Human-Agent Teaming: State of Assessments and Selected Issues

Grace Teo, Lauren Reinerman-Jones, Maartje Hidalgo
University of Central Florida, Institute for Simulation and Training
Orlando, FL
gteo@ist.ucf.edu, lreinerm@ist.ucf.edu, mhidalgo@ist.ucf.edu

Clayton Burford
Army Research Laboratory
Orlando, FL
clayton.w.burford.civ@mail.mil

ABSTRACT

Progress in computing and robotics technologies has fueled research in Human-Agent Teaming. More than before, robots, machines, and systems are seen as viable agent teammates that work alongside humans as force-multipliers to enhance performance, to ensure safety, and to improve efficiency. However, even as the demand for research and development in HAT by both the military and industry continues to increase, there is a growing concern over the current and foreseeable future challenges with assessments in this relatively new domain. For instance, it is difficult to compare results of assessments that are purported to examine the common HAT construct relationships, such as that between reliability and workload, but use different definitions and measures for the constructs. Moderation of the effects of multiple HAT innovations by contextual factors is also not well understood because many assessments have been done using various tasks and testbeds. All this constrains the extent to which study findings can be generalized, and the establishment of knowledge base about the critical factors and relationships in HAT. In this paper, analyses were conducted on the metadata of 74 HAT research studies from ten researchers from military labs. Results show patterns and trends in the metadata that illustrate which constructs tend to be examined together, which measures seem to cluster, and which constructs had the most and least diverse measures and definitions. Implications of these findings on assessment quality and the utility of assessment outcomes for informing a variety of critical decisions are also discussed.

ABOUT THE AUTHORS

Grace Teo, Ph.D., is a Research Associate at the University of Central Florida's Institute for Simulation and Training. Her research includes work on decision making, individual differences, human-technology interactions, and various assessments of human performance.

Lauren Reinerman-Jones, Ph.D., is the Director of Prodigy, which is one lab at the University of Central Florida's Institute for Simulation and Training, focusing on assessment for explaining, predicting, and improving human performance and systems.

Maartje Hidalgo, M.S., is a Graduate Research Assistant at the University of Central Florida's Institute for Simulation and Training and an Industrial Engineering and Management Systems (IEMS) doctoral student at the University of Central Florida with a special interest in neuroergonomics and performance optimization.

Clayton Burford, M.S., is a Science and Technology Manager with the U.S. Army Research Laboratory (ARL) Human Research and Engineering Directorate (HRED), and leads technical teams in the areas of modeling, simulation, and training.

Human-Agent Teaming: State of Assessments and Selected Issues

Grace Teo, Lauren Reinerman-Jones, Maartje Hidalgo
University of Central Florida, Institute for Simulation and Training
Orlando, FL
gteo@ist.ucf.edu, lreinerm@ist.ucf.edu, mhidalgo@ist.ucf.edu

Clayton Burford
Army Research Laboratory
Orlando, FL
clayton.w.burford.civ@mail.mil

INTRODUCTION

With emerging computing and robotics technologies, the domain of human-agent teaming (HAT) has seen tremendous growth, both in the military and in industry. Due to their increasing intelligence, systems and machines are acquiring behaviors that render them more like agents possessing a level of autonomy (Franklin & Graesser, 1996) in certain tasks, and some are slowly moving into the role of a teammate or peer (Steinfeld et al., 2006) to the human operator. However, as research and development on agents and opportunities for HAT grows, several assessment challenges that limit the generalizability of study findings to the real world have been identified. Many assessments are based on constructs that have multiple definitions and operationalizations. Studies investigating the effects in HAT are conducted in a variety of tasks and contexts. These challenges have in part been minimized in domains where strong theories typically describe the relationships among multiple concepts and explain some of the variability in findings. However, there are few of such theories in HAT as the domain is relatively new. As the HAT research continues to be in demand, there is a need to examine the impact of these challenges on efforts to build a research and knowledge base in HAT. By looking into the constructs most assessed in HAT research, how these have been operationalized, and the tasks and contexts used in the studies, we would have an understanding of how divergent HAT research is. Efforts to unify and consolidate research findings can then be undertaken.

Different Operationalizations of the Same Construct

Many common constructs in HAT research have been defined and operationalized differently. One of most widely-assessed constructs in human factors engineering is workload. In the context of HAT, the interest in the construct is in the effects of the agent(s) on workload. The research focus is on the extent to which task support from agents mediates workload, which has been defined as 'the level of attentional resources required to meet both objective and subjective performance criteria, which may be mediated by task demands, external support, and past experience' (Young and Stanton, 2005, chap. 39-1). More recently, definition of workload included "the interaction of multiple contextual variables and a human operator's perception of those variables" (p. 12) (Hooey, Kaber, Adams, Fong, & Gore, 2017). Both definitions illustrate the complexity of the workload construct, and how context-dependent it may be. Workload is a multi-dimensional construct with many different measures, each differing from the others in their sensitivity to distinct, meaningful properties of workload (Matthews, Reinerman-Jones, Barber, & Abich IV, 2015). For instance, eye tracking metrics are specifically sensitive to single- versus dual-tasking, whereas heart rate variability (HRV) and frontal cortex oxygenation (rSO_2) are more indicative of effort (Matthews et al., 2015), and the Multiple Resource Questionnaire (MRQ) (Boles & Adair, 2001) is more sensitive than the NASA-TLX (Hart & Staveland, 1988) to workload that is related to source complexity (Finomore Jr, Shaw, Warm, Matthews, & Boles, 2013). Although having a variety of measures for the same construct is necessary, especially for multi-dimensional constructs, they can lead to difficulties in understanding the relationships and effects concerning the construct. To illustrate, Calhoun, Ruff, Spriggs, and Murray (2012) measured the effects of levels of automation (LOA) on operator workload in an imaging and weapon release task. They utilized an experimenter-developed rating scale to measure workload and found that LOA affected operator workload. When Lin and colleagues (2015) used the same task but a different measure of workload, they found that LOA did not significantly affect workload. Based on these contrasting results, we cannot make any inferences about the effect of LOA on workload, although it is likely that the different measures of the unitary construct are sensitive to the construct in different ways (Matthews et al., 2015). Still, examples like this raise questions on the generalizability of research findings.

