g., children, prisoners, mentally disabled persons).

d) Study Endpoints

NOTE: This section is only required for biomedical research. It is generally not applicable to social or behavioral research.

- *N/A*

e) Study Timelines

- Anticipated time to complete the study is approximately 180 minutes.
- The researchers anticipate that we will need approximately 6 to 9 months to complete data collection, data coding, and preliminary analyses.

f) Procedures involved in the Human Research.

- Deception will not be used in this study.
- No audio or video recording of this research or research participants will be conducted without participant consent. Participants who do not agree to audio recording will be able to participate in the study. No video is recorded for this study.
- The foreseeable risk to participants is minimal to none; therefore procedures to minimize magnitude of risk will not be taken.
- However, there may be concern that a military scenario or the suggestion of a robotic teammate may invoke a negative response to those sensitive to issues associated with military conflict, police investigation, crime, or artificial intelligence.
- Participants will be allowed to withdraw from the study at any time should they feel it necessary. Further, they will be credited for the amount of time that they took part in the study prior to choosing to withdraw.
- No source records will be used.
- No long term follow-up data will be collected.
- No medical records will be used.
- A 2 x (Robot Type: Wheeled - W, Legged - L) x 2 (Visual Complexity: Low - LVC, High - HVC) mixed design with repeated measures for Robot Type will be used to identify the appropriate Visual Complexity to maintain performance on the Signal Detection Task (SDT - insurgent identification) and Information Reporting Task (IRT - working memory, information recall) and understand how Robot affects the user's perception and expectation for the robot and ultimately human-robot team performance.

Questionnaires:

- *Biographical Data questionnaire*. A software generated questionnaire gathers background information regarding age, gender, visual acuity, academic education, military experience, computer use, video game exposure/experience, and robotics knowledge.
- *Ishihara color deficiency test.* This consists of a number of colored plates containing a circle of dots randomized in color and size. Within the randomized pattern on each plate are dots that form a number visible to those with normal color vision and invisible, or difficult to see, for those with a red-green color vision defect.

- *Spatial orientation survey.* This spatial orientation test (adapted from Thurstone's Cubes) assesses the ability to mentally rotate and compare objects in space (Ekstrom, French, & Harman, 1979).
- *Verbal-Spatial Ability Rating (VSAR):* Self-report measure on two items that asks participants to rate their verbal and spatial ability separately.
- *Verbal-Visual Learning Style Rating (VVLSR):* Self-report measure on a single item using a 7-point scale asks participants to rate the degree to which they are more verbal or visual learners.
- *Santa Barbara Learning Style Questionnaire (SBLSQ):* Self-report measure on six items using a 7-point scale asks participants to rate the degree to which they are more verbal or visual learners.
- *Reading Span (RSPAN):* This software generated working memory task requires participants to read aloud sentences, each of which are followed by an upper case letter (Kane et al., 2004). Participants must recall the letters in correct serial order after a set of sentence-letter strings.
- *Trust between people and automation questionnaire.* This 12-item checklist is a self-report measure of human trust in automation created by Jian, Bisantz, and Drury's (2000).

- *NASA-Task Load Index (TLX).* The TLX is a multi-dimensional scale comprised of six subscales with three focusing on demand imposed on the participant (mental, physical, and temporal demand) and three on the interactions with the task or system (effort, frustration, and performance level; Hart & Staveland, 1988).

- *System usability survey (SUS).* This 10-item questionnaire focused on perceived usability of the system (i.e. hardware, software, equipment; Brooke, 1996).

- *Ratings of Expectation and Perceived Importance (REPI).* This 17-item questionnaire focused on perceptions of the user's expectations and perceived importance of the robot's behavior and functionality before and after interaction (Lohse, 2011).

- *Perceived awareness of the research hypothesis (PARH).* This scale is a quick and convenient quantitative method for measuring the potential influence of demand characteristics in psychology research situations (Rubin, Paolini, & Crisp, 2010).

- *Interaction Reflection.* Items of this measure cover positive and negative aspects of their interaction with the device, and ask to provide any suggestions for improvement.

- *Simulator Sickness Questionnaire (SSQ).* Beginning, mid-point, and end of experiment. Given at set time-intervals during the experiment (Kennedy, Lane, Berbaum, & Lilienthal, 1993).

Hardware:

- *Physiological assessment*.
    - The Microsoft Band 2, non-invasive, low-cost consumer-grade, wearable wristband monitors cardiac activity. Measures of heart-rate (HR), heart-rate variability (HRV), inerbeat-interval (IBI), and galvanic skin response (GSR), and skin temperature will be collected.
    - The Empatica E4, non-invasive, research-grade wearable wristband monitors cardiact activity. Measures of heart-rate (HR), heart-rate variability (HRV), interbeat-interval (IBI) and galvanic skin response (GSR), and skin temperature will be collected.
- *Virtual reality headset*. The HTC Vive will be used to display the virtual environment used within the simulation and emulate a mixed-reality heads-up display.

Procedure:

Upon arrival, participants will be assigned to a group for a corresponding visual complexity level (low or high).

Phase 1: Biographical Data

Participants will be asked to fill out the following measures:

- Biographical Data questionnaire
- RSPAN
- VSAR, VVLSR, SBLSQ
- Spatial orientation survey
- Trust between people and automation questionnaire (Pre-test)
- Ratings of Expectation and Perceived Importance (REPI; Pre-test)
- SSQ (baseline)

Phase 2: Training

Participants will then be asked to view a PowerPoint presentation that will familiarize them with the tasks they will be asked to perform and the subsequent practice exercises. Participants will be asked to complete the following training presentations and practice exercises.

- *Background information.* Participants will be given background information to provide a context for the given scenarios. The backstory will be validated by subject matter experts (SMEs) to ensure contextual credibility.
- *PowerPoint training on signal detection task.* Participants will be asked to view a PowerPoint presentation for training on the signal detection task which will include the identification of threat items in a simulated environment. Threat items will include models of potential improvised explosive devices (IEDs), weapons cache, as well as models of potential insurgents and enemy forces. The training will include which items are classified as threat items and how to identify them in the simulation.
- *Signal detection practice exercise.* After completing the training on the signal detection task, the participant will be asked to complete a practice signal detection task.

- *PowerPoint training on robot reporting.* Participants will then be asked to view a PowerPoint presentation for training on when and what type of information the robot will provide in the robot reports, how to access reports, how those reports will be displayed, and how to respond to related questions during the scenarios.
- *Robot reporting practice exercise*. Participants will then be asked to complete practice trials including the robot reports.
- *Questionnaire exposure*. Throughout the training, participants will also be given information describing what the questionnaires are, how they will look, and how to respond to them. They will be given an opportunity to practice answering the questionnaires.
- *Physiological assessment*. Participants will wear the Microsoft Band 2 and the Empatica E4 on their wrists (one on left, one on right) for the duration of the experiment. It will be explained to them what the Band 2 and E4 are and what information they collect.
- *Combined practice exercise*. Participants will then be asked to complete two practice trials one for the legged and another for the wheeled platform that includes robot reporting and signal detection tasks.

Phase 3: Experimental Scenarios

After completing all of the training materials and the training exercises, participants will be asked to complete two experimental scenarios. The scenarios will vary in robot type and visual complexity depending on assigned group.

Scenarios will be presented using a custom 3D Virtual Reality simulation testbed that emulates the operational area of a dismounted Soldier. The simulation will be a completed using suite of gaming tools available for customization to meet investigational needs.

Scenarios will take approximately 15 minutes each to complete. In general, during the scenario, participants will be playing the role of a human teammate in a simulated Soldier-robot team and will be at a fixed location searching for target items while responding to communicated messages from a robot. The simulated robot will be performing a similar task. The participant will be responsible to recall/recognize information from the robots communicated messages.

- ▪ Scenario :

    - Simulation: In a simulated environment, participants will take on the role of a team leader within a dismounted squad performing a cordon and search operation. The view will be from the first-person perspective of the Soldier. The Soldier will automatically be placed in the proper orientation, distance, and viewing angle to perform the signal detection task. Each scenario building location will be subject to a cordon and search operation.
    - Signal Detection (SD) task: The event rate will be 15 events/min with a 13.33% probability of a signal present. Based on previous research this should correspond to a low task level. An event will be the presence of a person (both enemies and friendlies) that is entering, exiting, or approaching the cordoned area. A signal is the presence of an enemy. The participant will identify and select enemies using the HTC Vive controllers.
- Conditions:
    - There will be two groups of participants.
        - Group 1 (G1) will experience both Robot Types (Wheeled, Legged) for Low Visual Complexity.
        - Group 2 (G2) will experience both Robot Types (Wheeled, Legged) for High Visual Complexity.
    - Wheeled Robot + Low Visual Complexity for G1:
        - One wheeled robot (part of the hit team) will send reports via audio (i.e. synthesized speech radio message) and visual (i.e. virtual text box) communicated simultaneously. In addition, a Basic Marker (i.e. symbol and location in 3D space) will be placed at the location of the report. This marker will remain in the scene until the end of the scenario.
        - The reports will be initiated automatically. The messages will contain information regarding distance, direction, and description (i.e. 3 D’s) of threats, IEDs, weapons cache, hostages, or currency bins outside of and within the building (i.e. out of line of sight).

        - The visual reports will contain the same information that is conveyed through auditory reports but in text format. In addition, the Basic Marker will also show the description of symbol and spatial location in the scene. The participant must later recall this information and provide a report back to the squad leader which will come in the form of questions that are prompted on the screen at varying intervals (e.g. after receiving varying number of reports from the robots). The questions will be in regards to the 3 D's and priority intelligence requirement (PIR) reports. Participants will verbally respond to questions that will be collected using an automatic speech recognizer (ASR). This task will provide a measure of situation awareness (SA) and level of recall.
    - Legged Robot + Low Visual Complexity for G1:
        - This scenario will be the same as above except the robot will be a legged robot.
    - Wheeled Robot + High Visual Complexity for G2:
        - This scenario is the same as Wheeled Robot + Low Visual Complexity for G1 but will contain more visual display elements (high complexity).
        - There will be a 2D top-down minimap in the bottom left corner of the visual display that shows the soldiers, robot and markers.
        - Instead of Basic Markers, Enhanced Markers will be displayed. Enhanced Markers display a symbol, the quantity, the location (direction or floor inside a building) and spatial location within the scene.
    - Legged Robot + High Visual Complexity for G2:
        - This scenario will be the same as above except the robot will be a legged robot.
- During and after completing each scenario, participants will remove the HTC Vive heads and asked to complete the NASA-TLX, SUS, Automation Trust, REPI measure, and SSQ on a standard desktop computer.

- Lastly, to account for potential extraneous effects of the presentation order of experimental scenarios, they will be counterbalanced across participants. There will 2 scenes which will place the participant at different locations/viewpoints in the 3D virtual environment. The scenes will also be counterbalanced across participants.

Phase 4: Post scenario questionnaires

Upon completion of all experimental scenarios and associated questionnaires, participants will be asked to complete

- Free-response questionnaire
- PARH
- Participants will then be provided with the post participation information form and the optional researcher evaluation form.

g) Data and specimen management

NOTE: Data confidentiality issues are a separate topic that is addressed in section 11 below.

- See procedures and provisions sections.
- No data will be sent out or received
- No specimens or data will be transported.
- All survey material identification shall be done through a participant id number that cannot be traced back to the participants. In addition, participants will sign up for the study using a Sona ID number that is only known to the participant. This is done to avoid any member of the research team accidentally finding out the identity of the research participants when they grant participation credit to participants via Sona systems. Through this ID number system, researchers granting credit to research participants cannot identify participants or potential participants via their name. Only de-identifiable summary results (e.g., mean ages, age ranges, number of males and females) have the potential to be published in technical research reports.
- All the sub and co-investigators are responsible for collecting and preserving data. Data will be kept for a period of five years and secured in a locked file cabinet that is compliant with human participant's research. Digital recorded data (e.g. audio recording, simulation logs) will be stored indefinitely in a secured network drive in which folder access will be restricted to those listed and approved in this protocol.

- Data shall be managed carefully by monitoring each of the survey items to ensure that they are filled out completely and that the survey items for each participant are combined together. If participants chose not to respond to items, researchers will determine whether certain items are systematically unanswered by study participants and consider removing those items. Participants will not be penalized for choosing not to respond to a question/item.
- Researchers will carefully monitor the data to determine if certain items are systematically unanswered by participants. As this situation could be a case of having “bad items” included in our item pool, we will work to ensure that these items receive additional scrutiny and are removed as necessary.
- Further, if participants are found to be malingering or “Christmas Treeing” items, our research team will take the following steps:
  - Politely tell participants, “It is very important that you try your best during the experiment. If you feel that you cannot give your full effort, I will have to end the experiment early.” Participants will be granted credit for all of the time that they participated in the study.
  - The researchers will have the right to ask participants to withdraw from the study if they are disrupting the participation of other participants, being disrespectful to other participants, the research staff, or research equipment, or engage in conduct that is not compliant with the University’s Golden Rule policy. In the event that participants are asked to withdraw, they will be granted credit for all of the time that they participated in the study.
- Data analysis plan will include but is not limited to the use of correlation, regression, and ANOVA statistical techniques as well as analyzing data for mean trends or otherwise useful patterns. The independent and dependent variables are listed below in Table 1.

Independent	**Dependent**
Biographical data • Gender/Sex • Video game experience • Virtual reality experience • Computer usage • Multilingual • Military experience (e.g. rank, deployment, time in service, etc.) • Education level • Robotics Experience	Correlated with • Mental workload (TLX score) • Usability preference (SUS rating) • Working memory (recall probe score) • Situation awareness (SA probe scores) • Robot expectations and perceived importance (REPI score) • Trust automation (trust score)
Visual Complexity • Low • High **Platform Type** • Legged • Wheeled	Effects on: • Mental workload (TLX score) • Usability preference (SUS rating) • Working memory (recall probe score) • Situation awareness (SA probe scores) • Robot expectations and perceived importance (REPI score) • Trust robots (trust score) • Physiological response – Microsoft Band 2 & Empatica E4 (HRV, IBI, HR) • Response time (IRT, SDT) • Identification percent accuracy (SDT) • Identification error rate (SDT) • Effects on task performance: o Percent accuracy o Error rate o Response time

h) Provisions to monitor the data for the safety of participants

NOTE: This section is only required when Human Research involves more than minimal risk to participants. It is not applicable to research that is not more than minimal risk.

- No more than minimal risk is anticipated
- The research team will not attempt to recruit participants from vulnerable populations. All volunteers will indicate that they are of legal age (18+ years of age) by answering a prescreening questionnaire via Sona Systems. Our study will not be visible as a participation option to students who do not indicate that they are at least 18 years of age.

i) Withdrawal of participants

- Individuals will be informed that participation in the study is voluntary and that they may withdraw at any time without penalty.
- Researchers believe that the likelihood of participant risk is very low. However, there may be concern that a military scenario or the suggestion of a robotic teammate may invoke a negative response to those sensitive to issues associated with military conflict, police investigation, crime, or artificial intelligence.
- Participants will be allowed to withdraw from the study at any time should they feel it necessary. Further, they will be credited for the amount of time that they took part in the study prior to choosing to withdraw.
- In addition, participants have the right to leave items or measures unanswered if they feel that answering the items or measures is not in their best interest, could cause unforeseen psychological or physical discomfort, or could compromise the confidentiality of their data. Researchers will not force participants to answer survey items or partake in filling out survey measures if they do not chose to do so.
- Participants may be asked to withdraw from the research without their consent in circumstances in which participants are found to be malingering or “Christmas Treeing” items (After being asked to stop this behavior by the researchers), or if the researchers determine that continuing participation is not in the best interest of the participant (e.g., in the event of tornado warning in the building, participant is falling asleep, etc.). Participants may be withdrawn from the study if

they are disrupting the participation of other participants, being disrespectful to other participants, the research staff, or research equipment, or if participants engage in conduct that is not compliant with the University's Golden Rule policy.

- In the event that participants are asked to withdraw, they will be granted credit for all of the time that they participated in the study.

8) Risks to participants

- Researchers believe that the likelihood of participant risk is very low. However, there may be concern that the suggestion of a military scenario or a robotic teammate may invoke a negative response to those sensitive to issues associated with military conflict, police investigation, crime, or artificial intelligence, or participation in research that is funded by the U.S government, Department of Defense, or the U.S. Army.
- Participants will be informed that this research is funded by the U.S. Army on the Informed Consent Form document.
- The Microsoft Band 2 and Empatica E4 physiological sensors used is a commercial wearable product that simply goes on the wrist like a watch. There is no foreseeable risks associated with wearing the sensor. All the equipment is unobtrusive, non-invasive, and has been fully tested and inspected to maintain safety. The researchers performing this study have completed training on the use and safety of each of the pieces of equipment used in the experiment.
- There is a slight risk of participants being affected by simulator sickness using the HTC Vive Virtual Reality headset. However, breaks from interactions with the virtual environment are built into the study design to avoid extended periods of VE interaction and lessen the likelihood of experiencing simulator sickness.

