5-12-2023

A generalizable method and case application for development and use of the Aviation Systems – Trust Survey (AS-TS).

Jamison Hicks
Mississippi State University, jamison.hicks@gmail.com

A generalizable method and case application for development and use of the Aviation Systems – Trust Survey (AS-TS).

Jamison Hicks

Approved by:

Reuben F. Burch V (Major Professor)
Lesley Strawderman
Kari Babski-Reeves
Daniel W. Carruth
Mohammad Marufuzzaman (Graduate Coordinator)
Jason M. Keith (Dean, Bagley College of Engineering)

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Industrial and Systems Engineering
in the Department of Industrial and Systems Engineering

Mississippi State, Mississippi

May 2023

Name: Jamison Hicks

Date of Degree: May 12, 2023

Institution: Mississippi State University

Major Field: Industrial and Systems Engineering

Committee Chair: Reuben F. Burch V

Title of Study: A generalizable method and case application for development and use of the Aviation Systems – Trust Survey (AS-TS).

Pages in Study 186

Candidate for Degree of Doctor of Philosophy

Automated systems are integral in the development of modern aircraft, especially for complex military aircraft. Pilot Trust in Automation (TIA) in these systems is vital for optimizing the pilot-vehicle interface and ensuring pilots use the systems appropriately to complete required tasks.

The objective of this research was to develop and validate a TIA scale and survey methodology to identify and mitigate trust deficiencies with automated systems for use in Army Aviation testing. There is currently no standard TIA assessment methodology for U.S. Army aviation pilots that identifies trust deficiencies and potential mitigations.

A comprehensive literature review was conducted to identify prominent TIA factors present in similar studies. The compiled list of factors and associated definitions were used in a validation study that utilized the Analytic Hierarchy Process (AHP) as a pair-wise comparison tool to identify TIA factors most relevant to Army pilots.

A notional survey, the Aviation Systems – Trust Survey (AS-TS), was developed from the identified factors and pilots were used as subjects in scenario-based testing to establish

construct validity for the survey. Exploratory factor analysis was conducted after data collection and a validated survey was produced.

A follow-on study interviewed Army test and evaluation experts to refine the survey methodology and ensure appropriate context for the recommended mitigations. A final packet was developed that included instructions for the rating scale, associated item definitions, and recommended mitigations for trust deficiencies. Future research will focus on other Army demographics to determine the generalizability of the AS-TS.

Keywords: Trust factors, Automation, Aviation, Trust survey, Analytic Hierarchy Process

ACKNOWLEDGEMENTS

I would like to thank the Mississippi State University advisory committee members for their time and guidance during the dissertation process. Specifically, Dr. Reuben Burch who was instrumental in preparing me for the PhD candidacy process.

I am grateful to the personnel and supervisors at the DEVCOM Analysis Center – Human Systems Integration Division for their support during both the research and PhD requirements process. Special thanks to Derek Millard for assistance in developing the Unreal software virtual scenarios used in this research.

Finally, I would like to thank all the participants that took the time to provide their expertise and input, making this research a success.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

RESEARCH INTRODUCTION

Automation can be defined as a technology that "actively selects data, transforms information, makes decisions, or controls processes" (Lee & See, 2004). The concept of an automated system traces definitively back to the first century AD, with the invention of an automatic vending machine developed by Heron of Alexandria (Danner, 2019). Around 2,000 years have passed since the invention of the first vending machine, since then, the world has seen the development of automated systems that have changed the course of history. Jumping ahead to the 1700's societies began to experience a series of Industrial Revolutions that capitalized on new energy technologies (e.g., steam, electricity) that transformed automated processes by improving efficiency for industrial production (Britannica, 2021). In more recent history, complex automation is now included in many consumer products from sensor-based wrist watches to self-driving cars. Since 2013, industry has used the buzzword "Automation Revolution" to describe the near future of work (Scholl & Hanson, 2020). These advanced automated systems and trends will continue to proliferate through our daily lives in both work and leisure.

As automation capabilities and use cases are being rapidly expanded, several ideas related to human-autonomy teaming interactions (e.g., trust, satisfaction, frustration, etc.) become relevant to both designers and users of automated systems (Shahrdar, Menezes, & Nojoumian, 2018). Trust in Automation (TIA) is considered one of the primary challenges for

successful integration of automation, Artificial Intelligence (AI), and humans (Beer, Fisk, & Rogers, 2014). Siau and Wang (2018) define relational trust as containing the following elements: (1) trusting beliefs (i.e., benevolence, competence, integrity, and predictability); (2) trusting intention (i.e., willingness to depend on another party during risky situations); (3) some combination of trusting beliefs and trusting intention. Trust can be contextually defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004). In the case of TIA, the agent can be considered an automated system. Automation system complexity often demands user trust for cooperative use. TIA is a useful construct when attempting to optimize and assess these human-automation interactions. TIA has been correlated to how often humans rely on and accept new technologies, key indicators of effective human-automation design (Lyons et al., 2016).