Contextual Specificity: Tasks, Contexts, and Testbeds

Another assessment challenge in the HAT domain is the wide range of tasks and contexts in studies that make it difficult to glean an understanding of effects from multiple studies. The same task can yield contrasting results in different contexts as demonstrated in a study on human-computer interactions to inform the design of computer agents (Gal, Grosz, Pfeffer, Shieber, & Allain, 2007). For instance, outcomes on a target detection task performed with an agent can differ substantially depending on context such as whether the target is a threat character in a simulated task or simulated cyber threat, the type of response required in the task (e.g., sustained attention to response task (SART) or traditional vigilance format (TVF); (Dillard et al., 2014), or the testbed used to administer the task. Many HAT studies utilize testbeds to investigate the impact of actual or simulated (i.e., "Wizard of Oz" studies (WoZ); (Maulsby, Greenberg, & Mander, 1993)) system/agent behaviors on the human operator. Testbeds are a major component in assessments in HAT since the type and amount of assessments are constrained by the system capabilities and functionalities of the testbed. One testbed may enable the simulation of a single human to a multiple agent task and support multiple teaming metrics, while another testbed may support the use of physiological assessments during a HAT dyad task. Although a meta-analysis can summarize the effect sizes across studies with different samples, study designs, and results (Borenstein, Hedges, Higgins, & Rothstein, 2011), combining HAT studies that differ on all of these as well on construct operationalizations, tasks, contexts, and testbeds would be challenging for any meta-analyst. It may also not yield useful results especially when there are relatively few theories or taxonomies to guide the analyses. Before meta-analyses can be performed, it is necessary to understand the research that has been conducted on HAT and the state of HAT assessments. Examining the metadata from research studies is a step towards this objective.

Goal

The present work is an analysis of the metadata from studies in the HAT domain. It aims to summarize the state of assessments to obtain an understanding of the knowledge base and research areas within the domain. We described the method for selecting the HAT studies and summarized the research in terms of the constructs most commonly assessed, how they have been operationalized, as well as information about the tasks and contexts used in HAT research. Lastly, we identify gaps and recommendations for future HAT research.

METHOD

Sample of Studies

A total of 74 studies were reviewed. We defined studies to be in the HAT domain if they involved human operator(s) working on task(s) with one or more machines (e.g., robot, agent, unmanned system). The HAT may be a human-agent dyad, as well as different number of humans and agents working in conjunction and/or collaboration. As this study was part of a larger military project, only empirical studies by authors and laboratories from the US Department of Defense (DoD) were included, i.e., United States Air Force (AFRL), Army (AFL), and Navy (NRL). While we recognize that this constraint is an artificial one, we also acknowledge the fact that most HAT studies are directly or indirectly a result of military initiatives given the resources required for such studies, and the military's interest and investment in autonomous systems and HAT. To keep the review current, the literature search only included recent studies from a search on Google Scholar.

Data Extraction

As the purpose of this paper is to uncover trends in HAT assessments, a large amount of data was extracted. The focus was on constructs, and how these were operationalized and measured. Extracted metadata included study title, authors, year, keywords, theory/model, task, testbed, experimental design, experimental conditions, sample (size and resource pool), statistical analysis used, assignment of variables in analysis, operationalization of constructs, and results obtained.

Inconsistencies in Terminology and Creation of Thesaurus

While every attempt was made to preserve the original construct names and terminology reported in the studies, it was necessary to group certain concepts to facilitate discovery of specific trends in the metadata. Within the studies selected, there were multiple inconsistencies in the use of construct names. For example, where one researcher would use the term "task difficulty", another researcher would refer to the same concept as "task load". To a degree, these inconsistencies are indicative of the level of a lack of standardization in HAT assessments. We collated keywords and constructs across the HAT studies and created a thesaurus of synonyms of those keywords and constructs (see Table 1 for an excerpt).