9) Potential direct benefits to participants

- Participants will be immersed in an environment of scholarly research during the duration of participation. This may help to augment their research education.
- No benefits have been promised or are expected to be given to the volunteers who participate in this study. However, the data resulting from this research will be the primary information used to inform designers of robotic systems, specifically in human-robot interaction.

10) Provisions to protect the privacy interests of participants

- Researchers do not foresee privacy interests being comprised by participating, entering into our study, or coming into our research facilities.
- Research facilities are located on the main campus of the University and its adjacent research park. As both facilities are associated with official university business and activities, we do not anticipate privacy interests to be compromised.
- Data in any form will be kept either in a locked cabinet or maintained on a password protected computer with limited access. Only persons listed on the IRB will have access to the information.
- Participant data will not be disseminated outside of the researchers and their immediate assistants. However, summary statistics of participant's de-identifiable data (e.g., mean age, age range, number of male and females) may be reported in technical publications including technical reports and peer reviewed submissions. Again, specific data will be used to inform the development of a follow up study.

11) Provisions to maintain the confidentiality of data

- Individual data will not be revealed to anyone other than the researchers and their immediate assistants.
- Only UCF researchers listed on this protocol will have access to immediate data in paper or electronic form.
- Instead of using names and personal information, data will be identified by assigned numbers participant numbers. Research credit will be granted using a different set of identification numbers determined by Sona Systems. This will ensure that the research team is not able to link participant data with participant names. Thus, the data cannot become identifiable.
- Participant IP addresses will not be available to researchers and will not be sought by researchers.
- Only group means scores and standard deviations, but not individual scores, will be published or reported.
- Data in paper form will be stored in a locked cabinet to which only researchers and immediate assistants will have access for five years. Digital data will be stored in a secured network drive in which folder access will be restricted to those listed and approved in this protocol indefinitely.

12) **Medical care and compensation for injury**

NOTE: This section is not applicable for research that involves no more than minimal risk.

- *N/A*

13) **Cost to participants**

- Participants will not incur any costs for participation

14) **Consent process**

NOTE: The process of obtaining informed consent is distinct from the informed consent document itself.

- Once in the lab, participants will be presented with the Informed Consent form that includes the details of the study, information on the rights of research participants, and contact information for the research team and internal review boards. The informed consent process will be conducted by the research assistants who will be facilitating this study and supervised by the sub investigators (Listed in the Investigators section of this document). After reviewing the form, participants will be given the opportunity to ask for clarification on any of the study details and/or ask questions about the research. Once this opportunity has passed and all questions and concerns have been addressed, participants will be asked if they would like to continue with their participation in the study. Participants will indicate their consent by signing their name on the informed consent form. If they chose not to participate, they will be thanked for their time and instructed to the exit. Informed consent will not be attempted in any language other than English. In accordance with University policy that dictates students demonstrate an adequate level of English language comprehension, researchers will anticipate participants to be able to read and write in English.
- Because this research is funded by the U.S. ARMY, the informed consent process will also comply with U.S. ARMY standards for ethical research. Meaning that, in the event that this research is considered "exempt" by the UCF institutional review board, the researchers will still seek a signed informed consent document, so as to be compliant with both UCF's IRB and the U.S. ARMY's HLAR/AHRPO review process.

15) **Process to document consent in writing**

- Although this study is of minimal risk and may qualify for a waiver of written documentation of consent, in compliance with DOD standards, participants will indicate their consent to participate by signing their name on the Informed Consent form. The research assistant conducting the study will also sign as the person obtaining consent. A copy of this document will be made and given to the participant to keep for their own records. The research team will also keep a copy of this document that will be stored in a secure locked filing cabinet away from other study materials so as to avoid any chance of linking participant names to other study materials.

16) Vulnerable populations

- The research team will not attempt to recruit participants from vulnerable populations. All volunteers will indicate that they are of legal age (18+ years of age) by answering a prescreening questionnaire via Sona Systems. Our study will not be visible as a participation option to students who do not indicate that they are at least 18 years of age.

17) Drugs or Devices

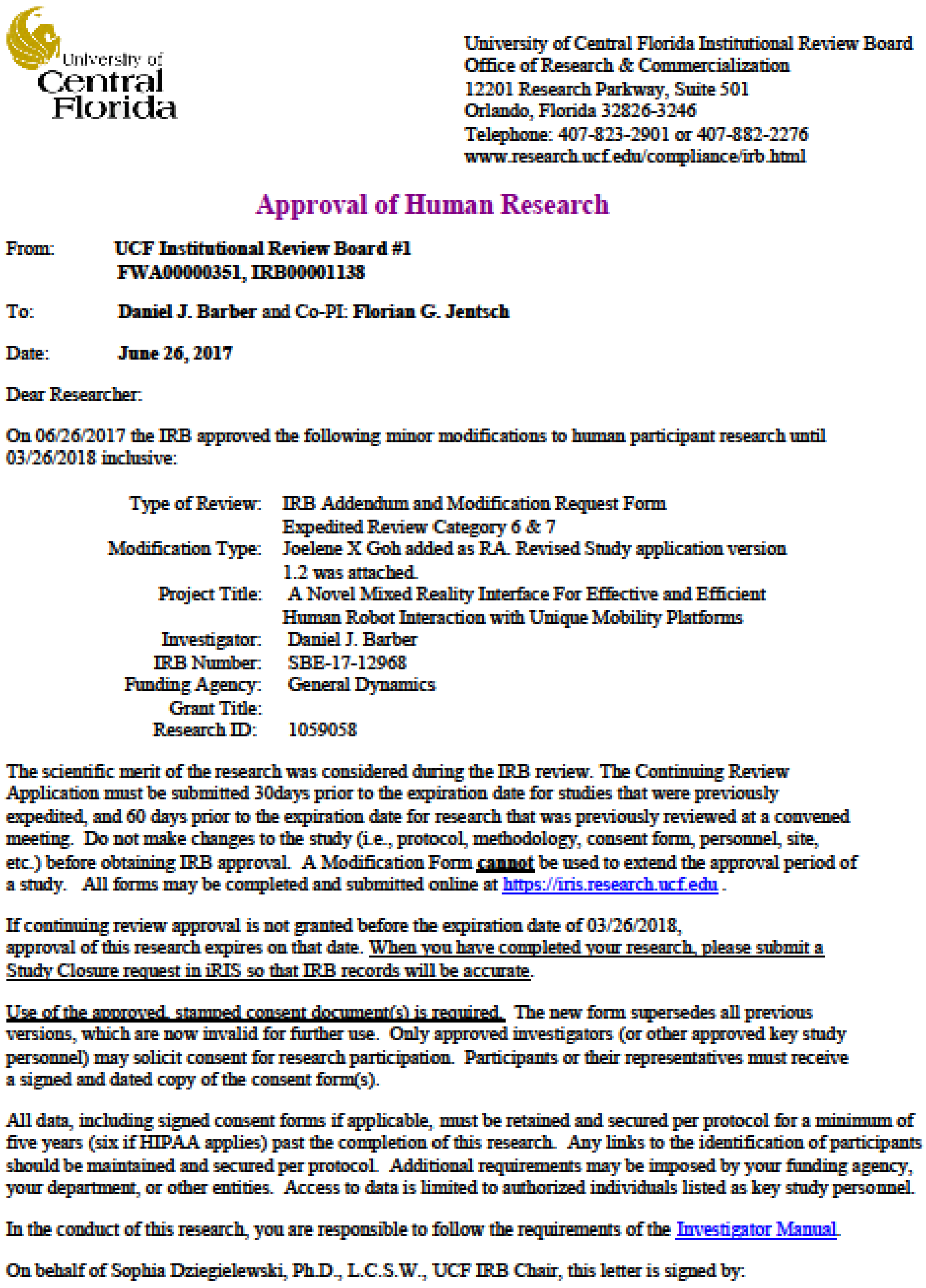
- *N/A*

18) Multi-site Human Research

- N/A

19) Sharing of results with participants

- Participants will have the option to inquire about the results of the study by contacting the experimenters.
- Experimenter contact information will be provided to the participants on the post participation information form provided upon the completion of the study.



University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **Daniel J. Barber** and Co-PI: **Florian G. Jentsch**

Date: **June 26, 2017**

Dear Researcher:

On 06/26/2017 the IRB approved the following minor modifications to human participant research until 03/26/2018 inclusive:

Type of Review:	IRB Addendum and Modification Request Form Expedited Review Category 6 & 7
Modification Type:	Joelene X Goh added as RA. Revised Study application version 1.2 was attached.
Project Title:	A Novel Mixed Reality Interface For Effective and Efficient Human Robot Interaction with Unique Mobility Platforms
Investigator:	Daniel J. Barber
IRB Number:	SBE-17-12968
Funding Agency:	General Dynamics
Grant Title:	
Research ID:	1059058

The scientific merit of the research was considered during the IRB review. The Continuing Review Application must be submitted 30days prior to the expiration date for studies that were previously expedited, and 60 days prior to the expiration date for research that was previously reviewed at a convened meeting. Do not make changes to the study (i.e., protocol, methodology, consent form, personnel, site, etc.) before obtaining IRB approval. A Modification Form cannot be used to extend the approval period of a study. All forms may be completed and submitted online at https://iris.research.ucf.edu .

If continuing review approval is not granted before the expiration date of 03/26/2018, approval of this research expires on that date. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

Use of the approved, stamped consent document(s) is required. The new form supersedes all previous versions, which are now invalid for further use. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Participants or their representatives must receive a signed and dated copy of the consent form(s).

All data, including signed consent forms if applicable, must be retained and secured per protocol for a minimum of five years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants should be maintained and secured per protocol. Additional requirements may be imposed by your funding agency, your department, or other entities. Access to data is limited to authorized individuals listed as key study personnel.

In the conduct of this research, you are responsible to follow the requirements of the Investigator Manual.

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:

APPENDIX H: IRB FOR BORROWED EXPERIMENTAL STUDY B

APPENDIX H: IRB FOR BORROWED EXPERIMENTAL STUDY B

- **If you believe your activity may not meet the definition of "Human Research" subject to IRB oversight, contact the IRB Office prior to developing your protocol**
- **Be sure that all study materials are correct and consistent with the information in this protocol.**
- **The italicized bullet points below serve as general guidance to investigators on the kinds of information that may be applicable to include in each section. Please DELETE the italicized text in your protocol.**
- **Note that, depending on the nature of your research, some sections below will not be applicable. Indicate this as "N/A."**
- **For any items described in the sponsor's protocol or other documents submitted with the application, investigators may simply reference the page numbers of these documents.**

1) Protocol Title

- Squad Level Soldier-Robot Communication Exchanges

2) Principal Investigator

- Principle Investigator: Daniel J. Barber
- Co-PI: Florian Jentsch
- Research Assistants: Julian Abich IV, Jonathan Harris, Samuel Cosgrove, Elizabeth Phillips, Andrew Talone

3) Objectives

- Collect Soldier feedback on types of information desired from a robot teammate
- Collect Soldier feedback on how robots should request confirmation regarding route planning and how robots should move when en route
- Collect information regarding frequency and type of information a Soldier requests from a robot teammate

- Collect information regarding how frequency and type of information affects a Soldier’s SA
- Collect information regarding a Soldier’s ability to interpret image data from a robot asset

4) Background

The future vision of a Soldier—robot (S-R) team is one in which humans and robots complete distributed but interdependent tasks to meet team goals. This vision of robotic teammates is one in which robots will be expected to be active participants in facilitating situation awareness (SA) among S-R teams. Military doctrine specifies that “Every Soldier is a sensor” on the battlefield (United States Army, 2012, pp. 9-1), therefore, Soldiers will expect robots to contribute to operator SA by understanding information that is relevant to the task at hand and sharing this information in an effective, proactive way (Robotics Collaborative Technology Alliance, 2012; Schuster, Keebler, Zuniga, & Jentsch, 2012). Emerging Soldier systems include advanced sensors that can penetrate walls, detect thermal signatures, localize enemy fire through 3D audio, and detect/recognize moving entities (U.S. Army Evaluation Center, 2013). They also include advanced networks for inter-and intra-squad communications. Robots will be expected to have some of these capabilities and engage in situation assessment behaviors, to perceive and understand surroundings, share information and report status (Endsley, 1995), in order to achieve SA within the team.

A robot’s ability to engage in these behaviors, and consequently aid in the development of team SA, will be guided by mental models to determine what information is relevant and when to share said information. However, based on what is currently known about the state of the art (SOA) in human—robot teams and team performance in human—human teams, we know that humans and robots have different levels of complexity with regard to mental models for engaging in situation assessment behaviors (i.e., information sharing). Assuming a robot system with some level of AI, the task-goal architecture is nevertheless still simple (e.g., using ladar and camera, recon the interior of a building) and lacking in contingencies/nuances. In high performing human—human teams, human team members often draw on highly complex mental models that enable members to “push” information in anticipation of team member information needs (Johannesen, Cook, & Woods, 1994; MacMillan, Entin, & Serfaty, 2004).

For this effort, investigations of Soldier SA will be based on realistic simulation-based scenarios with SA questions relevant to scenario events. Investigations of team-member SA benefit from careful construction of scenario events that elicit and document team communications and decision making, which in turn demonstrate the critical role of communication in shared SA (Elliott, Serfaty, & Schiflett, 1998 Elliott, Coovert, Barnes, & Miller, 2003). This communication strategy is dictated by knowledge of teammate expectations of information sharing. As a result, members transfer information to teammates, without explicit prompting. In SOA human—robot teams, robots share information based on their internal programming, dictated not by an understanding of when humans will expect or need information, but on design decisions bounded by practical and functional limitations. In order to reconcile mental model and design differences in situation assessment behaviors, like information sharing, research is required to determine Soldier expectations of robot information sharing and the degree to which these expectations and behaviors can best support team SA, leading to more efficient and effective team performance and increased Soldier safety.

Previous research has provided insight into perceived mental models of robotic teammates along several dimensions. Dimensions include the human's perception of the robot's own knowledge of its operating procedures, system limitations, interaction patterns, as well as the robot's knowledge of its human teammates (e.g., teammate specific knowledge, skills, and attitudes) (Ososky, Phillips, Swigert, & Jentsch, 2012). While this research has a wealth of insight into what novices infer about their robotic partner's understanding of tasks and teammates, it has not provided insight into mental models of specific robotic behaviors. As a result, we do not yet have an understanding of the mental models that humans hold of robot situation assessment behaviors. With this research, we would like to investigate human expectations and preferences for frequency of information sharing, type of information, and presentation of robot queries for information, that robots should communicate to Soldiers in a mission environment. We are also interested in the degree to which these information sharing behaviors influence a Soldier's SA.

In this effort, we will gather Soldier feedback through a simulation based assessment approach, to identify Soldier expectations of robot information sharing and information requesting behaviors. The results will inform the design of robot mental models of information sharing (i.e., robot-to-human communication protocols) and interfaces for facilitating S-R

References

Barber, D. J., Leontyev, S., Sun, B., Davis, L., Nicholson, D., & Chen, J. Y. (2008). The Mixed Iniative Experimental (MIX) Testbed for Collaborative Human Robot Interactions. *Army Science Conference.* Orlando: DTIC.

Chen, J. Y., & Barnes, M. J. (2012, April). Supervisory Control of Multiple Robots: Effects of Imperfect Automation and Individual Differences. *The Journal of the Human Factors and Ergonomics Societ*, 154-174.

Chen, J. Y., Barnes, M. J., & Qu, Z. (2010). RoboLeader: an agent for supervisory control of multiple robots. *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI'10)*, (pp. 81-82).

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for Kit of Factor-Referenced Cognitive Tests.* Princeton, New Jersey: Educational Testing Service.

Elliott, L. R., Serfaty, D., & Schiflett, S. G. (1998). Theory-based Development of Synthetic Team Task Environments for C3 Team Performance and Training Research. Proceedings of the 1998 Command and Control Research and Technology Symposium, Naval Postgraduate School, Monterey, CA, June, 1998.

Elliott, L.; Coovert, M.; Barnes, C.; Miller, J. (2003). Modeling Performance in C4ISR Sustained Operations: A Multi-level Approach. Proceedings of the 8th International Command and Control Research and Technology Symposium, National Defense University, Washington DC

Endsley, M. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors*, *37*(1), 32-64.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. (P. A. Hancock, & N. Meshkati, Eds.) *Human mental workload, 1*(3), 139-184.

Ishihara, S. (1917). *Tests for color-blindness.* Handaya, Tokyo, Hongo Harukicho.

Johannesen, L. J., Cook, R. I., & Woods, D. D. (1994). Cooperative communications in dynamic fault management. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 38(4)*, pp. 225-229.

MacMillan, J., Entin, E. E., & Serfaty, D. (2004). Communication overhead: The hidden cost of team cognition. In E. Salas, & S. M. Fiore, *Team Cognition: Understanding the Factors that Drive Process and Performance* (pp. 61-82). Washington, DC: American Psychological Association.

Ososky, S., Phillips, E., Swigert, B., & Jentsch, F. (2012). *The influence of training and mental model formation of robot behavior and intention in a simulated environment.* Contract No. W911NF-10-2-0016.