Trust Measurement

Measuring TIA is very difficult due to its multi-faceted nature (Brzowski & Nathan-Roberts, 2019). In general, the measurement of TIA is often associated with the subjective user perception of trust between the user and the automated system (Parasuraman & Riley, 1997). There are numerous factors that can influence the TIA relationship. Factors ranging from human perception of automation characteristics, system performance, mental model expectations, personality, cultural influences, and mental workload have all been found to play a potential role in TIA (Brzowski & Nathan-Roberts, 2019; Hoff & Bashir, 2015).

Establishing a level of TIA in a system can be accomplished with several different methods dependent on the context of the automation and researcher intent. Brzowski & Nathan-Roberts (2019) conducted a systematic literature review to show the frequency of types of trust measurements (e.g., subjective, objective) related to automated systems. The primary conclusion

of the Brzowski & Nathan-Roberts (2019) research was that a majority (75%) of the relevant articles utilized subjective measurement to collect trust data. The primary measurement tools were subjective rating scales that examined relevant trust factors (e.g., reliability, competency, understanding) to establish user perception of the system. An initial measurement of TIA can be accomplished through the methodology of utilizing a valid survey instrument in context with the appropriate system (Brzowski & Nathan-Roberts, 2019). By using a valid survey instrument, researchers can identify a level of perceived trust between the user and the automation.

In addition to identifying a user trust level, it is often helpful to ensure that the user trust level is appropriately calibrated for the application. Inappropriate levels of trust are often defined as instances of improper trust calibration through over or under reliance on the automation (Duez, Zuliani, & Jamieson, 2006). In the case of inappropriate trust, objective measurements are typically used to evaluate the reliance relationship between the user and the automated system (Wang, Jamieson, & Hollands, 2008). The two concepts of trust and reliance are interrelated. Additionally, research has found that user reliance on automated systems can be affected by affective influences (i.e., emotional state), as well as trust (Merritt, 2011).

Trust in Aviation Systems

While there are many examples of automated systems requiring human trust, aviation related human-automation applications have been at the foremost position of the research efforts. Technologies such as autopilot, fly-by-wire, automated route planners, global positioning systems (GPS), and advanced inertial measurement units (IMU), are all part of the growing multitude of automated systems found in modern aircraft. Military aviation applications often utilize even more complex systems than general or commercial aviation, due to special equipment and stressful mission requirements (Pamplona & Alves, 2020). U.S. Army rotorcraft

(helicopter) operations are particularly complex when considering the mission demands and environmental conditions where these operations occur. Helicopter pilots often operate in lower altitudes, complex terrain, hostile environments, and within Degraded Visual Environments (DVE) (Helfrich, 2020). These unique circumstances drive the need for robust human-automation designs that enable the pilots to effectively utilize the aircraft systems for mission accomplishment.

Future Vertical Lift

The U.S. Army is at the forefront of technology development for advanced autonomy and human-interface design. The Army Future Vertical Lift (FVL) program is one of the Army's three major modernization priorities and includes initiatives for both a Future Attack Reconnaissance Aircraft (FARA) and a Future Long-Range Assault Aircraft (FLRAA) (Mayfield, 2021). A significant design consideration for FVL is optimizing the human-automation interactions that will occur during system use. FVL is required to provide multiple levels of supervised autonomy within the aircraft (e.g., autonomous takeoffs/landings, cueing, and adaptive interventions). Korber, Baseler, & Bengler (2018) states that TIA is a key determinant for the adoption of automated systems and their appropriate use. Appropriate levels of TIA for the pilots will be extremely important for FVL platforms due to a significant focus on automated processes and automated assistance in high-speed and DVE rotorcraft operations (Freedberg, 2020).

Future Vertical Lift Roles

Within the FVL aircraft concept, two roles stand out as having direct interaction with the FVL systems. The traditional pilot role of aircraft state monitoring and *hands-on* flight control,

and the Air Mission Commander (AMC). The AMC is a rated crewmember designated as the primary tactical and operational decision maker to accomplish the mission intent (Antonides, 2014). In the case of FVL, the AMC role will likely manage much of the battlefield operations through Manned-Unmanned Teaming (MUM-T) operations, requiring advanced automation for adaptive decision making, workload reduction, sensor management, and attention allocation (Taylor & Turpin, 2015). The Synergistic Unmanned-Manned Intelligent Teaming (SUMIT) program examined the role of the AMC in FVL and their interactions with multiple user interfaces to control the battlefield, including numerous unmanned systems (Alicia, Hall, & Terman, 2020). While some of these actions could take place remotely, initial testing indicates that the AMC will be a crewmember in a FVL aircraft (Alicia, Hall, & Terman, 2020). Identifying the appropriate levels of TIA for the AMC is very important to ensure that the AMC's intent is carried out in a reliable manner and that the AMC maintains appropriate battlefield awareness by effectively utilizing the FVL systems.