Table 1: Some of the keywords/construct names and synonyms encountered

Keyword/Construct names	"Synonyms"
Human-agent teaming	Human-agent teams; human-robot interaction; human-robot teaming; human-robot teams
Unmanned systems	Unmanned vehicles (UVs); unmanned aerial vehicles (UAVs); unmanned aerial systems (UASs); unmanned ground vehicles (UGVs); remotely piloted aircrafts (RPAs)
Task load	Task difficulty; task demand
Cyber security	Cyber defense
Reliability	Automation reliability
Supervisory control	Multi-robot control
Decision making	Team decision making
Workload	Mental workload; real-time workload; operator workload; cognitive load
Stress	Task-induced stress; stress states

Analysis of Metadata from the Studies

In our summary of the HAT body of research, we focused on identifying the main constructs investigated. For each study, we identified the independent variables (IVs) and the outcomes, or dependent variables (DVs) that they purportedly influenced. We examined how these were measured and operationalized in the various studies. However, since there was little consistency or standardization in the use of construct names and terms, we needed to categorize and group constructs in order to identify research trends. For instance, performance outcome from threat detection and that from target detection were both classified as *detection performance*. In addition, to understand the contexts within which HAT research is conducted, we examined the tasks and testbeds used in the studies. Current HAT research is inevitably shaped by the ability of the testbed to simulate agent behaviors and administer tasks. Determining the capabilities of the testbeds used in current HAT research offers an understanding of the scope of HAT research. The analysis of the metadata comprised the following areas:

1. Outcomes (DVs) investigated
2. Factors (IVs) investigated
3. Task utilized
4. Testbeds utilized

RESULTS

Outcome Variable Categories

There was a large number of DVs assessed in the 74 studies. These were grouped into 86 different DV and DV categories. Figure 1 depicts the five most–studied outcomes.

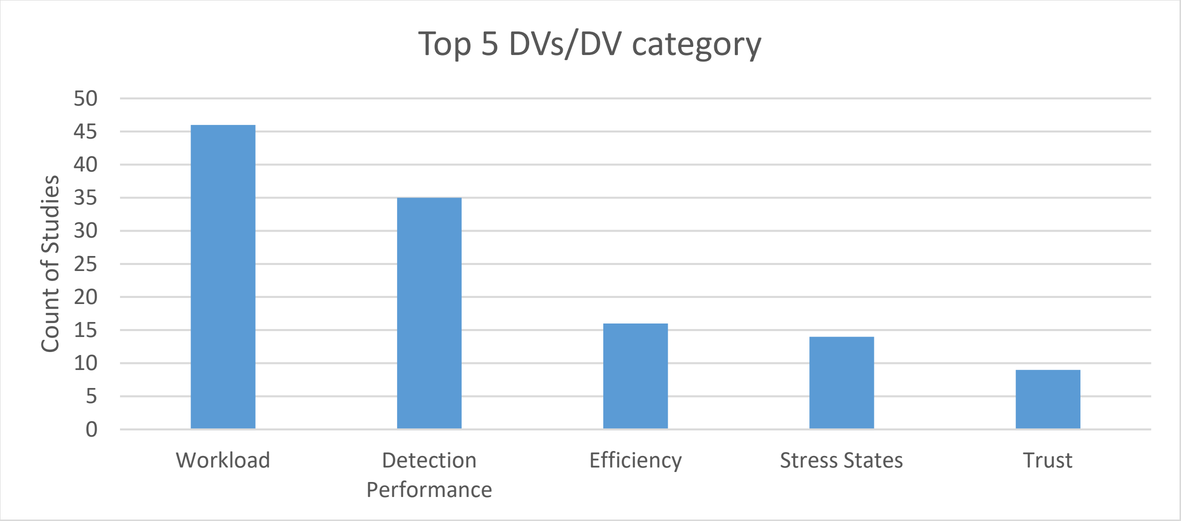

Figure 1: Top 5 DVs/DV category with the count of studies

Workload

Workload was the most studied DV, occurring in 46 of the 74 studies. The construct had 13 distinct operationalizations. Figure 2 shows the different ways in which *workload* was measured.