Robotics Collaborative Technology Alliance. (2012). FY 2012 Annual program plan. *FY 2012 Annual program plan (Contract No. W011NF-10-2-0016).*

Schuster, D., Keebler, J., Zuniga, J., & Jentsch, F. (2012). Individual differences in SA measurement and performance in human-robot teaming. New Orleans, LA: Poster session presented at the meeting of the IEEE Conference on Cognitive Methods in Situation Awareness and Decision Support.

Taylor, R. M. (1990). Situational Awareness Rating Technique (SART): The Development of a Tool for Aircrew Systems Design. *AGARD Situational Awareness in Aerospace Operations, 4*(17).

United States Army. (2012). *The warrior ethos and Soldier combat skills (Field Manual No. 3-21-75).* Washington, DC.

U.S. Army Evaluation Center (2013). *Army Expeditionary Warrior Experiement(AEWE) Spiral H Final Report. R*equests for this document must be referred to Commander, U.S. Army Test and Evaluation Command (CSTE-AEC-FFE), 2202 Aberdeen Boulevard, Third Floor, Aberdeen Proving Ground, MD 21005-5001

5) Setting of the Human Research

Experiments will be conducted at Fort Benning, GA in collaboration and under the supervision of the U.S. Army Research Laboratory (ARL). This is a field experiment and will be conducted in a designated area on base determined by ARL.

Permission has been confirmed by Linda Elliot, Ph.D., Human Research and Engineering Directorate, Human Factors Integration Division.

6) Resources available to conduct the Human Research

This project is funded by the RCTA FY2014 Task H5: Evaluating Tactical Command and Coordination Vocabulary and Protocols. Research staffing, testing equipment and testbeds are provided by the University of Central Florida's Institute for Simulation and Training.

7) Study Design

The study will be a 2 (constant demand, varying demand) x 2 (participant request, robot request) within-subject design. The first independent variable is signal detection task demand with two levels: Baseline (constant low) and varying (low to high). The second independent variable is a communication type with two levels with constant signal detection task demand: participant requests information and robot requests information.

j) Recruitment Methods

This experiment is going to be conducted on the Fort Benning, GA military base and up to 60 OCS Soldiers will participate in the study. This experiment is a joint collaboration with HRED-ARL, which will help provide the sample population. The project investigators will make clear to the unit that Soldier participation in the evaluation will be voluntary. The Soldiers will be informed that if they choose not to participate, they can convey that choice privately to a project investigator. The project investigator will inform that Soldier's unit supervisor, without elaboration, that the Soldier did not meet evaluation criteria.

k) Participant Compensation

Participants will be not be compensated for participantion.

l) Inclusion and Exclusion Criteria

Participants will be healthy American Citizens over age 18 years from the U.S. Army community. Participants will be notified by their unit leader that they may be excluded from the study before the study actually begins if they do not meet the inclusion criteria.

Reasons for exclusion are:

- Color-blindness

m) Study Endpoints

The results will inform the design of robot mental models of information sharing (i.e., robot-to-human communication protocols) and recommendations of display characteristics for facilitating S-R communications.

n) Study Timelines

Individual participation in the study will be about 2 hours. The duration anticipated to enroll all study participants will be 2 weeks. The estimated date to complete this study will be July 2015.

o) Procedures involved in the Human Research.

Upon arrival, participants will first complete the Informed Consent that details their rights as a research participant, the purpose of the study, overall procedure, and potential risks associated with participation. After reviewing and signing the Informed Consent, the participant will complete the Demographics Questionnaire to collect standard items such as age and gender, as well as items used to determine their level of training and experience. After completion of the Demographics Questionnaire, the participant will complete the Cube Comparison Test. Once all pre-questionnaires are completed, the participant will begin training for the experiment scenarios.

Participants will be shown a PowerPoint-based presentation instructing them on the tasks they will perform. It will include descriptions of threat and non-threat targets for the signal detection task. This presentation will include screenshots of the simulation environment with instructions on how to classify potential threats. After reviewing the PowerPoint information, the first training scenario will be administered to allow participants to practice performing the signal detection task. After completing the signal detection practice scenario, the participant will continue to the next phase of the PowerPoint presentation, which will include information regarding when and what type of information the robot will provide during Robot Reporting (RR) tasks. Next, a practice task will be administered providing practice of receiving audio cues, requesting robot status information, and answering SA questions. Upon practice task completion, the next training phase will be given using the PowerPoint presentation regarding what types of navigation questions and aides the robot may require during Robot Assistance (RA) tasks. Similar to the previous training phases, a practice scenario focused on requesting assistance from the robot will be administered. After completing training for each individual task, participants will then be given two additional practice tasks replicating the experimental scenarios. The first practice task will include both signal detection and RR tasks. The second practice involves signal detection and RA tasks. Project investigators will brief Soldiers on the purpose of the each task, and go through the training with each Soldier. Soldiers will be trained on simulation procedures until they demonstrate adequate proficiency to perform the simulation tasks. They will then be requested to provide feedback regarding their knowledge of experiment goals, quality and sufficiency of training content and practice, and indicate their level of confidence (self-efficacy) to perform simulation tasks.

After completion of all training materials and tasks, participants will perform the three experimental scenarios. Project investigators will randomize and counterbalance presentation order of experimental scenarios across participants. Participants will complete two RR scenarios and one RA. One RR scenario will have constant signal detection demand and the other varying from low to high. The level of demand of the signal detection task will be varied through manipulation of the signal to noise ratio, with demand changing half-way through the scenario. The RA scenario will have constant signal detection demand. After completing each experimental scenario, participants will complete the NASA-TLX followed by the SART. For the RR task within varying signal detection task demand, the NASA-TLX will be measured half-way through the scenario and at the end. Upon completion of all experimental scenarios, participants will be administered the Robot Movement Questionnaire.

p) Data and specimen management

See procedures and provisions sections.

q) Provisions to monitor the data for the safety of participants

No more than minimal risk is anticipated

r) Withdrawal of participants

Individuals will be informed that participation in the study is voluntary and that they may withdraw at any time without penalty.

8) Risks to participants

There are no foreseeable risks or discomforts other than those normally encountered in the daily lives of healthy persons. As in all studies, there is a potential risk to participants; however, in this study those risks are minimal. Specifically, there is always a chance of data loss or misplacement. This potential risk is reduced by keeping data separate from informed consents, in locked cabinets, and identifiable only by numerical ID numbers.

9) Potential direct benefits to participants

No benefits have been promised or are expected to be given to the volunteers who participate in this study. However, the data resulting from this research will be the primary information used to inform designers of robotic systems, specifically in robot communication behaviors.

10) Provisions to protect the privacy interests of participants

Data in any form will be kept either in a locked cabinet or maintained on a password protected computer with limited access. Only persons listed on the IRB will have access to the information.

11) Provisions to maintain the confidentiality of data

See above

12) Medical care and compensation for injury

N/A

13) Cost to participants

Participants will not incur any costs for participation

14) Consent process

When participants arrive for the experimental session, they will be briefed on the experimental procedure and asked to read an IRB-approved informed consent form. Participants will be allowed to ask questions of the experimenter at any time and all questions will be answered completely. Following completion of the informed consent form, participants will be assigned a participant number so that all data will remain anonymous. This number will be kept separate from the participant's name, so all data collected will be associated with only this number and will not be traceable to a specific individual.

Because this research is funded by the U.S. ARMY, the informed consent process will also comply with U.S. ARMY standards for ethical research. Meaning that, in the event that this research is considered "exempt" by the UCF institutional review board, the researchers will still seek a signed informed consent document, so as to be compliant with both UCF's IRB and the U.S. ARMY's HLAR/AHRPO review process.

15) Process to document consent in writing

Although this study is of minimal risk and may qualify for a waiver of written documentation of consent, in compliance with DOD standards, participants will indicate their consent to participate by signing their name on the Informed Consent form. The research assistant conducting the study will also sign as the person obtaining consent. A copy of this document will be made and given to the participant to keep for their own records. The research team will also keep a copy of this document that will be stored in a secure locked filing cabinet away from other study materials so as to avoid any chance of linking participant names to other study materials.

16) Vulnerable populations

N/A

17) Drugs or Devices

N/A

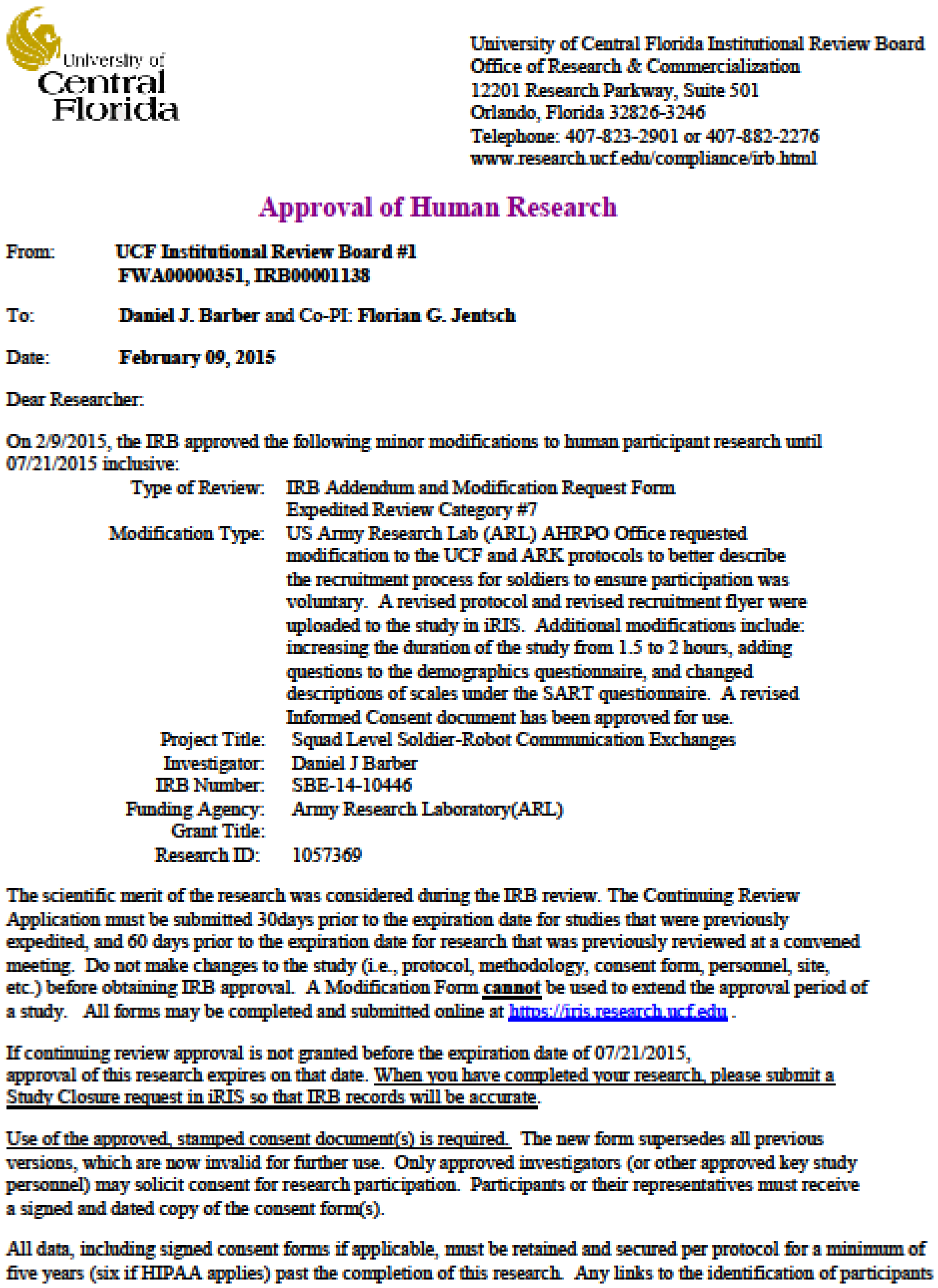
18) Multi-site Human Research

N/A

19) Sharing of results with participants

Results will not be shared with participants. Participants can obtain approved-publicly released reports such as journals articles and conference proceedings.

Experimenter contact information will be provided to the subjects on the post participation information form provided upon the completion of the study.



University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **Daniel J. Barber** and Co-PI: **Florian G. Jentsch**

Date: **February 09, 2015**

Dear Researcher:

On 2/9/2015, the IRB approved the following minor modifications to human participant research until 07/21/2015 inclusive:

Type of Review:	IRB Addendum and Modification Request Form Expedited Review Category #7
Modification Type:	US Army Research Lab (ARL) AHRPO Office requested modification to the UCF and ARK protocols to better describe the recruitment process for soldiers to ensure participation was voluntary. A revised protocol and revised recruitment flyer were uploaded to the study in iRIS. Additional modifications include: increasing the duration of the study from 1.5 to 2 hours, adding questions to the demographics questionnaire, and changed descriptions of scales under the SART questionnaire. A revised Informed Consent document has been approved for use.
Project Title:	Squad Level Soldier-Robot Communication Exchanges
Investigator:	Daniel J Barber
IRB Number:	SBE-14-10446
Funding Agency:	Army Research Laboratory(ARL)
Grant Title:	
Research ID:	1057369

The scientific merit of the research was considered during the IRB review. The Continuing Review Application must be submitted 30days prior to the expiration date for studies that were previously expedited, and 60 days prior to the expiration date for research that was previously reviewed at a convened meeting. Do not make changes to the study (i.e., protocol, methodology, consent form, personnel, site, etc.) before obtaining IRB approval. A Modification Form **cannot** be used to extend the approval period of a study. All forms may be completed and submitted online at https://iris.research.ucf.edu .

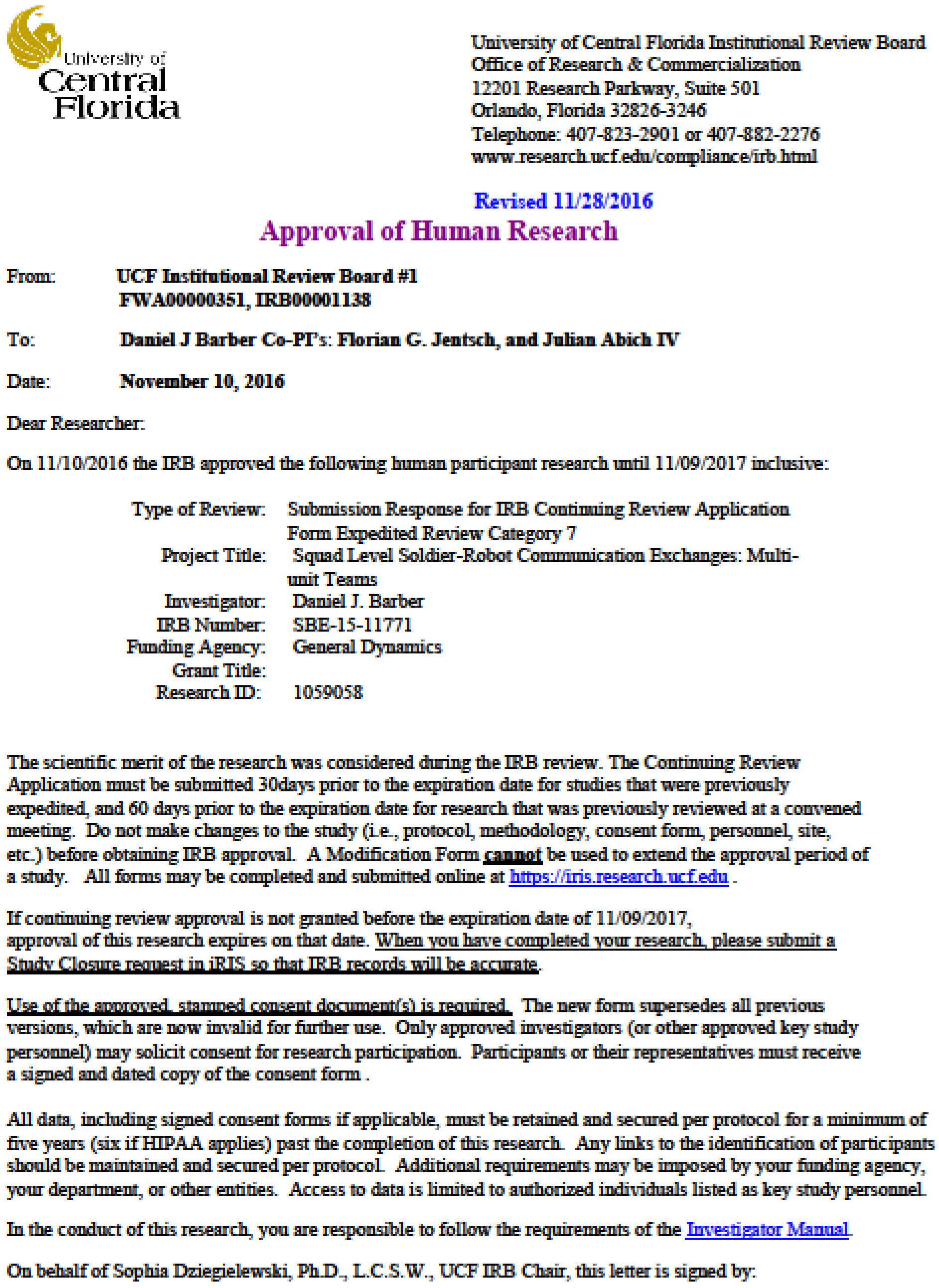
If continuing review approval is not granted before the expiration date of 07/21/2015, approval of this research expires on that date. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

Use of the approved, stamped consent document(s) is required. The new form supersedes all previous versions, which are now invalid for further use. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Participants or their representatives must receive a signed and dated copy of the consent form(s).