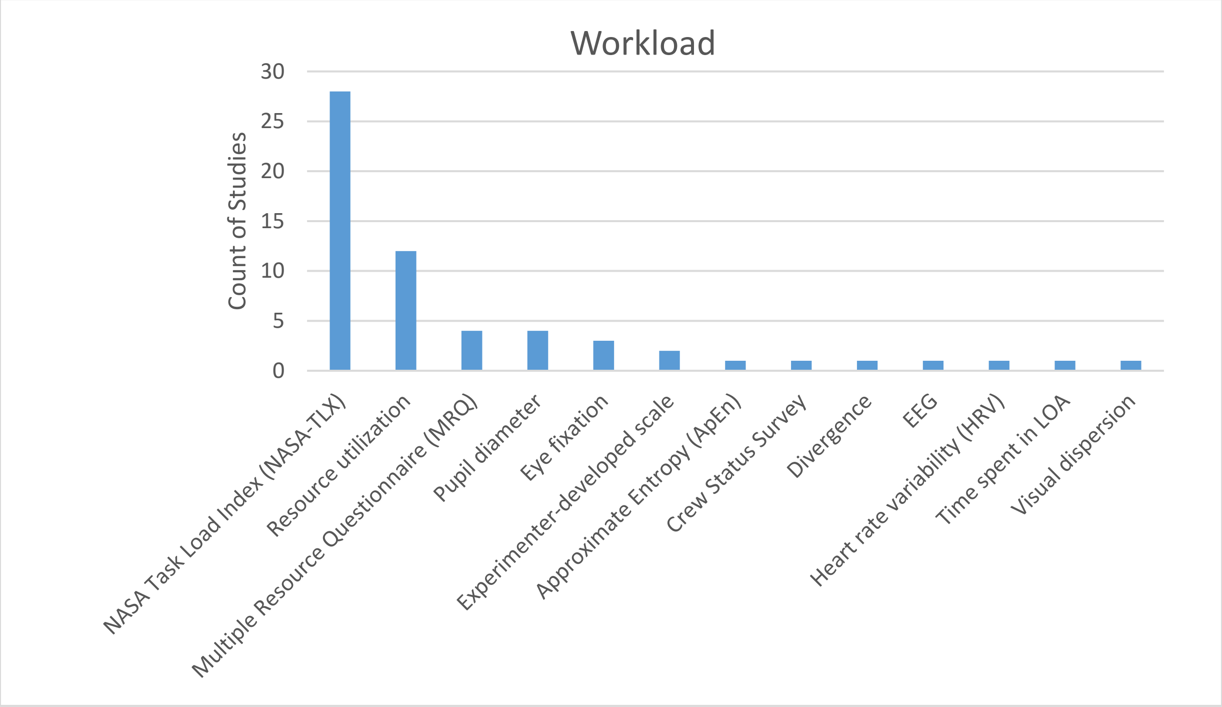

Figure 2: Different measures used for *workload* with a count of studies per measure

The NASA-TLX (Hart & Staveland, 1988) was by far the most utilized measure of workload. It was scored in three ways (i.e., overall score, unweighted score, and weighted score). Another popular subjective workload measure was the Multiple Resource Questionnaire (MRQ, (Boles & Adair, 2001). Others were physiological measures of workload, such as assessments of resource utilization with cerebral blood flow oxygenation (CBFV) and regional oxygenation saturation (rSO2), and measures of pupil diameter, eye fixation, EEG, and HRV. There were also other workload measures, but these were not common measures (e.g., Approximate Entropy (ApEn), Crew Status Survey (Ames & George, 1993), and visual dispersion.

Detection Performance

There were multiple measures of *detection performance*, which was the second most investigated study outcome. Specific metrics included proportion of correct detections/responses, correct rejections, false alarms, missed alarms, errors of omission and commission, and response time. *Detection performance* included detection of threat/hostile targets, detection of hostile emails or matching IP addresses (cyber threat detection), detection in collision avoidance tasks that required participants to detect when two RPAs/UVs/aircraft/tank are moving in an opposite direction or on a potential collision path, detection of critical phrases in communication-monitoring tasks, detection of orientation change in stimuli, as well detection of unspecified targets.

Efficiency

Efficiency is a time-based measure of performance (Steinfeld et al., 2006). There were multiple metrics of *efficiency*, corresponding to measures of performance on different tasks with respect to time (e.g., time to mission completion, average response time to correct detections).

Stress States

Stress states were mainly assessed with either the short or full version of the Dundee Stress State Questionnaire (DSSQ; Matthews et al., 2002), and was a DV of interest in 14 studies. The stress states assessed by the DSSQ were task engagement, distress, and worry. Task engagement refers to subjective state aspects of energy, motivation, and concentration, and has been linked to willingness to apply effort, where disengagement represents a withdrawal of effort (Matthews et al., 2002). Distress is associated with the affective aspects that relate to information processing overload, while Worry pertains to the state of evaluation of self or relevance of task to the self (Matthews et al., 2002). While most of the 14 studies assessed Stress States in terms of the three subscales, a few summarized Stress State using the overall DSSQ score.

Trust

Out of the 74 reviewed studies, *Trust* had been examined in nine studies, making it the fifth most assessed DV. There were multiple measures of *Trust*, and this was probably due to the fact that there are different definitions of the construct (Schaefer, 2013). However, the most utilized definition of trust was that proposed by Lee and See (2004), who posit that trust is “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” (p. 54). This construct had been assessed by the Function Specific Trust in Automated Systems Scale (Parasuraman, Sheridan, & Wickens, 2000) and its variants, the Trust in Automation Scale (Jian, Bisantz, & Drury, 1998), and the Human-Robot Trust Scale (Schaefer, 2013).

Independent Variable Categories

Across the 74 studies, there were many factors or IVs examined. However, the use of the IV names was very inconsistent, and required at some interpretation and categorization. The IVs were grouped into 89 different categories of IVs, with the top 5 as follows (see Fig. 3):

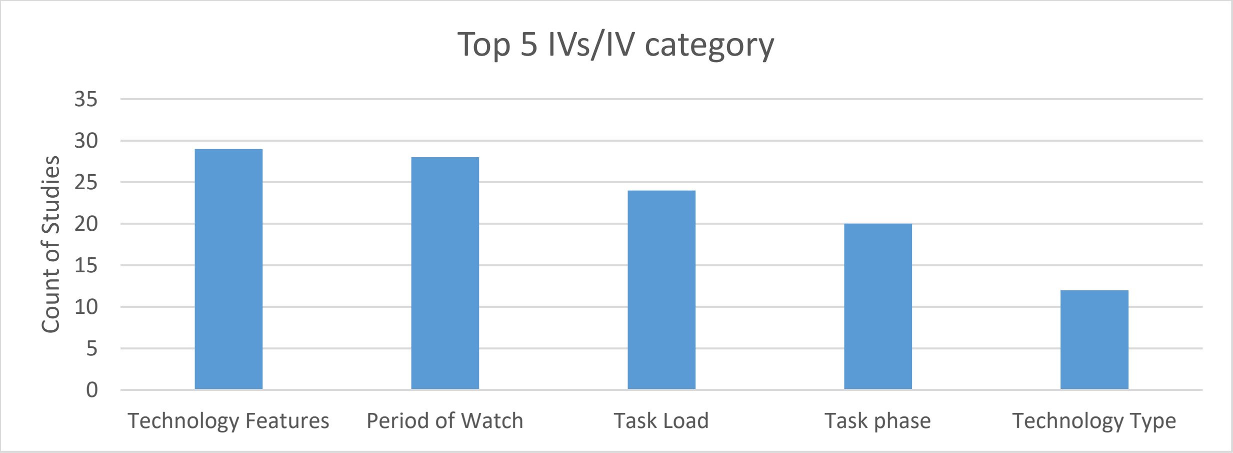

Figure 3: Top 5 IVs/IV category with the count of studies

Technology Features

This was a large, and heterogeneous category which comprised 12 very different IVs. The wide range of IVs in this IV category reflects the breadth of design opportunities for agents. "Technology Features" included IVs that pertained to system/agent interface details (e.g., *Presentation Mode*, *Background Color*), and agent capabilities (e.g., *Robot Autonomy, Reliability*) (see Fig. 4):

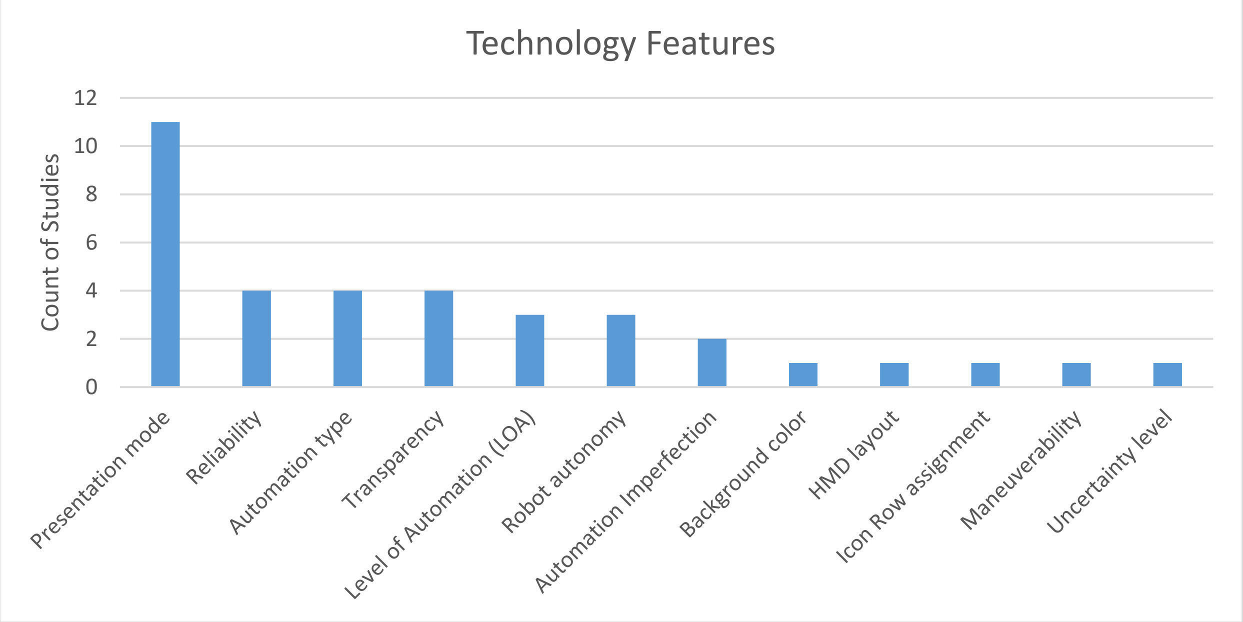

Figure 4: IVs in the IV category of "Technology Features" with the count of studies

The operationalization of the IV *Presentation Mode* included the use of spatial (3-D) audio, text, tactile, chat, combinations of these, within multimodal communication (MMC) (Donald, 2008). The construct of *Reliability* had been operationalized as the percentage of agent recommendations that were correct and was investigated in four studies. Researchers operationalized low reliability as 55% and 60%, whereas high reliability was operationalized as the agent being correct 86.7%, 90%, 93%, and 100% of the time. The construct *Transparency* pertained to the amount of information displayed from the robot to the human. It was an IV of interest in four studies and was operationalized in several ways including the agent possessing minimal, contextual, or constant information transparency levels. In the reviewed studies, *Automation type* was operationalized as static, adaptive, adaptable, no automation levels, while the construct of *Levels of Automation (LOA)* referred to the number of actions the agent can execute independently, and related to the extent to which the system or agent alleviated the human operator's information processing load (Parasuraman, Sheridan, et al., 2000). In our review, there were six different operationalizations of *LOA*. *Robot Autonomy* was distinct from constructs that related to automation since it had levels of auto, teleoperation, and monitor. The construct *Automation Imperfection* refers to the type of errors the automated system made, i.e., false alarm versus missed alarm. This has been manipulated in two studies. Other IVs relating to "Technology Features" included *Background Color*, *Icon Row Assignment*, *HMD Layout*, *Maneuverability*, and *Uncertainty Level*.