All data, including signed consent forms if applicable, must be retained and secured per protocol for a minimum of five years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants

APPENDIX I: IRB FOR BORRWED EXPERIMENTAL STUDY C

APPENDIX I: IRB FOR BORRWED EXPERIMENTAL STUDY C



University of Central Florida

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Revised 11/28/2016

Approval of Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **Daniel J Barber Co-PI's: Florian G. Jentsch, and Julian Abich IV**

Date: **November 10, 2016**

Dear Researcher:

On 11/10/2016 the IRB approved the following human participant research until 11/09/2017 inclusive:

Type of Review: Submission Response for IRB Continuing Review Application Form Expedited Review Category 7
Project Title: Squad Level Soldier-Robot Communication Exchanges: Multi-unit Teams
Investigator: Daniel J. Barber
IRB Number: SBE-15-11771
Funding Agency: General Dynamics
Grant Title:
Research ID: 1059058

The scientific merit of the research was considered during the IRB review. The Continuing Review Application must be submitted 30days prior to the expiration date for studies that were previously expedited, and 60 days prior to the expiration date for research that was previously reviewed at a convened meeting. Do not make changes to the study (i.e., protocol, methodology, consent form, personnel, site, etc.) before obtaining IRB approval. A Modification Form **cannot** be used to extend the approval period of a study. All forms may be completed and submitted online at https://iris.research.ucf.edu .

If continuing review approval is not granted before the expiration date of 11/09/2017, approval of this research expires on that date. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

Use of the approved, stamped consent document(s) is required. The new form supersedes all previous versions, which are now invalid for further use. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Participants or their representatives must receive a signed and dated copy of the consent form .

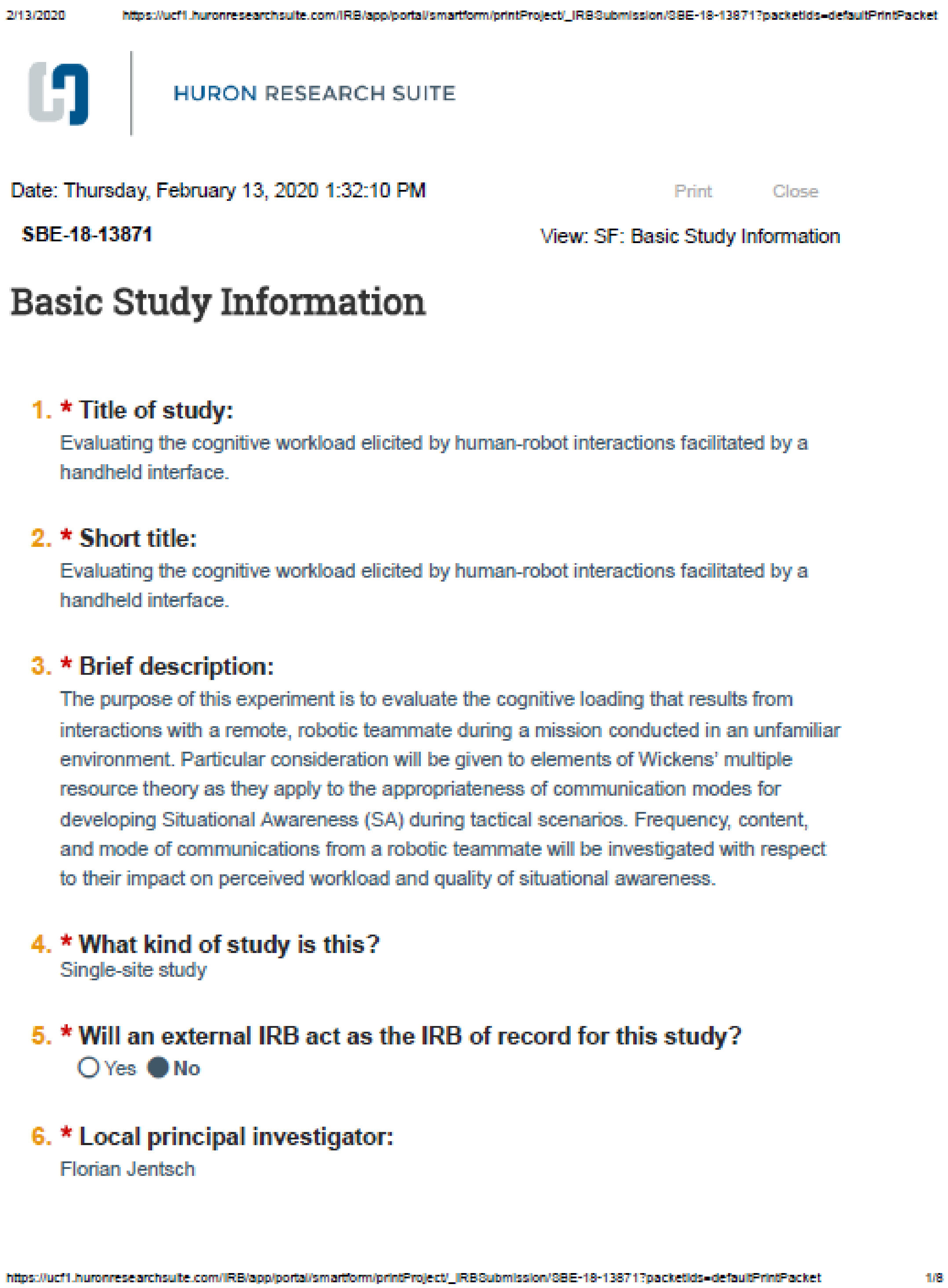
All data, including signed consent forms if applicable, must be retained and secured per protocol for a minimum of five years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants should be maintained and secured per protocol. Additional requirements may be imposed by your funding agency, your department, or other entities. Access to data is limited to authorized individuals listed as key study personnel.

In the conduct of this research, you are responsible to follow the requirements of the Investigator Manual.

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:

APPENDIX J: IRB FOR BORROWED EXPERIMENTAL STUDY D

APPENDIX J: IRB FOR BORROWED EXPERIMENTAL STUDY D



2/13/2020 https://ucf1.huronresearchsuite.com/IRB/app/portal/smartform/printProject/_IRBSubmission/SBE-18-13871?packetIds=defaultPrintPacket

HURON RESEARCH SUITE

Date: Thursday, February 13, 2020 1:32:10 PM

SBE-18-13871

View: SF: Basic Study Information

Basic Study Information

1. * **Title of study:**
Evaluating the cognitive workload elicited by human-robot interactions facilitated by a handheld interface.

2. * **Short title:**
Evaluating the cognitive workload elicited by human-robot interactions facilitated by a handheld interface.

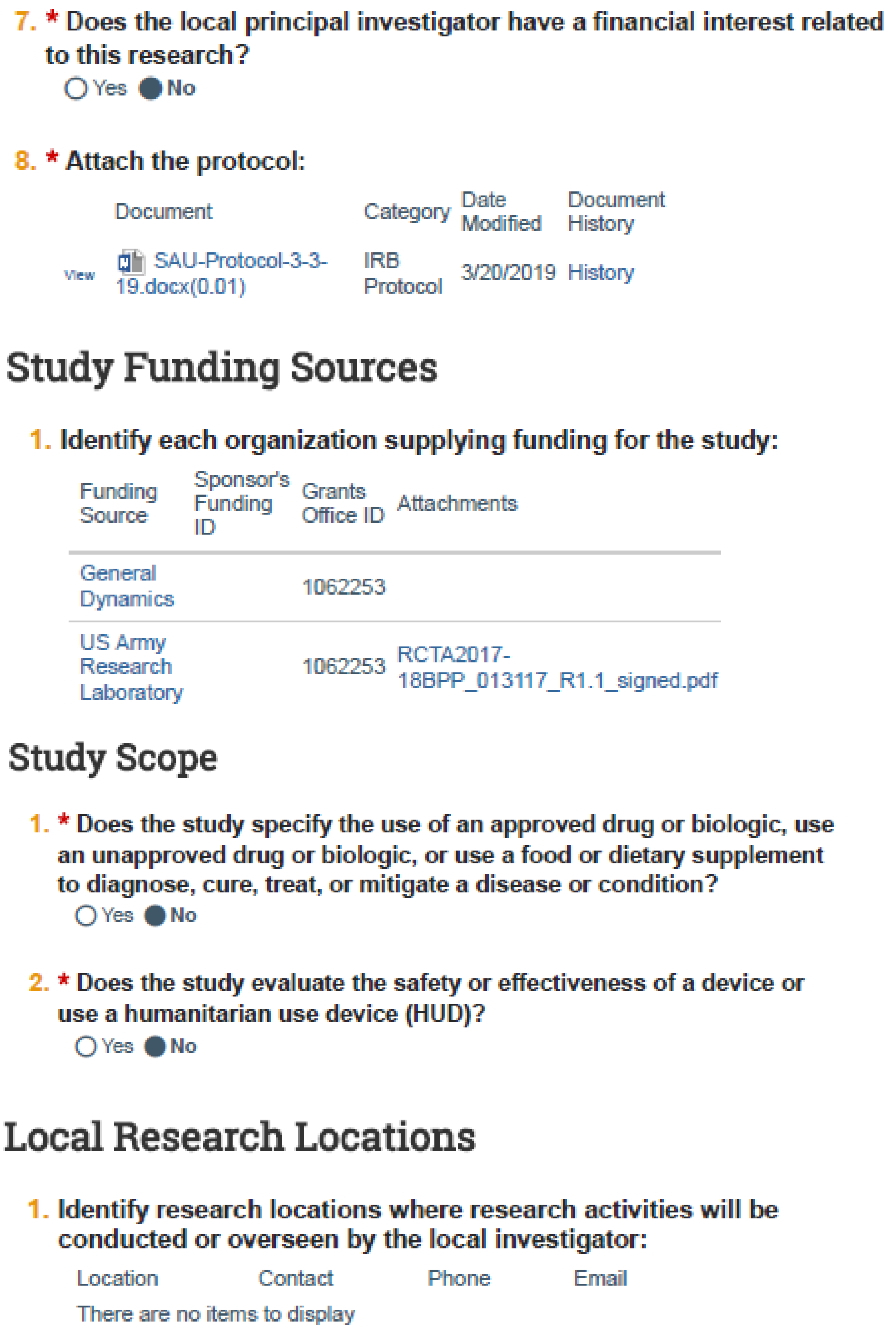
3. * **Brief description:**
The purpose of this experiment is to evaluate the cognitive loading that results from interactions with a remote, robotic teammate during a mission conducted in an unfamiliar environment. Particular consideration will be given to elements of Wickens' multiple resource theory as they apply to the appropriateness of communication modes for developing Situational Awareness (SA) during tactical scenarios. Frequency, content, and mode of communications from a robotic teammate will be investigated with respect to their impact on perceived workload and quality of situational awareness.

4. * **What kind of study is this?**
Single-site study

5. * **Will an external IRB act as the IRB of record for this study?**
○ Yes ● No

6. * **Local principal investigator:**
Florian Jentsch

https://ucf1.huronresearchsuite.com/IRB/app/portal/smartform/printProject/_IRBSubmission/SBE-18-13871?packetIds=defaultPrintPacket 1/8



7. * Does the local principal investigator have a financial interest related to this research?

○ Yes ● No

8. * Attach the protocol:

	Document	Category	Date Modified	Document History
View	SAU-Protocol-3-3-19.docx(0.01)	IRB Protocol	3/20/2019	History

Study Funding Sources

1. Identify each organization supplying funding for the study:

Funding Source	Sponsor's Funding ID	Grants Office ID	Attachments
General Dynamics		1062253	
US Army Research Laboratory		1062253	RCTA2017-18BPP_013117_R1.1_signed.pdf

Study Scope

1. * Does the study specify the use of an approved drug or biologic, use an unapproved drug or biologic, or use a food or dietary supplement to diagnose, cure, treat, or mitigate a disease or condition?

○ Yes ● No

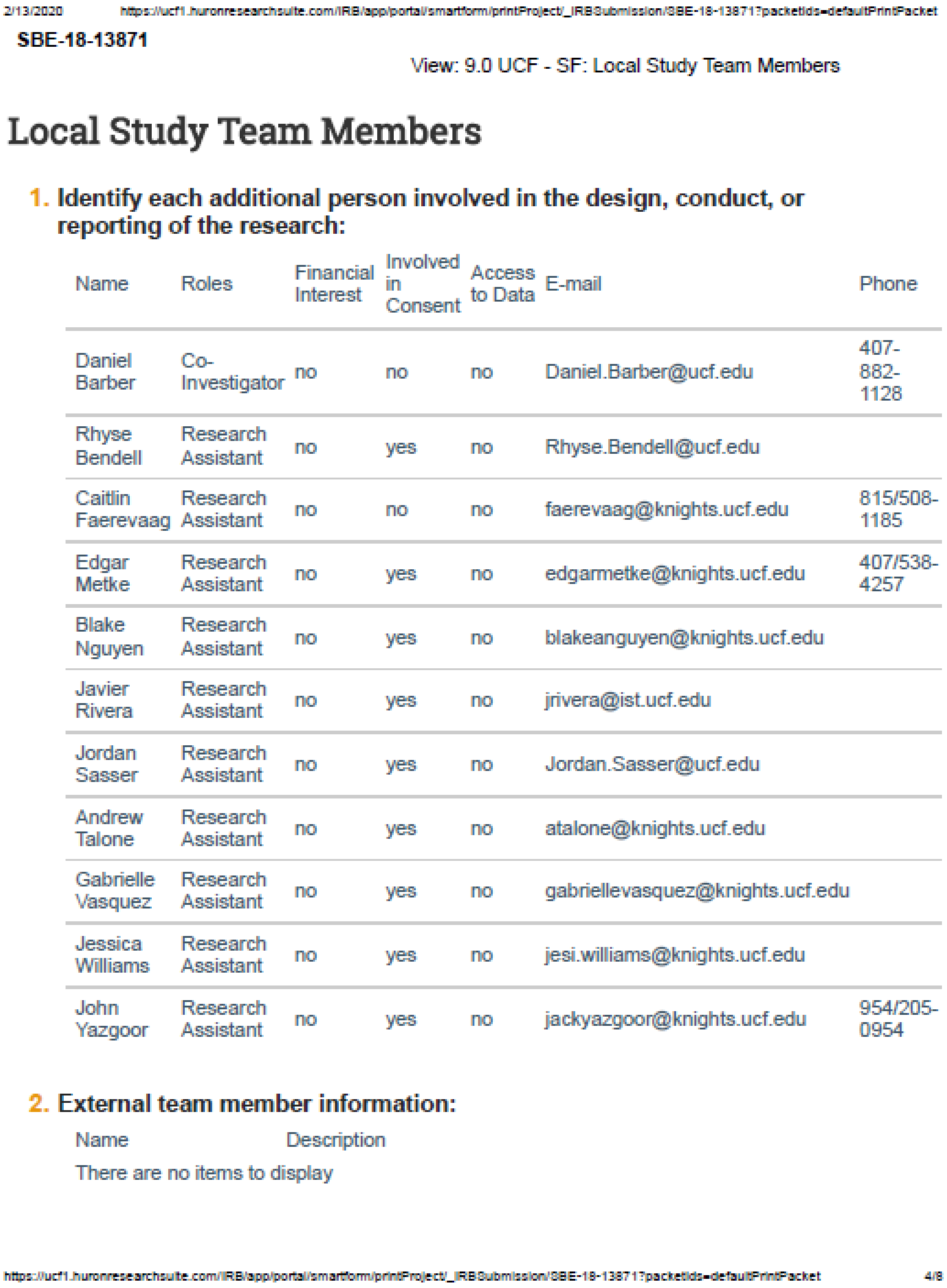
2. * Does the study evaluate the safety or effectiveness of a device or use a humanitarian use device (HUD)?

○ Yes ● No

Local Research Locations

1. Identify research locations where research activities will be conducted or overseen by the local investigator:

Location	Contact	Phone	Email
There are no items to display			



SBE-18-13871

View: 9.0 UCF - SF: Local Study Team Members

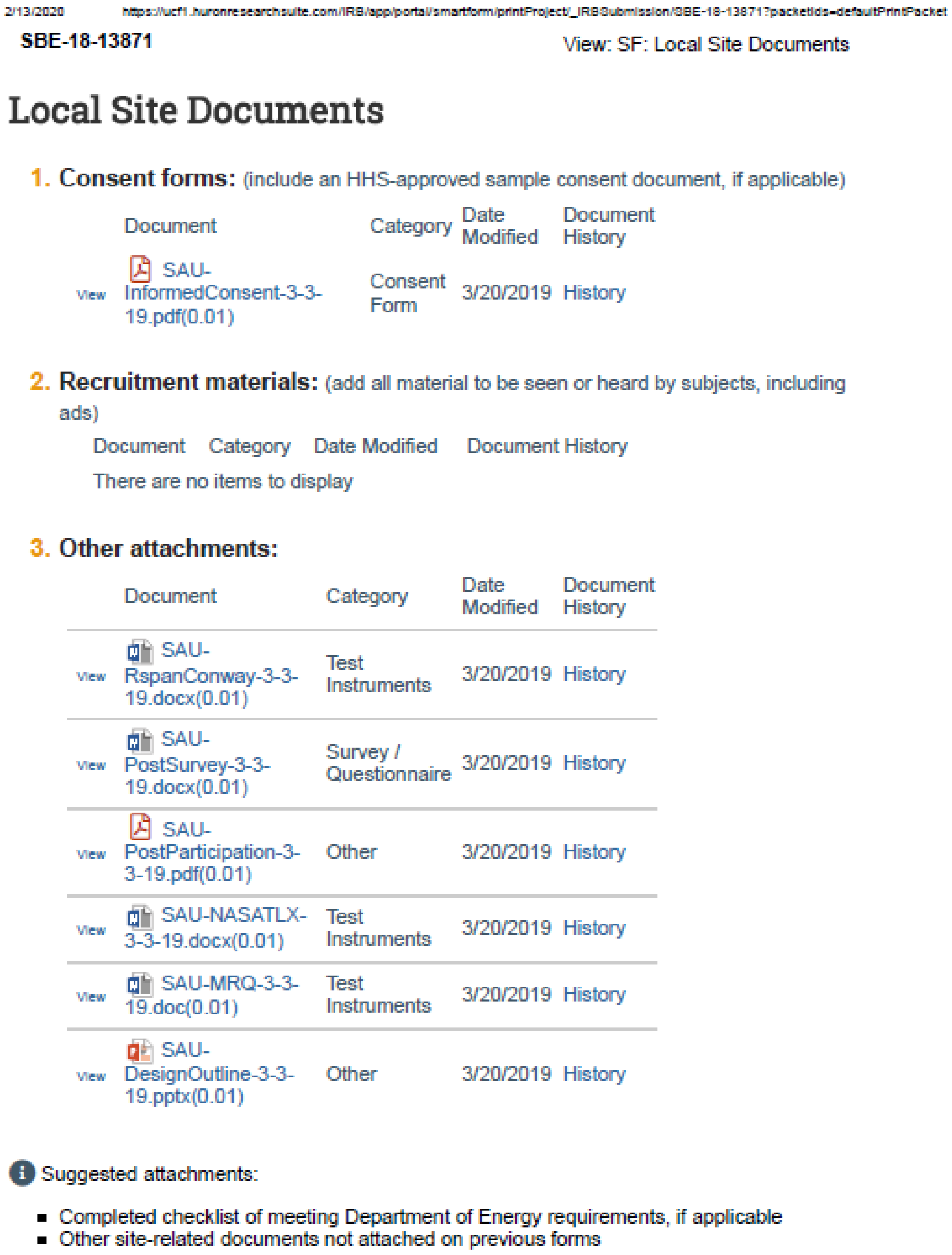
Local Study Team Members

1. Identify each additional person involved in the design, conduct, or reporting of the research:

Name	Roles	Financial Interest	Involved in Consent	Access to Data	E-mail	Phone
Daniel Barber	Co-Investigator	no	no	no	Daniel.Barber@ucf.edu	407-882-1128
Rhyse Bendell	Research Assistant	no	yes	no	Rhyse.Bendell@ucf.edu	
Caitlin Faerevaag	Research Assistant	no	no	no	faerevaag@knights.ucf.edu	815/508-1185
Edgar Metke	Research Assistant	no	yes	no	edgarmetke@knights.ucf.edu	407/538-4257
Blake Nguyen	Research Assistant	no	yes	no	blakeanguyen@knights.ucf.edu	
Javier Rivera	Research Assistant	no	yes	no	jrivera@ist.ucf.edu	
Jordan Sasser	Research Assistant	no	yes	no	Jordan.Sasser@ucf.edu	
Andrew Talone	Research Assistant	no	yes	no	atalone@knights.ucf.edu	
Gabrielle Vasquez	Research Assistant	no	yes	no	gabriellevasquez@knights.ucf.edu	
Jessica Williams	Research Assistant	no	yes	no	jesi.williams@knights.ucf.edu	
John Yazgoor	Research Assistant	no	yes	no	jackyazgoor@knights.ucf.edu	954/205-0954

2. External team member information:

Name	Description
There are no items to display	



2/13/2020 https://ucf1.huronresearchsuite.com/IRB/app/portal/smartform/printProject/_IRBSubmission/SBE-18-13871?packetIds=defaultPrintPacket

SBE-18-13871 View: SF: Local Site Documents

Local Site Documents

1. Consent forms: (include an HHS-approved sample consent document, if applicable)

	Document	Category	Date Modified	Document History
View	SAU-InformedConsent-3-3-19.pdf(0.01)	Consent Form	3/20/2019	History

2. Recruitment materials: (add all material to be seen or heard by subjects, including ads)

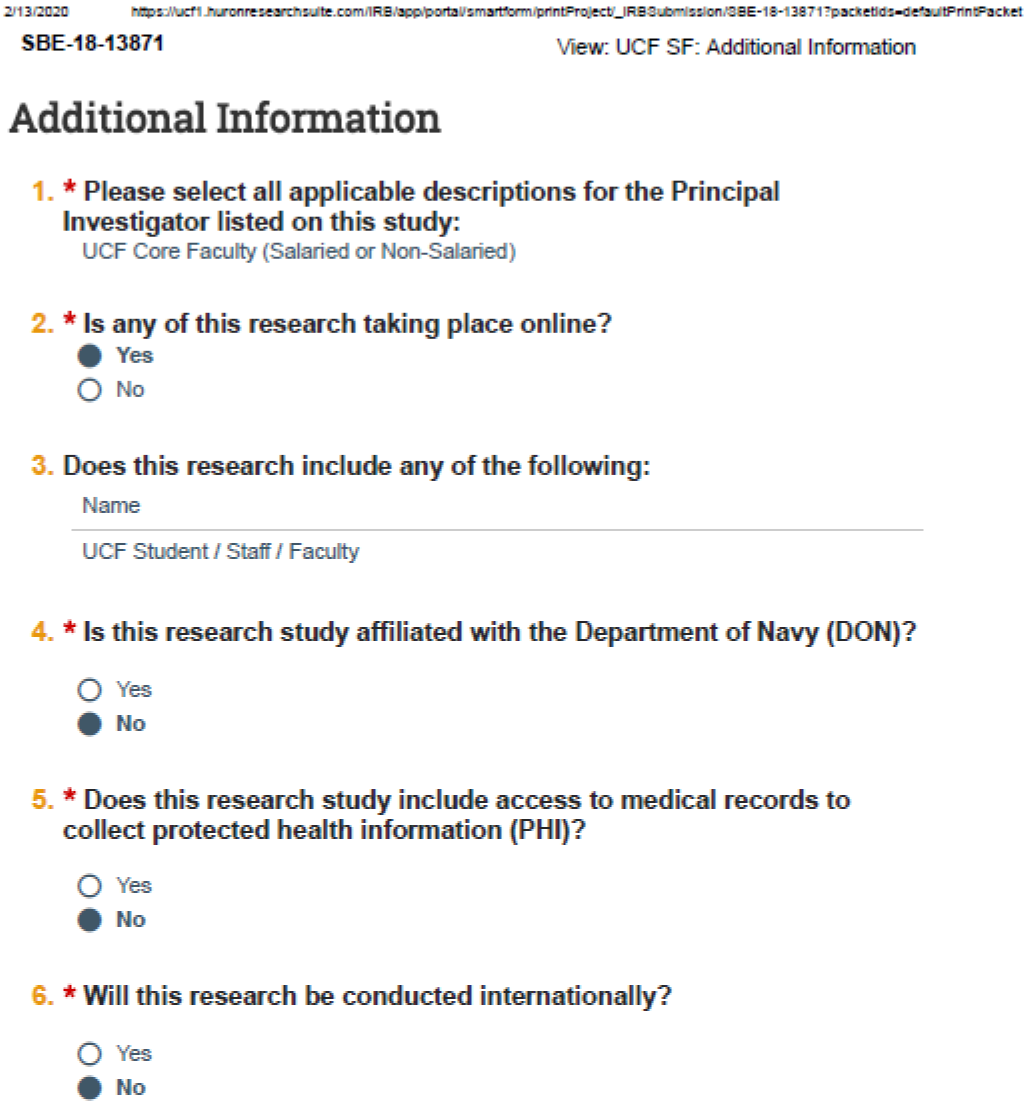
Document	Category	Date Modified	Document History
There are no items to display			

3. Other attachments:

	Document	Category	Date Modified	Document History
View	SAU-RspanConway-3-3-19.docx(0.01)	Test Instruments	3/20/2019	History
View	SAU-PostSurvey-3-3-19.docx(0.01)	Survey / Questionnaire	3/20/2019	History
View	SAU-PostParticipation-3-3-19.pdf(0.01)	Other	3/20/2019	History
View	SAU-NASATLX-3-3-19.docx(0.01)	Test Instruments	3/20/2019	History
View	SAU-MRQ-3-3-19.doc(0.01)	Test Instruments	3/20/2019	History
View	SAU-DesignOutline-3-3-19.pptx(0.01)	Other	3/20/2019	History

Suggested attachments:

- Completed checklist of meeting Department of Energy requirements, if applicable
- Other site-related documents not attached on previous forms



2/13/2020 https://ucf1.huronresearchsuite.com/IRB/app/portal/smartform/printProject/_IRBSubmission/SBE-18-13871?packetIds=defaultPrintPacket

SBE-18-13871 View: UCF SF: Additional Information

Additional Information

1. * **Please select all applicable descriptions for the Principal Investigator listed on this study:**
   UCF Core Faculty (Salaried or Non-Salaried)

2. * **Is any of this research taking place online?**
   ● **Yes**
   ○ No

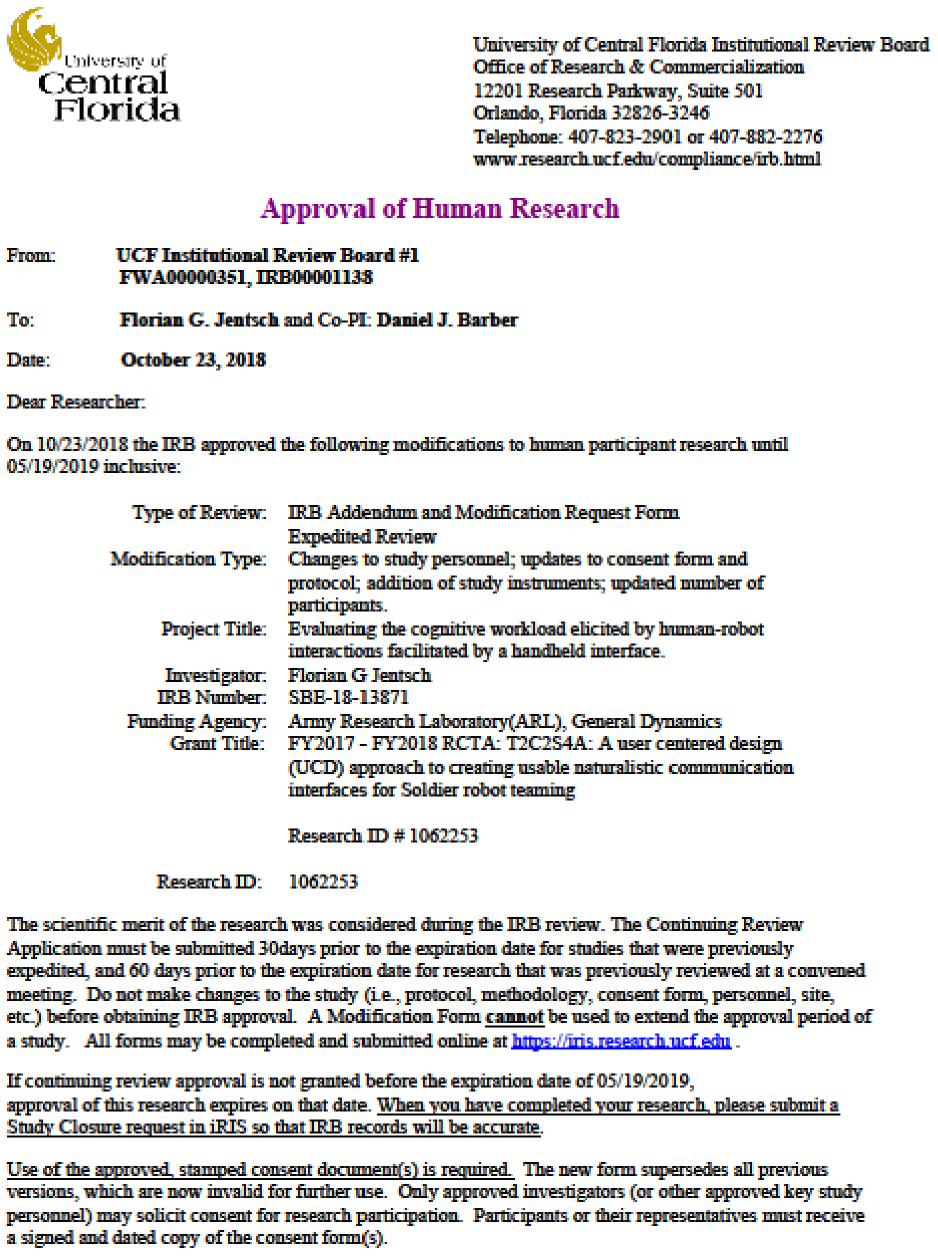
3. **Does this research include any of the following:**

Name
UCF Student / Staff / Faculty

4. * **Is this research study affiliated with the Department of Navy (DON)?**
   ○ Yes
   ● **No**

5. * **Does this research study include access to medical records to collect protected health information (PHI)?**
   ○ Yes
   ● **No**

6. * **Will this research be conducted internationally?**
   ○ Yes
   ● **No**



University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Human Research

From:	**UCF Institutional Review Board #1** **FWA00000351, IRB00001138**
To:	**Florian G. Jentsch** and Co-PI: **Daniel J. Barber**
Date:	**October 23, 2018**

Dear Researcher:

On 10/23/2018 the IRB approved the following modifications to human participant research until 05/19/2019 inclusive:

Type of Review:	IRB Addendum and Modification Request Form Expedited Review
Modification Type:	Changes to study personnel; updates to consent form and protocol; addition of study instruments; updated number of participants.
Project Title:	Evaluating the cognitive workload elicited by human-robot interactions facilitated by a handheld interface.
Investigator:	Florian G Jentsch
IRB Number:	SBE-18-13871
Funding Agency:	Army Research Laboratory(ARL), General Dynamics
Grant Title:	FY2017 - FY2018 RCTA: T2C2S4A: A user centered design (UCD) approach to creating usable naturalistic communication interfaces for Soldier robot teaming Research ID # 1062253
Research ID:	1062253

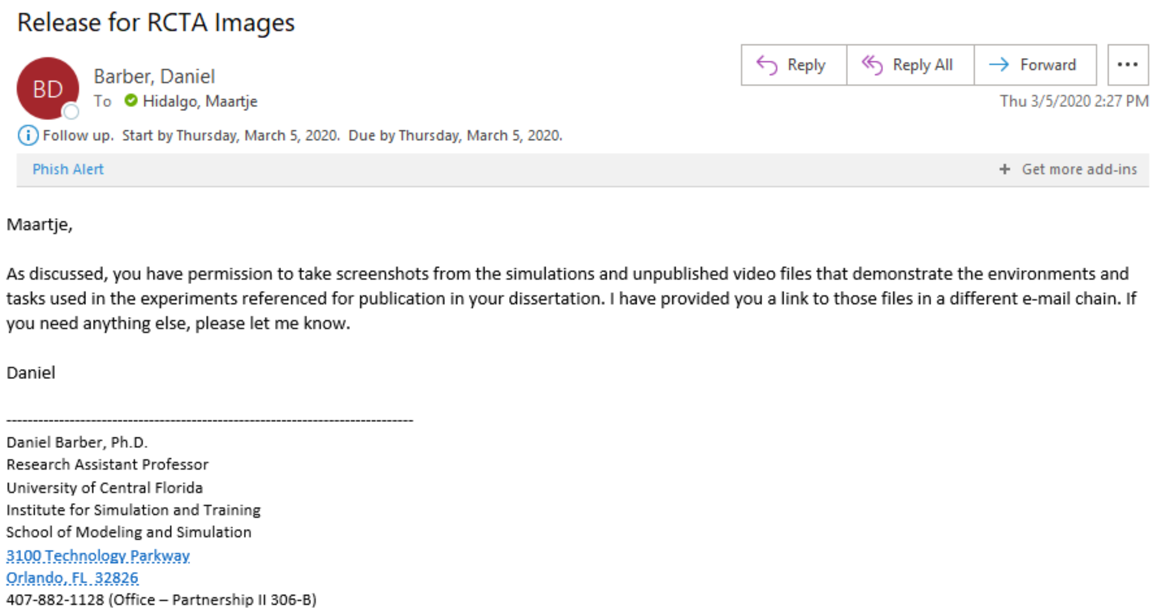
The scientific merit of the research was considered during the IRB review. The Continuing Review Application must be submitted 30days prior to the expiration date for studies that were previously expedited, and 60 days prior to the expiration date for research that was previously reviewed at a convened meeting. Do not make changes to the study (i.e., protocol, methodology, consent form, personnel, site, etc.) before obtaining IRB approval. A Modification Form **cannot** be used to extend the approval period of a study. All forms may be completed and submitted online at https://iris.research.ucf.edu .