Period of Watch

Period of watch was a construct that related to the effects of time on a relatively uniform task without distinct task phases (e.g., as in a vigilance task). Levels of this construct included 2-minute, 3-minute, 5-minute, 10-minute, or 15-minute periods on the task.

Task Load

The IV category of "Task load" was among the most frequently examined IV, appearing in 24 studies, and examined by all ten authors. The IVs in this category included constructs such as *Event Rate*, *Sensory Modality* (auditory vs. sensory levels), *Source Complexity* (single vs. multitask levels; Parasuraman & Davies, 1977), *Number of Parameters* (e.g., number of robots to supervise, number of digits to hold in memory, etc.), *Number of Tasks* (single vs. dual tasks levels), *Signal Probability* (i.e., the likelihood of the critical event occurring), and *Visual Density*.

Task Phase

The IV category of “Task Phase” included IVs with levels that reflected different phases of the task, such as *Task-Related Change* (i.e., pre- and post-task phases), *Learning Phase* (i.e., first few trials vs. last few trials that corresponded to extent of learning), *Task Phase* (i.e., braking, replying, recovery phases in a driving task). Other IVs included *Temporal Manipulation, Event Schedule,* and *Time Segmentation*, among others.

Technology Type

Unlike the IV category “Technology Features” which encompasses details on task parameters, the IV category “Technology Type” comprised IVs with levels that enabled comparisons of different types of agent capabilities. Some of the types entailed comparing 2-D vs. 3-D displays, coordinated versus uncoordinated displays, trackpoint versus touchpad, etc. In addition, there was also the IV of *Agent presence,* which contrasted the outcomes when an agent was present to that when there was no agent.

Tasks Utilized

Examination of the tasks utilized in HAT research shows the scope of application of HAT. The top ten most utilized tasks between authors and across studies are presented in Figure 5. These tasks corresponded to some extent to the subcategories of *Detection performance*.

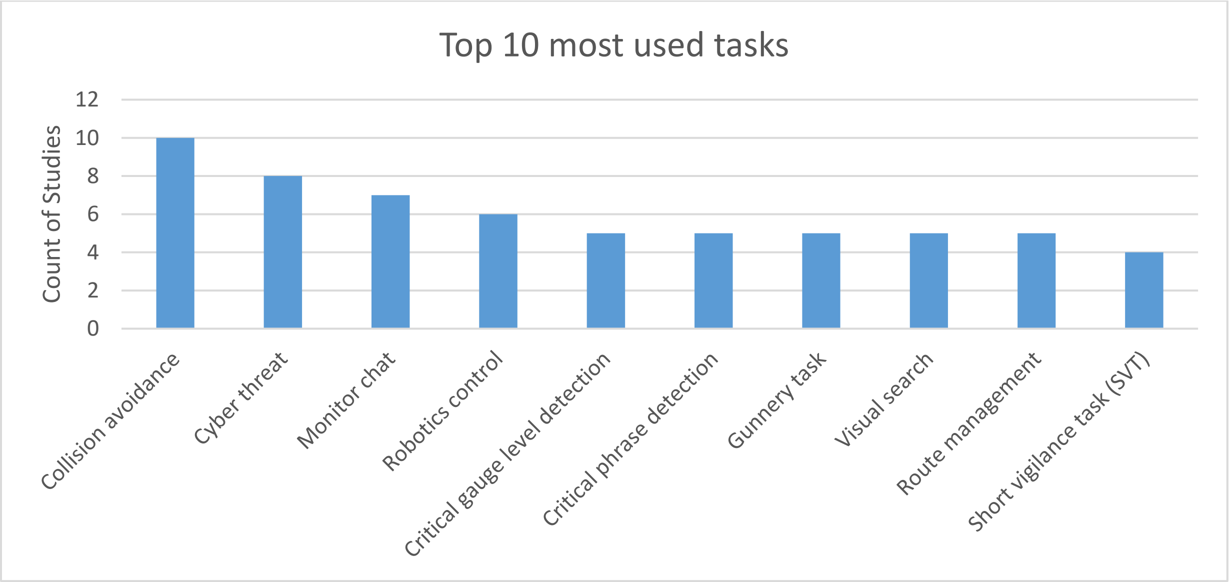

Figure 5: Top 10 most used tasks

“Collision Avoidance” was a task often performed by pilots, drivers, and UAV/UGV controllers. “Cyber Threat Detection” was a task involving detection of hostile emails or IP-addresses. “Chat Monitoring” was often employed as a concurrent task. The “Robotics Control” task required teleoperating a robot, while the “Critical Signal Detection” task entails the detection of an important signal, such as a barrel being longer, a line being longer, and auditory signal, etc. The “Route Management” task involved rerouting agents, whereas the goal in a “Visual Search” task was to search an area visually for specified targets. The “Gunnery Task” involved searching for hostile targets and reporting their location. During the “Critical Phrase Detection” task, operators were to detect a critical phrase from a stream of auditory information. For the “Critical Gauge Level Detection” task, operators were to detect when a gauge, often symbolizing battery life or fuel level, reached a critical level.

Testbeds Utilized

The testbed is an important part of the context within which assessments are conducted. Oftentimes, the capabilities of the testbed influence the type of tasks that can be administered. Table 2 shows an overview of the most utilized testbeds in the HAT studies reviewed, and their associated task capabilities.