If continuing review approval is not granted before the expiration date of 05/19/2019, approval of this research expires on that date. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

Use of the approved, stamped consent document(s) is required. The new form supersedes all previous versions, which are now invalid for further use. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Participants or their representatives must receive a signed and dated copy of the consent form(s).

APPENDIX K: COPYRIGHT

APPENDIX K: COPYRIGHT



Release for RCTA Images

Barber, Daniel
To Hidalgo, Maartje

Thu 3/5/2020 2:27 PM

Follow up. Start by Thursday, March 5, 2020. Due by Thursday, March 5, 2020.

Maartje,

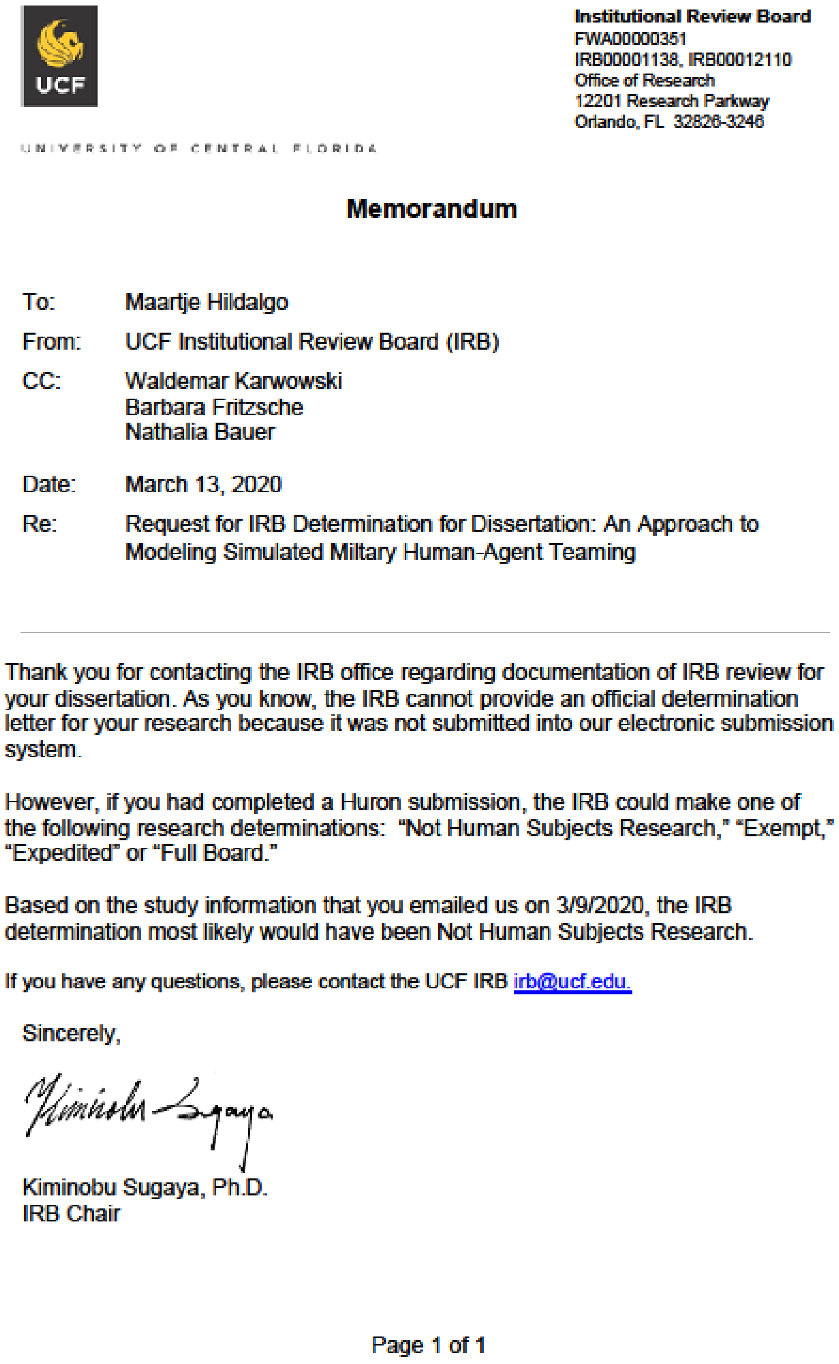
As discussed, you have permission to take screenshots from the simulations and unpublished video files that demonstrate the environments and tasks used in the experiments referenced for publication in your dissertation. I have provided you a link to those files in a different e-mail chain. If you need anything else, please let me know.

Daniel

Daniel Barber, Ph.D.
Research Assistant Professor
University of Central Florida
Institute for Simulation and Training
School of Modeling and Simulation
3100 Technology Parkway
Orlando, FL 32826
407-882-1128 (Office – Partnership II 306-B)

APPENDIX L: IRB DETERMINATION DISSERTATION

APPENDIX L: IRB DETERMINATION DISSERTATION



UCF

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138, IRB00012110
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

Memorandum

To: Maartje Hildalgo

From: UCF Institutional Review Board (IRB)

CC: Waldemar Karwowski
Barbara Fritzsche
Nathalia Bauer

Date: March 13, 2020

Re: Request for IRB Determination for Dissertation: An Approach to Modeling Simulated Miltary Human-Agent Teaming

Thank you for contacting the IRB office regarding documentation of IRB review for your dissertation. As you know, the IRB cannot provide an official determination letter for your research because it was not submitted into our electronic submission system.

However, if you had completed a Huron submission, the IRB could make one of the following research determinations: "Not Human Subjects Research," "Exempt," "Expedited" or "Full Board."

Based on the study information that you emailed us on 3/9/2020, the IRB determination most likely would have been Not Human Subjects Research.

If you have any questions, please contact the UCF IRB irb@ucf.edu.

Sincerely,

Kiminobu Sugaya, Ph.D.
IRB Chair

Page 1 of 1

REFERENCES

Abich IV, J., Reinerman-Jones, L., & Matthews, G. (2017). Impact of three task demand factors on simulated unmanned system intelligence, surveillance, and reconnaissance operations. *Ergonomics*, *60*(6), 791–809. https://doi.org/10.1080/00140139.2016.1216171

Abich, J., Barber, D. J., & Elliott, L. R. (2017). An initial investigation of exogenous orienting visual display cues for dismounted human-robot communication. In P. Savage-Knepshield & J. Chen (Eds.), *Advances in human factors in robots and unmanned systems* (pp. 27–38). Springer International Publishing.

Abich, J., Reinerman-Jones, L., & Taylor, G. (2013a). Establishing workload manipulations utilizing a simulated environment. In R. Shumaker (Ed.), *Virtual, augmented and mixed reality. Systems and applications.* (Vol. 8022, pp. 211–220). Springer.

Abich, J., Reinerman-Jones, L., & Taylor, G. S. (2013b). Investigating workload measures for adaptive training systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 2091–2095. https://doi.org/10.1177/1541931213571466

Adams, J. A. (2014). Shared mental models for human-robot teams. *Proceedings of AAAI Fall Symposium*, 5–6.

Allen, J. E., Guinn, C. I., & Horvtz, E. (1999). Mixed-initiative interaction. *IEEE Intelligent Systems and Their Applications*, *14*(5), 14–23. https://doi.org/10.1109/5254.796083

Allison, P. (2013, February 1). What's the best R-squared for logistic regression? *Statistical Horizons*. https://statisticalhorizons.com/r2logistic

Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 1078–1088.

Annett, J. (2002). Subjective rating scales: Science or art? *Ergonomics*, *45*(14), 966–987. https://doi.org/10.1080/00140130210166951

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods, 8*(2), 129–148.

Azen, R., & Budescu, D. V. (2006). Comparing predictors in multivariate regression models: An extension of dominance analysis. *Journal of Educational and Behavioral Statistics*, *31*(2), 157–180. https://doi.org/10.3102/10769986031002157

Azen, R., & Traxel, N. (2009). Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics*, *34*(3), 319–347.

Baber, C., Morin, C., Parekh, M., Cahillane, M., & Houghton, R. J. (2011). Multimodal control of sensors on multiple simulated unmanned vehicles. *Ergonomics*, *54*(9), 792–805. https://doi.org/10.1080/00140139.2011.597516

Bailer-Jones, Daniela M. (2003). When scientific models represent. *International Studies in the Philosophy of Science, 17*(1), 59–74. https://doi.org/10.1080/02698590305238

Bailer-Jones, D.M. (2002). Models, metaphors and analogies. In P. Machamer & M. Silberstein (Eds.), *The Blackwell guide to the philosophy of science* (pp. 108–127). Blackwell Publishers, Inc.

Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008). The effect of presence on human-robot interaction. *Proceedings of the RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 701–706.

Barber, D. J., Abich IV, J., Phillips, E., Talone, A. B., Jentsch, F., & Hill, S. G. (2015). Field assessment of multimodal communication for dismounted human-robot teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*, 921–925.

Barber, D. J., Abrich IV, J., Talone, A. B., Phillips, E., Jentsch, F., Pettitt, R., & Elliott, L. R. (2018). *Soldier-robot team communication: An investigation of exogenous orienting visual display cues and robot reporting preferences*. U.S. Army Research Laboratory. https://apps.dtic.mil/dtic/tr/fulltext/u2/1048418.pdf

Barber, D. J., Howard, T. M., & Walter, M. R. (2016). A multimodal interface for real-time soldier-robot teaming. *Proceedings of SPIE Unmanned Systems Technology XVIII*, *9837*, 1–12.

Barber, D. J., Reinerman-Jones, L. E., & Matthews, G. (2015). Toward a tactile language for human–robot interaction: Two studies of tacton learning and performance. *Human Factors*, *57*(3), 471–490.

Barber, D., Reinerman-Jones, L., & Hidalgo, M. (2019). Optimizing military human-robot teaming: An evaluation of task load and modality switch cost. *Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society*, 1751–1755.

Barber, Daniel. (2018). Assessing Multimodal Interactions with Mixed-Initiative Teams. *Proceedings of the International Conference on Human Interface and the Management of Information*, 175–184.

Barnes, M., Elliott, L. R., Wright, J., Scharine, A., & Chen, J. (2019). *Human–robot interaction design research: From teleoperations to human-agent teaming* (p. 54). CCDC Army Research Laboratory. https://apps.dtic.mil/dtic/tr/fulltext/u2/1079789.pdf

Barnes, Michael J, Chen, J. Y. C., & Hill, S. (2017). *Humans and autonomy: Implications of shared decision-making for military operations* (p. 42). U.S. Army Research Laboratory. https://apps.dtic.mil/dtic/tr/fulltext/u2/1024840.pdf

Barnes, Michael J, Lakhmani, S., Holder, E., & Chen, J. (2019). *Issues in human–agent communication* (p. 23). U.S. Army Research Laboratory. https://apps.dtic.mil/dtic/tr/fulltext/u2/1067050.pdf

Barnes, M.J., & Evans, A. W. (2010). Soldier-robot teams in future battlefields: An overview. In F. Jentsch & M. J. Barnes (Eds.), *Human-robot interactions in future military operations* (pp. 9–29). CRC Press.

Barnlund, D. C. (1979). A transactional model of communication. In C. D. Mortensen (Ed.), *Basic readings in communication theory* (2nd ed., pp. 47–57). Harper & Row. https://doi.org/10.4324/9781315080918-5

Bendell, R., Vasquez, G., Nguyen, B., Barber, D., & Jentsch, F. (2020). Designing naturalistic communication interfaces for human-robot teams: Report modalities and transmission schedules. *Unpublished Manuscript, University of Central Florida, FL*.

Bendell, Rhyse, Vasquez, G., & Jentsch, F. (2019a). Multiple resource loading and auditory preemption during a continuous signal detection task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 1244–1248. https://doi.org/10.1177/1071181319631478

Bendell, Rhyse, Vasquez, G., & Jentsch, F. (2019b). The influence of signal presentation factors on performance of an immersive, continuous signal detection task. In D. N. Cassenti (Ed.), *Advances in human factors and simulation* (pp. 37–48). Springer.

Berlo, D. K. (1960). *Communication: An introduction to theory and practice*. Holt, Rinehart and Winston, Inc.

Billings, D. R., Schaefer, K. E., Chen, J. Y., Kocsis, V., Barrera, M., Cook, J., Ferrer, M., & Hancock, P. A. (2012). *Human-animal trust as an analog for human-robot trust: A review of current evidence* (p. 36). U.S. Army Research Laboratory. https://apps.dtic.mil/dtic/tr/fulltext/u2/a559369.pdf

Bradshaw, J. M., Dignum, V., Jonker, C., & Sierhuis, M. (2012). Human-agent-robot teamwork. *IEEE Intelligent Systems*, *27*(2), 8–13. https://doi.org/10.1109/MIS.2012.37

Bradshaw, Jeffrey M, Acquisti, A., Allen, J., Breedy, M. R., Bunch, L., Chambers, N., Feltovich, P., Galescu, L., Goodrich, M. A., & Jeffers, R. (2004). Teamwork-centered autonomy for extended human-agent interaction in space applications. *Proceedings of the AAAI 2004 Spring Symposium*, 22–24.

Breazeal, C. (2004). Social interactions in HRI: The robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *34*(2), 181–186. https://doi.org/10.1109/TSMCC.2004.826268

Breazeal, Cynthia, & Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots*, *12*(1), 83–104.

Breazeal, Cynthia, Dautenhahn, K., & Kanda, T. (2016). Social robotics. In B. Siciliano & O. Khatib (Eds.), *Springer handbook of robotics* (pp. 1935–1972). Springer. https://doi.org/10.1007/978-3-319-32552-1_72

Breazeal, Cynthia, Gray, J., & Berin, M. (2010). Mindreading as a foundational skill for socially intelligent robots. In M. Kaneko & Y. Nakamura (Eds.), *Robotics research* (Vol. 66, pp. 383–394). Springer. https://doi.org/10.1007/978-3-642-14743-2_32

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, *42*(3), 361–377.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, *114*(3), 542–551.

Bustos, C., & Countinho, F. (2019). *Package dominanceanalysis* (Version 1.2.0) [R package]. https://cran.r-project.org/web/packages/dominanceanalysis/dominanceanalysis.pdf

Cannon-Bowers, J., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In N. J. Castellan Jr. (Ed.), *Individual and group decision making: Current issues* (pp. 221–246). Erlbaum.

Carpenter, J. (2016). *Culture and human-robot interaction in militarized spaces: A war story*. Routledge. https://doi.org/10.4324/9781315562698

Cartwright, N. (1983). *How the laws of physics lie*. Oxford University Press. https://doi.org/10.1093/0198247044.001.0001

Chandarana, M., Meszaros, E. L., Trujillo, A., & Allen, B. D. (2017). Natural language based multimodal interface for UAV mission planning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*, 68–72.

Che, J., Yang, Y., Li, L., Bai, X., Zhang, S., & Deng, C. (2017). Maximum relevance minimum common redundancy feature selection for nonlinear data. *Information Sciences*, *409–410*, 68–86. https://doi.org/10.1016/j.ins.2017.05.013

Chen, J. Y. C. (2010). UAV-guided navigation for ground robot tele-operation in a military reconnaissance environment. *Ergonomics*, *53*(8), 940–950. https://doi.org/10.1080/00140139.2010.500404

Chen, J. Y. C., & Barnes, M. J. (2012). Supervisory control of multiple robots: Effects of imperfect automation and individual differences. *Human Factors*, *54*(2), 157–174.

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. J. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, *19*(3), 259–282.

Chen, J. Y. C., & Terrence, P. I. (2009). Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics*, *52*(8), 907–920. https://doi.org/10.1080/00140130802680773

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency* (pp. 1–30). U.S. Army Research Laboratory. https://apps.dtic.mil/dtic/tr/fulltext/u2/a600351.pdf

Childers, M., Lennon, C., Bodt, B., Pusey, J., Hill, S., Camden, R., Oh, J., Dean, R., Keegan, T., & Diberardino, C. (2016). *US army research laboratory (ARL) robotics collaborative technology alliance 2014 capstone experiment* (p. 62). U.S. Army Research Laboratory. https://apps.dtic.mil/dtic/tr/fulltext/u2/1012767.pdf

Cooke, N. J. (2015). Team cognition as interaction. *Current Directions in Psychological Science*, *24*(6), 415–419. https://doi.org/10.1177/0963721415602474

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science*, *37*(2), 255–285.

Costa Jr, P. T., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality; Volume 2 Personality measurement and testing* (pp. 179–198). Sage Publications, Inc. http://dx.doi.org/10.4135/9781849200479.n9

Cox, D. R., & Snell, E. J. (2018). *Analysis of binary data* (2nd ed.). Routledge. https://doi.org/10.1201/9781315137391

Cribari-Neto, F., & Zeileis, A. (2010). *Beta regression in R* (No. 98; Research Report Series, p. 24). University of Economics and Business Department of Statistics. http://www.jstatsoft.org/v34/i02/

Cuevas, H. M., Fiore, S. M., Caidwell, B. S., & Strater, L. (2007). Augmenting team cognition in human-automation teams performing in complex operational environments. *Aviation, Space, and Environmental Medicine*, *78*(5), B63–B70.

Dambacher, M., & Hübner, R. (2013). Investigating the speed-accuracy trade-off: Better use deadlines or response signals? *Behavior Research Methods, 45*(3), 702-717.

Davis, N. (2016). What is the fourth industrial revolution? *World Economic Forum*. https://www.weforum.org/agenda/2016/01/what-is-the-fourth-industrial-revolution/

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2019). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-019-00596-x

de Winter, J. C. F. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective. *Cognition, Technology & Work, 16*(3), 289–297. https://doi.org/10.1007/s10111-014-0275-1

de Winter, J. C. F., & Dodou, D. (2014). Why the Fitts list has persisted throughout the history of function allocation. *Cognition, Technology & Work, 16*(1), 1–11. https://doi.org/10.1007/s10111-011-0188-1

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: A meta-analysis. *Journal of Applied Psychology, 95*(1), 32.

Demir, M., McNeese, N. J., & Cooke, N. J. (2016). Team communication behaviors of the human-automation teaming. *Proceedings of the IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 28–34. https://doi.org/10.1109/COGSIMA.2016.7497782

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*(1), 417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221

Dillard, M. B., Warm, J. S., Funke, G. J., Funke, M. E., Finomore Jr, V. S., Matthews, G., Shaw, T. H., & Parasuraman, R. (2014). The sustained attention to response task (SART) does not promote mindlessness during vigilance performance. *Human Factors, 56*(8), 1364–1379.

Dobrišek, S., Gajšek, R., Mihelič, F., Pavešić, N., & Štruc, V. (2013). Towards efficient multi-modal emotion recognition. *International Journal of Advanced Robotic Systems*, (*10*), 53.

D’Orazio, D. (2015). Valve’s VR headset is called the Vive and it’s made by HTC. *The Verge*. https://www.theverge.com/2015/3/1/8127445/htc-vive-valve-vr-headset

Driskell, J. E., Salas, E., & Driskell, T. (2018). Foundations of teamwork and collaboration. *American Psychologist*, *73*(4), 334–348. https://doi.org/10.1037/amp0000241

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3–4), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3

Dumas, B., Lalanne, D., & Oviatt, S. (2009). Multimodal interfaces: A survey of principles, models and frameworks. In L. Denis & J. Kohlas (Eds.), *Human machine interaction* (pp. 3–26). Springer.

Efron, B. (1981). Nonparametric estimates of standard error: The Jackknife, the bootstrap and other methods. *Biometrika*, *68*(3), 589–599. https://doi.org/10.2307/2335441

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 54–75.

Endsley, Mica R. (1988). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society Annual Meeting*, *32*, 97–101.

Endsley, Mica R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64.

Endsley, M.R., & Jones, W. M. (2001). A model of inter- and intrateam situation awareness: Implications for design, training and measurement. In M. McNeese, E. Salas, & M. Endsley (Eds.), *New trends in cooperative activities: Understanding system dynamics in complex environments* (pp. 46–67). Human Factors and Ergonomics Society.

Epic Games, Inc. (2019). *What is Unreal Engine 4*. https://www.unrealengine.com/en-US/what-is-unreal-engine-4

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Estrella, A. (1998). A new measure of fit for equations with dichotomous dependent variables. *Journal of Business & Economic Statistics*, *16*(2), 198–205.

Feigh, K. M., & Pritchett, A. R. (2014). Requirements for effective function allocation: A critical review. *Journal of Cognitive Engineering and Decision Making*, *8*(1), 23–32. https://doi.org/10.1177/1555343413490945

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. https://doi.org/10.1080/0266476042000214501

Fiore, S. M., Wiltshire, T. J., Lobato, E. J., Jentsch, F. G., Huang, W. H., & Axelrod, B. (2013). Toward understanding social cues and signals in human–robot interaction: Effects of robot gaze and proxemic behavior. *Frontiers in Psychology*, *4*, 859.

Fitbit. (2019). *Fitbit official site for activity trackers & more*. https://www.fitbit.com/home

Fitts, P. M. (1951). *Human engineering for an effective air-navigation and traffic-control system.* (p. 84). National Research Council.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3–4), 143–166. https://doi.org/10.1016/S0921-8890(02)00372-X

Fox, J., & Bouchet-Valat, M. (2019). *Rcmdr: R Commander* (Version version 2.6-1) [R package]. http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/

Friendly, M., Fox, J., & Friendly, M. M. (2017). *Candisc* (Version 0.8-0) [R package]. https://cran.r-project.org/web/packages/candisc/candisc.pdf

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Giere, R. N. (1988). *Explaining science: A cognitive approach*. University of Chicago Press.

Góngora Alonso, S., Hamrioui, S., de la Torre Díez, I., Motta Cruz, E., López-Coronado, M., & Franco, M. (2018). Social robots for people with aging and dementia: A systematic review of literature. *Telemedicine and E-Health*, *25*(7), 533–540. https://doi.org/10.1089/tmj.2018.0051

Grammarly. (2019). *Grammarly: Free writing assistant*. https://www.grammarly.com/

Green, G. K. (1993). Meta-analysis of multiple-task performances: Cumulating the first two decades of research findings across studies. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *37*(17), 1147–1151. https://doi.org/10.1177/154193129303701705

Grier, R. A., Warm, J. S., Dember, W. N., Matthews, G., Galinsky, T. L., Szalma, J. L., & Parasuraman, R. (2003). The vigilance decrement reflects limitations in effortful attention, not mindlessness. *Human Factors*, *45*(3), 349–359.

Grosz, B. J. (1996). Collaborative systems. *AI Magazine*, *17*(2), 67–67.

Guznov, S., Nelson, A., Lyons, J., & Dycus, D. (2015). The effects of automation reliability and multi-tasking on trust and reliance in a simulated unmanned system control task. In C. Stephanidis (Ed.), *HCI international 2015 communications in computer and information science* (Vol. 529, pp. 616–621). Springer.

Hancock, P. A. (2017). Imposing limits on autonomous systems. *Ergonomics*, *60*(2), 284–291.

Hancock, P.A. (2009). *Mind, machine and morality: Toward a philosophy of human-technology symbiosis*. CRC Press.

Hancock, P.A., & Matthews, G. (2019). Workload and performance: Associations, insensitivities, and dissociations. *Human Factors, 61*(3), 374–392.

Hancock, P.A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527.

Hancock, P.A., & Warm, J. (1989). A dynamic model of stress and sustained attention. *Human Factors*, *31*(5), 519–537.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society*, *159*(3), 445–473. https://doi.org/10.2307/2983326

Hanna, N., & Richards, D. (2018). The impact of multimodal communication on a shared mental model, trust, and commitment in human–intelligent virtual agent teams. *Multimodal Technologies and Interaction*, *2*(3), 48. https://doi.org/10.3390/mti2030048

Harrell Jr, F. E. (2019). *Hmisc* (Version 4.3-1) [R package]. https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf

Harris, J., & Barber, D. (2014). Speech and gesture interfaces for squad-level human-robot teaming. *Proceedings Unmanned Systems Technology XVI*, *9084*, B1–B12. https://doi.org/DOI - 10.1117/12.2052961

Harris, J., & Barber, D. (2014). *Speech and gesture interfaces for squad-level human-robot teaming*. *9084*, 90840B.

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

Headquarters Department of the Army. (2006). *The infantry battalion*. Headquarters Department of the Army. https://fas.org/irp/doddir/army/fm3-21-20.pdf

Ho, C., Tan, H. Z., & Spence, C. (2005). Using spatial vibrotactile cues to direct visual attention in driving scenes. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(6), 397–412.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology*, *45*(1–3), 73–93.

Hoffman, G., & Breazeal, C. (2004). Collaboration in human-robot teams. *Proceedings of the AIAA 1st Intelligent Systems Technical Conference*, 1–18.

HTC Vive. (2019). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=HTC_Vive&oldid=907041262

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

Hyde, J. S. (1990). Meta-analysis and the psychology of gender differences. *Signs: Journal of Women in Culture and Society*, *16*(1), 55–73.

Imenda, S. (2014). Is there a conceptual difference between theoretical and conceptual frameworks? *Journal of Social Sciences*, *32*(2), 185-195.

Institute for Digital Research & Education Statistical Consulting. (2011, October 20). *FAQ: What are pseudo R-squareds?* https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/

Jackson, J. J., Thoemmes, F., Jonkmann, K., Lüdtke, O., & Trautwein, U. (2012). Military training and personality trait development: Does the military make the man, or does the man make the military? *Psychological Science*, *23*(3), 270–277.

Jiang, S., & Arkin, R. C. (2015). Mixed-initiative human-robot interaction: Definition, taxonomy, and survey. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, 954–961. https://doi.org/10.1109/SMC.2015.174

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Jonker, C. M., Van Riemsdijk, M. B., & Vermeulen, B. (2010). Shared mental models. In M. de Vos, N. Fornara, J. V. Pitt, & G. Vouros (Eds.), *Coordination, organizations, institutions, and norms in agent systems VI* (Vol. 6054, pp. 132–151). Springer.

Jordan, N. (1963). Allocation of functions between man and machines in automated systems. *Journal of Applied Psychology*, *47*(3), 161–165.

Kaber, D. B. (2017). A conceptual framework of autonomous and automated agents. *Theoretical Issues in Ergonomics Science*, 1–25.

Keebler, J. R., Jentsch, F., Fincannon, T., & Hudson, I. (2012). Applying team heuristics to future human-robot systems. *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 169–170.

Kiesler, S., & Goetz, J. (2002). Mental models and cooperation with robotic assistants. *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, 576–584.

Kim, S., Miller, M. E., Rusnock, C. F., & Elshaw, J. J. (2018). Spatialized audio improves call sign recognition during multi-aircraft control. *Applied Ergonomics*, *70*, 51–58.

Klein, G., Feltovich, P. J., Bradshaw, J. M., & Woods, D. D. (2005). Common ground and coordination in joint activity. *Organizational Simulation*, *53*, 139–184.

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

Kopinsky, R. J. (2017). *Novel mixed reality interface for effective and efficient human robot interaction with unique mobility platforms* [Dissertation, Florida State University]. https://fsu.digital.flvc.org/islandora/object/fsu%3A552323/

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer Science & Business Media. https://link.springer.com/content/pdf/10.1007/978-1-4614-6849-3.pdf

Lackey, S., Barber, D., Reinerman-Jones, L., Badler, N. I., & Hudson, I. (2011). Defining next-generation multi-modal communication in human robot interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*, 461–464.

Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable agency for intelligent autonomous systems. *Proceedings of the Twenty-Ninth AAAI Conference on Innovative Applications*, 4762–4764.

Latorella, K. A. (1998). Effects of modality on interrupted flight deck performance: Implications for data link. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 42*, 87–91.

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer Publishing Company.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Lee, S., Lau, I. Y., Kiesler, S., & Chiu, C.-Y. (2005). Human mental models of humanoid robots. *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, 2767–2772.

Lewis, J. E., & Neider, M. B. (2016). Through the Google Glass: The impact of heads-up displays on visual attention. *Cognitive Research: Principles and Implications*, *1*(1), 13. https://doi.org/10.1186/s41235-016-0015-6

Lin, J., Wohleber, R., Matthews, G., Chiu, P., Calhoun, G., Ruff, H., & Funke, G. (2015). Video game experience and gender as predictors of performance and stress during supervisory control of multiple unmanned aerial vehicles. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*, 746–750.

Lips-Wiersma, M., Haar, J., & Wright, S. (2018). The effect of fairness, responsible Leadership and worthy work on multiple dimensions of meaningful work. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-018-3967-2

Lu, S. A., Wickens, C. D., Prinet, J. C., Hutchins, S. D., Sarter, N., & Sebok, A. (2013). Supporting interruption management and multimodal interface design: Three meta-analyses of task performance as a function of interrupting task modality. *Human Factors*, *55*(4), 697–724.

Luo, W., & Azen, R. (2013). Determining predictor importance in hierarchical linear models using dominance analysis. *Journal of Educational and Behavioral Statistics*, *38*(1), 3–31.

Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. *AAAI Spring Symposium Series*, 48–54.

Lyons, J. B., & Havig, P. R. (2014). Transparency in a human-machine context: Approaches for fostering shared awareness/intent. In R. Shumaker & S. Lackey (Eds.), *Virtual, augmented and mixed reality: Designing and developing virtual and augmented environments* (pp. 181–190). Springer International Publishing. https://doi.org/10.1007/978-3-319-07458-0_18

Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., Smith, D., Johnson, W., & Shively, R. (2017). Shaping trust through transparent design: Theoretical and experimental guidelines. In P. Savage-Knepshield & J. Chen (Eds.), *Advances in human factors in robots and unmanned systems* (Vol. 499, pp. 127–136). Springer International Publishing. https://doi.org/10.1007/978-3-319-41959-6_11

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*(1), 6–21.

Maeda, Y., & Yoon, S. Y. (2013). A meta-analysis on gender differences in mental rotation ability measured by the Purdue Spatial Visualization Tests: Visualization of rotations (PSVT:R). *Educational Psychology Review*, *25*(1), 69–94. https://doi.org/10.1007/s10648-012-9215-x

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, *85*(2), 273–283.

Matthews, G., de Winter, J., & Hancock, P. A. (2019). What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical Issues in Ergonomics Science*, 1–31. https://doi.org/10.1080/1463922X.2018.1547459

Matthews, G., & Campbell, S. E. (2009). Sustained performance under overload: Personality and individual differences in stress and coping. *Theoretical Issues in Ergonomics Science*, *10*(5), 417–442. https://doi.org/10.1080/14639220903106395

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

Matthews, G., Davies, D. R., Stammers, R. B., & Westerman, S. J. (2000). *Human performance: Cognition, stress, and individual differences*. Psychology Press.

Matthews, G., Deary, I. J., & Whiteman, M. C. (2003). *Personality traits* (2nd edition). Cambridge University Press.

Matthews, G., & Reinerman-Jones, L. (2017). *Workload assessment: How to diagnose workload issues and enhance performance*. Human Factors and Ergonomics Society.

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

Matthews, G., Wohleber, R. W., Lin, J., Matthews, G., Wohleber, R. W., & Lin, J. (2019). Stress, skilled performance, and expertise: Overload and beyond. In P. Ward, J. M. Schaagen, J. Gore, & E. M. Roth (Eds.), *The Oxford handbook of expertise* (pp. 1–39). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198795872.013.22

Maurtua, I., Fernandez, I., Tellaeche, A., Kildal, J., Susperregi, L., Ibarguren, A., & Sierra, B. (2017). Natural multimodal communication for human-robot collaboration. *International Journal of Advanced Robotic Systems*, *14*(4), 1–12. https://doi.org/10.1177/1729881417716043

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 104–142). Academic Press.

Mehrabian, A. (1979). Communication without words. In C. D. Mortensen (Ed.), *Basic readings in communication theory* (2nd ed., pp. 193–208). Harper & Row. https://doi.org/10.4324/9781315080918-5

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human Factors*, *58*(3), 401–415.

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2015). Intelligent agent transparency in human-agent teaming for multi-UxV management. *Human Factors*, *58*(3), 401–415. https://doi.org/10.1177/0018720815621206

Merriam-Webster. (2019). www.merriam-webster.com

Mesmer-Magnus, J., Niler, A. A., Plummer, G., Larson, L. E., & DeChurch, L. A. (2017). The cognitive underpinnings of effective teamwork: A continuation. *Career Development International*, *22*(5), 507–519.

Microsoft. (2019). *Microsoft Speech Platform—SDK* (Version 11) [Computer software]. Microsoft. https://www.microsoft.com/en-us/download/details.aspx?id=27226

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley. *IEEE Robotics Automation Magazine*, *19*(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Morris, N. M., & Rouse, W. B. (1986). *Adaptive aiding for human-computer control: Experimental studies of dynamic task allocation*. Harry G. Armstrong Aerospace Medical Research Laboratory. https://apps.dtic.mil/dtic/tr/fulltext/u2/a166704.pdf

Morrow, D., Leirer, V., Altiteri, P., & Fitzsimmons, C. (1994). When expertise reduces age differences in performance. *Psychology and Aging*, *9*(1), 134–148. https://doi.org/10.1037/0882-7974.9.1.134

Mostafa, S. A., Ahmad, M. S., & Mustapha, A. (2019). Adjustable autonomy: A systematic literature review. *Artificial Intelligence Review*, *51*(2), 149–186. https://doi.org/10.1007/s10462-017-9560-8

Murphy, R. R. (2004). Human-robot interaction in rescue robotics. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34*(2), 138–153.

Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., & Hagita, N. (2009). Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 69–76. https://doi.org/10.1145/1514095.1514110

Mutlu, B., Roy, N., & Šabanović, S. (2016). Cognitive human–robot interaction. In B. Siciliano & O. Khatib (Eds.), *Springer Handbook of Robotics* (pp. 1907–1934). Springer International Publishing. https://doi.org/10.1007/978-3-319-32552-1_71

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.

Naval Education and Training Command. (2009). *Task-based curriculum development manual 130B Volume I* (NAVEDTRA 130-B). Naval Education and Training Command. https://www.public.navy.mil/netc/ile/documents/NAVEDTRA130B/NAVEDTRA_130B_(Vol_I).pdf

Negretti, R. (2012). Metacognition in student academic writing: A longitudinal study of metacognitive awareness and its relation to task perception, self-regulation, and evaluation of performance. *Written Communication*, *29*(2), 142–179.

Neubauer, C., Dillard, M. B., Warm, J. S., Funke, G. J., Funke, M., Matthews, G., Epling, S. L., Dukes, A. W., Force, W.-P. A., & Base, O. (2013). Effects of event rate on cerebral blood flow velocity during vigilance performance. *Proceedings of International Symposium on Aviation Psychology*, 609–614.

Nikolaidis, S., & Shah, J. (2012). Human-robot teaming using shared mental models. *ACM IEEE*, 1–6.

Nolan, E., & Santos, P. (2019). Genetic modification and yield risk: A stochastic dominance analysis of corn in the USA. *PLoS ONE*, *14*(10). https://doi.org/10.1371/journal.pone.0222156

Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. Basic Books.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff & J. P. Thomas (Eds.), *Handbook of perception and human performance* (2nd ed., Vol. 2, pp. 1–49). Wiley.

O'Neill, T. A., & Allen, N. J. (2011). Personality and the prediction of team performance. *European Journal of Personality*, *25*(1), 31–42. https://doi.org/10.1002/per.769

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, *56*(3), 476–488.

Ososky, S., Schuster, D., Jentsch, F., Fiore, S., Shumaker, R., Lebiere, C., Kurup, U., Oh, J., & Stentz, A. (2012). The importance of shared mental models and shared situation awareness for transforming robots from tools to teammates. *Proceedings of SPIE Unmanned Systems Technology XIV*, *8387*, 838710:1-838710–838712.

Oviatt, S. (2012). Multimodal interfaces. In J. A. Jacko (Ed.), *The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications* (3rd ed., pp. 405–429). CRC Press: Taylor & Francis Group.

Pachella, R.G., & Pew, R.W. (1968). Speed-accuracy tradeoff in reaction time: Effect of discrete criterion times. *Journal of Experimental Psychology, 76*(1), 19-24.

Pan, K., Zhang, Y., Wang, Y., Wang, Y., & Xu, H. (2016). A systematic review and meta-analysis of conventional laparoscopic sacrocolpopexy versus robot-assisted laparoscopic sacrocolpopexy. *International Journal of Gynecology & Obstetrics*, *132*(3), 284–291.

Parasuraman, R., & Davies, D. (1977). A taxonomic analysis of vigilance performance. In R. R. Mackie (Ed.), *Vigilance* (Vol. 3, pp. 559–574). Springer.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 30*(3), 286–297.

Pedhazur, E. J. (1973). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Harcourt Brace & Company.

Peeters, M. A. G., Tuijl, H. F. J. M. van, Rutte, C. G., & Reymen, I. M. M. J. (2006). Personality and team performance: A meta-analysis. *European Journal of Personality*, *20*(5), 377–396. https://doi.org/10.1002/per.588

Phillips, E., Ososky, S., Grove, J., & Jentsch, F. (2011). From tools to teammates: Toward the development of appropriate mental models for intelligent robots. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*, 1491–1495.

Powers, A., & Kiesler, S. (2006). The advisor robot: Tracing people's mental model from a robot's physical attributes. *Proceeding of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction - HRI '06*, 218–225. https://doi.org/10.1145/1121241.1121280

Prewett, M. S., Elliott, L. R., Walvoord, A. G., & Coovert, M. D. (2012). A meta-analysis of vibrotactile and visual information displays for improving task performance. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *42*(1), 123–132.

Quine, W. V., & Ullian, J. S. (1998). Hypothesis. In E. D. Klemke, R. Hollinger, & D. W. Rudge (Eds.), *Introductory readings in the philosophy of science* (pp. 404–414). Prom.

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

Rahimi, M., & Hancock, P. (1986). Optimization of hybrid production systems: The integration of robots into human-occupied work environments. In O. Brown Jr. & H. Hendrick (Eds.), *Human factors in organizational design and management II* (pp. 39–54). Elsevier.

Reinerman-Jones, L, Barber, D., Szalma, J., & Hancock, P. (2017). Human interaction with robotic systems: Performance and workload evaluations. *Ergonomics, 60*(10), 1351–1368.

Reinerman-Jones, L., Taylor, G., Sprouse, K., Barber, D., & Hudson, I. (2011). Adaptive automation as a task switching and task congruence challenge. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 55*, 197–201.

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., & Robinson, P. (2009). How anthropomorphism affects empathy toward robots. *Proceedings of the ACM/IEEE International Conference on Human Robot Interaction*, 245–246. https://doi.org/10.1145/1514095.1514158

Ross, J. M., Szalma, J. L., Hancock, P. A., Barnett, J. S., & Taylor, G. (2008). The effect of automation reliability on user automation trust and reliance in a search-and-rescue scenario. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 52*(19), 1340–1344. https://doi.org/10.1177/154193120805201908

Rouse, W.B. (1994). Twenty years of adaptive aiding: Origins of the concept and lessons learned. In M. Mouloua & Parasuraman, R (Eds.), *Human performance in automated systems: Current research and trends* (pp. 28–32). Erlbaum.

Rouse, W. & Morris, N. M. (1986). *On looking into the black box: Prospects and limits in the search for mental models.* (p. 61). Center for Man-Machine Systems Research. https://apps.dtic.mil/dtic/tr/fulltext/u2/a159080.pdf

Russell, S. J., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd ed.). Upper Prentice Hall.

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a “big five” in teamwork? *Small Group Research*, *36*(5), 555–599.

Sanders, T. L., Wixon, T., Schafer, K. E., Chen, J. Y., & Hancock, P. (2014). The influence of modality and transparency on trust in human-robot interaction. *Proceedings of the IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 156–159.

Savela, N., Turja, T., & Oksanen, A. (2018). Social acceptance of robots in different occupational fields: A systematic literature review. *International Journal of Social Robotics*, *10*(4), 493–502. https://doi.org/10.1007/s12369-017-0452-5

Sawyer, B. (2015). *Effects of signal probability on multitasking-based distraction in driving, cyberattack & battlefield simulation* [Dissertation]. University of Central Florida.

Sawyer, B. D., Finomore, V. S., Funke, G. J., Mancuso, V. F., Funke, M. E., Matthews, G., & Warm, J. S. (2014). Cyber vigilance: Effects of signal probability and event rate. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*, 1771–1775.

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, *12*(1), 13–24.

Scassellati, B. M. (2001). *Foundations for a theory of mind for a humanoid robot* [Dissertation]. Massachusetts Institute of Technology.

Scerbo, M. W. (2007). Adaptive automation. In R. Parasuraman & M. Rizzo (Eds.), *Neuroergonomics: The brain at work* (pp. 239–252). Oxford University Press.

Schaefer, K. E., Billings, D. R., & Hancock, P. A. (2012). Robots vs. machines: Identifying user perceptions and classifications. *Proceedings of the IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, 138–141. https://doi.org/10.1109/CogSIMA.2012.6188366

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schaefer, K. E., Hill, S. G., & Jentsch, F. G. (2019). Trust in human-autonomy teaming: A review of trust research from the US Army Research Laboratory Robotics Collaborative Technology Alliance. In J. Chen (Ed.), *Advances in human factors in robots and unmanned systems* (pp. 102–114). Springer International Publishing. https://doi.org/10.1007/978-3-319-94346-6_10

Schaefer, K. E., Straub, E. R., Chen, J. Y. C., Putney, J., & Evans, A. W. (2017). Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research*, *46*, 26–39. https://doi.org/10.1016/j.cogsys.2017.02.002

Scheutz, M., DeLoach, S. A., & Adams, J. A. (2017). A framework for developing and using shared mental models in Human-Agent Teams. *Journal of Cognitive Engineering and Decision Making, 11*(3), 203–224. https://doi.org/10.1177/1555343416682891

Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Larmarange, J. (2017). *GGally: Extension to 'ggplot2'(R Package Version 1.3. 1)*.

Scholtz, J. (2003). Theory and evaluation of human robot interactions. *Proceedings of the Annual International Conference on System Sciences*, 1–10. https://doi.org/10.1109/HICSS.2003.1174284

Schramm, W. (1954). How communication works. In W. Schramm (Ed.), *The process and effects of mass communication* (pp. 3–26). University of Illinois Press.

Schwab, K., & Davis, N. (2018). *Shaping the future of the fourth industrial revolution*. Currency.

Sebok, A., & Wickens, C. D. (2017). Implementing lumberjacks and black swans into model-based tools to support human–automation interaction. *Human Factors*, *59*(2), 189–203.

See, J. E., Howe, S. R., Warm, J. S., & Dember, W. N. (1995). Meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin*, *117*(2), 230.

Selkowitz, A. R., Lakhmani, S. G., & Chen, J. Y. (2017). Using agent transparency to support situation awareness of the Autonomous Squad Member. *Cognitive Systems Research*, *46*, 13–25.

Shah, P., Livingston, L. A., Callan, M. J., & Player, L. (2019). Trait autism is a better predictor of empathy than alexithymia. *Journal of Autism and Developmental Disorders*, 1–9.

Shakarchi, A. F., Mihailovic, A., West, S. K., Friedman, D. S., & Ramulu, P. Y. (2019). Vision parameters most important to functionality in glaucoma. *Investigative Ophthalmology & Visual Science*, *60*(14), 4556–4563. https://doi.org/10.1167/iovs.19-28023

Sheridan, T. B., & Parasuraman, R. (2005). Human-Automation Interaction. *Reviews of Human Factors and Ergonomics*, *1*(1), 89–129. https://doi.org/10.1518/155723405783703082

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (p. 186). Massachusetts Institute of Technology: Man-Machine Systems Lab.

Shou, Y., & Smithson, M. (2015). Evaluating predictors of dispersion: A comparison of dominance analysis and Bayesian model averaging. *Psychometrika*, *80*(1), 236–256.

Sims, V. K., Chin, M. G., Sushil, D. J., Barber, D. J., Ballion, T., Clark, B. R., Garfield, K. A., Dolezal, M. J., Shumaker, R., & Finkelstein, N. (2005).

Anthropomorphism of robotic forms: A response to affordances? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(3), 602–605.

Singer, P.W. (2009). Wired for war: The future of military robots. *Brookings*. https://www.brookings.edu/opinions/wired-for-war-the-future-of-military-robots/

Singer, P.W. (2012). The robotics revolution. *Brookings*. https://www.brookings.edu/opinions/the-robotics-revolution/

Smith, R. L., Ager Jr, J. W., & Williams, D. L. (1992). Suppressor variables in multiple regression/correlation. *Educational and Psychological Measurement*, *52*(1), 17–29.

Smithson, M., & Merkle, E. C. (2013). *Generalized linear models for categorical and continuous limited dependent variables*. CRC Press, Taylor & Francis Group.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71.

Springer, P. J. (2013). *Military robots and drones: A reference handbook*. ABC-CLIO.

Stanton, N. A., Salmon, P. M., Rafferty, L. A., Walker, G. H., Baber, C., Jenkins, D. P., Salmon, P. M., Rafferty, L. A., Walker, G. H., Baber, C., & Jenkins, D. P. (2017). *Human factors methods: A practical guide for engineering and design* (2nd ed.). CRC Press. https://doi.org/10.1201/9781315587394

Steamworks. (2019). *SteamVR Tracking*. https://partner.steamgames.com/vrlicensing

Sukthankar, G., Shumaker, R., & Lewis, M. (2012). Intelligent agents as teammates. In E. Salas, S. M. Fiore, & M. P. Letsky (Eds.), *Theories of team cognition: Cross-disciplinary perspectives* (pp. 313–343). Routledge Taylor & Francis Group.

Suppes, P. (1967). What is a scientific theory? In S. Morgenbesser (Ed.), *Philosophy of science today* (pp. 55–67). Basic Books.

Sutherland, J., Baillergeon, R., & McKane, T. (2010). Cordon and search operations: A deadly game of hide and seek. *Air Land Sea Bulletin*, *3*, 4–10.

Szalma, J., Hancock, P., & Quinn, S. (2008). A meta-analysis of the effect of time pressure on human performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*, 1513–1516.

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The five factor model of personality. *Journal of Experimental Psychology: Applied*, *17*(2), 71–96. https://doi.org/10.1037/a0024170

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics* (6th edition). Pearson.

Talone, A. B., Phillips, E., Ososky, S., & Jentsch, F. (2015). An evaluation of human mental models of tactical robot movement. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*, 1558–1562.

Taylor, G. S., Reinerman-Jones, L. E., Szalma, J. L., Mouloua, M., & Hancock, P. A. (2013). What to automate: Addressing the multidimensionality of cognitive resources through system design. *Journal of Cognitive Engineering and Decision Making*, *7*(4), 311–329.

Teo, G., Reinerman-Jones, L., Hidalgo, M., & Burford, C. (2018). Human-agent teaming: State of assessments and selected issues. *Proceedings of the 2018 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, 1–12.

Thayer, J. F., Åhs, F., Fredrikson, M., Sollers, J. J., & Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, *36*(2), 747–756. https://doi.org/10.1016/j.neubiorev.2011.11.009

Tighe, E. L., & Schatschneider, C. (2014). A dominance analysis approach to determining predictor importance in third, seventh, and tenth grade reading comprehension skills. *Reading and Writing*, *27*(1), 101–127.

Tonidandel, S., & LeBreton, J. M. (2010). Determining the relative importance of predictors in logistic regression: An extension of relative weight analysis. *Organizational Research Methods*, *13*(4), 767–781. https://doi.org/10.1177/1094428109341993

van Fraassen, B. C. (1987). The semantic approach to scientific theories. In N. J. Nersessian (Ed.), *The process of science: Contemporary philosophical approaches to understanding scientific practice* (pp. 105–124). Springer Netherlands. https://doi.org/10.1007/978-94-009-3519-8_6

Vasquez, G., Bendell, R., Talone, A., & Jentsch, F. (2018). Development of a signal fetection-based task for research on distributed human-robot teaming within immersive virtual reality. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*, 1479–1483.

Vidulich, M. A., & Tsang, P. S. (2012). Mental workload and situation awareness. *Handbook of Human Factors and Ergonomics*, *4*, 243–273.

VIVE. (2019). *VIVE Virtual Reality System*. https://www.vive.com/us/product/vive-virtual-reality-system/

Warm, J. S., & Jerison, H. J. (1980). The psychophysics of vigilance. *Proceedings of the Human Factors Society Annual Meeting*, *24*, 605–605.

White, T. L. (2010). *Suitable body locations and vibrotactile cueing types for dismounted soldiers*. Army Research Laboratory.

Wickelgren, W.A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*, 67-85.

Wickens, C.D., Prinet, J., Hutchins, S., Sarter, N., & Sebok, A. (2011). Auditory-visual redundancy in vehicle control interruptions: Two meta-analyses. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*, 1155–1159.

Wickens, C.D., & Hollands, J. G. (2000). *Engineering psychology and human performance* (3rd ed.). Prentice-Hall Inc.

Wickens, C.D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177.

Wickens, C.D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455.

Wickens, C.D., & Dixon, S. R. (2005). *Is there a magic number 7 (to the minus 1)?: The benefits of imperfect diagnostic automation: A synthesis of the literature*. University of Illinois.

Wickens, C.D., Dixon, S. R., & Seppelt, B. (2005). Auditory preemption versus multiple resources: Who wins in interruption management? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 49*, 463–466.

Wickens, C.D., & Harris, D. (2017). Head-up displays. In *Engineering psychology and cognitive ergonomics: Volume 1 Transportation systems* (pp. 3–22).

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Wickham, H.. (2017). *tidyverse: Easily Install and Load the'Tidyverse'* (Version 1.3.0) [R package]. https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf

Wiltshire, T. J., Smith, D. C., & Keebler, J. R. (2013). Cybernetic teams: Towards the implementation of team heuristics in HRI. In R. Shumaker (Ed.), *Virtual augmented and mixed reality: Designing and developing augmented and virtual environments* (Vol. 8021, pp. 321–330). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39405-8_36

Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of Performance and Subjective Measures of Workload. *Human Factors*, *30*(1), 111–120. https://doi.org/10.1177/001872088803000110

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research, 5*, 1205–1224.

Zalta, E. N. (2020). *The Stanford encyclopedia of philosophy*. Stanford University. https://plato.stanford.edu/

Zeileis, A., Cribari-Neto, F., Gruen, B., Kosmidis, I., Simas, A. B., & Rocha, A. V. (2019). *Package betareg* (Version 3.1-2) [R package].