Table 2: Excerpt of testbeds used in HAT studies and their capabilities

Testbed Task Capabilities	ALOA*	MIX†	MMC‡	SCOUT§	TCU‖	VBS2¶
SUPPORTING COMMUNICATIONS AMONG TEAM MEMBERS (e.g., supports “Chat Monitor”, “Coordinate Response Measure (CRM)”, “Critical Phrase”, “Simple Communication”)	✓	✓	✓		✓	
ADMINISTERING DETECTION TASKS (e.g., “Gunnery Task”, “Threat Search”, “Visual Search”, “Change Detection Task”, “Insurgent Search”, “Critical Gauge”, “Critical Phrase”)	✓	✓	✓	✓	✓	✓
DISPLAYING MAP LOCATIONS		✓				
SUPPORTING TELEOPERATIONS (e.g., “Robotics Control”)		✓			✓	✓
SUPPORTING ROUTE MANAGEMENT		✓		✓		

*ALOA: Adaptive Levels of Autonomy, v.3 research testbed developed by OR Concepts Applied (ORCA) (Johnson, Leen, Goldberg, & Chiu, 2005)
†MIX: Mixed Initiative Experimental testbed (Barber, Davis, Nicholson, Finkelstein, & Chen, 2008)
‡MMC: Multi-Modal Communication, a network-centric communication management suite (Finomore et al., 2011)
§SCOUT: Supervisory Control Operations User Testbed (Sibley, Coyne, & Thomas, 2016)
‖TCU: Tactile Control Unit, developed by the U.S. Army Research Laboratory’s Robotic Collaborative Technology Alliance (RCTA) (Chen & Terrence, 2009)
¶VBS2: Virtual Battlespace 2 (Morrison, 2012)

DISCUSSION

This metadata study was conducted in consideration of the current and foreseeable assessment challenges in HAT research. Although by no means exhaustive or comprehensive, we hoped to describe the state of assessments in the HAT domain by understanding how HAT research has been conducted, and the contexts within which the research has been conducted. This information can help the research community understand how best to build upon the research that has been done, as well as identify research gaps.

In the studies examined, the outcomes (DVs) that were assessed the most were *workload*, *detection performance*, *efficiency*, *stress states*, and *trust*. Across the HAT studies, the operationalization and measurement of workload involved a wide range of measures such as multiple subjective, performance, and physiological measures. In the 74 studies, there were 13 different measures of *workload*. While the range of measures for workload reflects the multi-dimensional nature of the construct, it can also signal the problem with articulating what some of these dimensions may be, or they may require specification of other factors, in order to define and measure the construct more precisely. HAT assessments may benefit from guidelines to help determine which workload measure is most suited for what study. While the effects of technological advances on workload and stress states have been a common concern, the impact of technology on trust is relatively recent and shows the transition of technology as tools to partners or teammates. The inclusion of *detection performance* as the second-most commonly studied outcome in the HAT studies reviewed suggests that many applications of HAT teaming involve agents assisting with target/signal detection of some kind across a variety of tasks (e.g., detection of threat characters, cyber threats, general targets which comprise different stimuli). With advances in artificial intelligence (AI) for face and image recognition, there is a real possibility of developing agents with this capability. Developmental efforts for agents in this area should focus on AI which aid in search and detection tasks, and machine learning algorithms that improve signal-to-noise ratios to boost sensitivity, and assist the human operator with appropriate criterion shifts in signal detection. The other performance outcome most assessed was *efficiency*, which is a measure of performance with respect to time. While timely performance is generally desirable, there has not been much work done to understand the processes or strategies involved in untimely performance. For instance, for any given level of performance, the operators who took less time may engage in different cognitive processes for performance compared to operators who took more time. In developing agents that support tasks, it may be necessary to understand the different methods and strategies human operators use when they execute tasks.

The factors (IVs) that were most widely examined in HAT studies include *technology features, period of watch, task load, task phase,* and *technology type*. Unsurprisingly, HAT studies have focused on the effects of the technology features since such studies influence the development and design of agents for various tasks. *Technology features* range from interface details and features (i.e., information presentation mode, HMD layout, color, position of icons), to specific agent capabilities (i.e., reliability, automation, maneuverability), including its ability to be independent from the human (i.e., robot autonomy, uncertainty level, transparency). In contrast, studies on *technology type* typically compared outcomes from types of agent capabilities (e.g., systems with coordinated vs. uncoordinated displays). The inclusion of *task load* and *task phase* underscores the importance of understanding the task to be performed by the HAT. As expected, there are several different operationalizations of *task load*, reflecting the range of task parameters across HAT studies to some degree (e.g., modality, event rate, visual density). The factor *period of watch* relates the effects of time on HAT outcomes and acknowledges that there can be different short- and long-term effects of factors as well.

For a number of studies, HAT has been applied to *collision avoidance tasks* (e.g., Automatic Ground Collision Avoidance Technology (AGCAS)), *cyber threat* and *other detection tasks*, suggesting that these are the agent capabilities that are prioritized at this juncture. While there is definitely a need for agents' assistance with such tasks (e.g., need to process large volumes of sensor data that humans are unable to process), HAT research can be extended to incorporate other tasks that exploit other agent capabilities (e.g., amassing data from multiple sources, including large volumes of historical data, and extracting patterns and relationships with various data mining methods for the purpose of prediction and adaptation). However, all these are contingent on the capabilities and features of the testbeds at hand. Our analysis of testbeds indicates that most of the testbeds support some form of communications and different detection tasks, but it remains to be seen what other tasks can be supported by these testbeds.

LIMITATIONS & FUTURE DIRECTION

Due to project constraints, the present analysis of metadata was limited to Department of Defense (DoD) research labs. While this study provides an initial overview of the assessments in HAT research, extension of these analyses to other non-DoD HAT studies is necessary to obtain a more comprehensive understanding of HAT assessments and applications. Such understanding of DoD and non-DoD HAT studies (i.e., the conceptual similarity of assessments and construct operationalizations) can then inform the strategy for future meta-analysis that will shed more light on the sizes of the effects and construct relationships that are of most importance in HAT. All these will contribute to our knowledge of how best to team humans and agents for various applications.

CONCLUSION

The present study established a baseline of what common constructs in HAT are and how these have been operationalized. We found that in some studies, multiple measures have been used for the same construct without sufficient specification of what the scores on those measures mean within the context of use. While we recognize that the use of multiple measures for a construct is necessary as different measures are sensitive to the constructs they measure in different ways, more work needs to be done to better define and understand how multi-dimensional constructs are manifesting in the specific context (e.g., workload). Psychometric and factor analytic methods will be useful for this purpose. Ongoing construct validation studies are also needed to keep track of how construct meanings change with different contexts. Without such a foundation, the impact of future work may be limited.

ACKNOWLEDGEMENTS

This research was sponsored by the U.S. Army Research Laboratory (ARL) and was accomplished under Cooperative Agreement Number W911NF-15-2-0100. The views and conclusions contained in this document are those of the author's and should not be interpreted as representing the official policies, either expressed or implied, of ARL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation herein.

REFERENCES

Ames, L. L., & George, E. J. (1993). *Revision and verification of a seven-point workload estimate scale*. Air Force Test Center Edwards AFB CA.

Barber, D., Davis, L., Nicholson, D., Finkelstein, N., & Chen, J. Y. C. (n.d.). The Mixed Initiative Experimental (MIX) Testbed for Human Robot Interactions With Varied Levels of Automation, 7.

Boles, D. B., & Adair, L. P. (2001). The Multiple Resources Questionnaire (MRQ). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *45*(25), 1790–1794. https://doi.org/10.1177/154193120104502507

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Calhoun, G. L., Ruff, H. A., Spriggs, S., & Murray, C. (2012). Tailored performance-based adaptive levels of automation (Vol. 56, pp. 413–417). Presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications Sage CA: Los Angeles, CA.

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.

Dillard, M. B., Warm, J. S., Funke, G. J., Funke, M. E., Finomore Jr, V. S., Matthews, G., … Parasuraman, R. (2014). The sustained attention to response task (SART) does not promote mindlessness during vigilance performance. *Human Factors*, *56*(8), 1364–1379.

Donald, F. M. (2008). The classification of vigilance tasks in the real world. *Ergonomics*, *51*(11), 1643–1655.

Finomore Jr, V., Popik, D., Dallman, R., Stewart, J., Satterfield, K., & Castle, C. (2011). Demonstration of a network-centric communication management suite: multi-modal communication. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 55, pp. 1832–1835). SAGE Publications Sage CA: Los Angeles, CA.

Finomore Jr, V. S., Shaw, T. H., Warm, J. S., Matthews, G., & Boles, D. B. (2013). Viewing the workload of vigilance through the lenses of the NASA-TLX and the MRQ. *Human Factors*, *55*(6), 1044–1063.

Franklin, S., & Graesser, A. (1996). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents (pp. 21–35). Presented at the International Workshop on Agent Theories, Architectures, and Languages, Springer.

Gal, Y., Grosz, B., Pfeffer, A., Shieber, S., & Allain, A. (2007). The Influence of Task Contexts on the Decision-Making of Humans and Computers. In B. Kokinov, D. C. Richardson, T. R. Roth-Berghofer, & L. Vieu (Eds.), *Modeling and Using Context* (Vol. 4635, pp. 206–219). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74255-5_16

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

Hooey, B. L., Kaber, D. B., Adams, J. A., Fong, T. W., & Gore, B. F. (2017). The underpinnings of workload in unmanned vehicle systems. *IEEE Transactions on Human-Machine Systems*.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (1998). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Johnson, R., Leen, M., Goldberg, D., & Chiu, M. (2005). *Adaptive levels of autonomy (ALOA) for UAV supervisory control*. OR CONCEPTS APPLIED WHITTIER CA.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

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 (Vol. 59, pp. 746–750). Presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications Sage CA: Los Angeles, CA.

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. A., Huggins, J., Gilliland, K., … Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion*, *2*(4), 315.

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.

Maulsby, D., Greenberg, S., & Mander, R. (1993). Prototyping an intelligent agent through Wizard of Oz. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 277–284). ACM.

Morrison, P. (2012). White Paper: VBS2 Release Version 2.0. *Nelson Bay, Australia*.

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.

Schaefer, K. (2013). The perception and measurement of human-robot trust.

Sibley, C., Coyne, J., & Thomas, J. (2016). Demonstrating the Supervisory Control Operations User Testbed (SCOUT). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 1324–1328. https://doi.org/10.1177/1541931213601306

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., & Goodrich, M. (2006). Common metrics for human-robot interaction (pp. 33–40). Presented at the Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction, ACM.