The AMC mission focus is primarily on battlefield management. Automation that assists in sensor management (e.g., Unmanned Aircraft System (UAS) cameras), asset control (e.g., Air Launched Effects (ALE)), and battlefield element (e.g., enemy and friendly forces) interactions is especially valuable in this role. AMC's need to maintain high levels of TIA in reliable systems to ensure accurate execution of tasks for battlespace management and specific asset control (e.g., ALE swarms). High levels of TIA are required to ensure AMC's trust that assets can be successfully employed to accomplish critical portions of the intended mission.

Problem Statement

As autonomy continues to play a major role in aircraft system use and development, users must trust that the automation is performing to standard. There is currently no standard

methodology that is in use for the Army to assess TIA for pilots as a holistic measurement that identifies trust deficiencies and the relationship of trust to user reliance with follow-up actions. Several surveys are currently in use, but only used as a general indicator of system trust with no recommendations to improve the user-automation trust relationship. Additionally, many of the TIA scales only consider a limited number of potential TIA factors that may have an impact on systems outside of the purpose of the initially targeted research. Similar issues exist across other research areas. In McKnight & Chervany (1996), the authors express concern over comparing different types of trust constructs (e.g., personality-based vs. structural) across studies. Much of the TIA research is focused on a specific area of interest, where the research may not be generalizable to other domains (Pak et al., 2016). Not having a standard theoretical and empirical trust construct for system measurement, results in incompatible comparisons across the research applications. Standardization and a comprehensive analysis of TIA factors that are directly relevant for Army Aviation can help to provide focused assessments of TIA across the lifecycle of a system.

The term *trust deficiency* is not readily found in TIA research. The use of the term *trust deficiency* in this research is intended to describe instances of the user-automation relationship that indicate a subjective lack of trust in the system and/or an inappropriate reliance on the system. In this context, a trust deficiency can be defined as an inadequate trust relationship between the user and the automated system, where the user perceives the system as untrustworthy and/or the user is unable to reasonably calibrate their trust in the system. The definition is derived from references related to the factors that influence perceived trust and associated trust levels as they relate to reliance on the automation (Brzowski & Nathan-Roberts, 2019; Hoff & Bashir, 2015).

Purpose

The purpose of this research is to produce a measurement tool and assessment methodology for TIA assessment using an Army Aviation helicopter system use case that can help to identify *trust deficiencies*. A comprehensive literature review was conducted based on key words that address TIA and TIA factor characteristics across a variety of industries (e.g., aviation, driving, and maritime). The research informs readers on current TIA concepts and various measurement techniques, as well as providing an overview of the relationship between user trust and automated systems. A concentrated analysis of which factors are relevant to Army Aviation pilots was conducted to ensure contextual application of the proposed TIA factors. A measurement tool (i.e., survey) and methodology were developed based on effective research implementations identified during the systematic literature review. Development and utilization of a survey tool, based on the identified TIA factors, can provide the relevant information required for analysts to evaluate perceived user trust of the system.

Survey tools often rely on psychometric measurement techniques (e.g., Likert scale) where they are commonly used to quantify traits like ability, perceptions, qualities, and outlooks in research (Joshi et al., 2015). The use of a Likert scale can be used to measure a latent variable through questionnaire items to address specific dimensions and composite measurement of the variable under investigation (Joshi et al., 2015). The survey tool will be validated and assessed for reliability based on subject-matter expert and pilot feedback.

The primary purpose of the methodology is to identify the dimensional and overall level of trust and trust-related deficiencies within the system (e.g., reliability, transparency, training) that affect user trust perception along with recommendations on steps to correct or mitigate the deficient area.

Finally, the measurement tool was evaluated by Army Aviation testing (e.g., simulations, developmental, and operational tests) experts to assess the efficacy of the tool and provide feedback related to the methodology. Data were used to refine and validate the methodology prior to full implementation as an accepted tool for TIA data collection and analysis.

Aviation Systems-Trust Survey (AS-TS) is the proposed title of the developed survey. This title helps to scope the use case for the survey. The developmental methodology in this research is generalizable for other TIA survey developments and use cases. It is highly likely that many of the survey items identified in this study will overlap with other automated systems (e.g., ground-based vehicles, workstations). As further research is conducted, a more generalized survey naming convention could be applied to reflect the survey capabilities of TIA measurement.

Additionally, factors and items included in the survey tool and collected ratings should enhance the fidelity of the information provided to decision-makers. The AS-TS provides several TIA items included with established definitions for independence from similar factors. The AS-TS can provide increased fidelity to human factors engineers and product managers regarding decisions on required improvements or mitigations for low scoring trust items.

For all research activities henceforth, participant recruitment and follow-on interactions were conducted in accordance with guidelines required by the Mississippi State University and U.S. Army DEVCOM Analysis Center – Human Research Protection Protocol's defined by each institutions Internal Review Board. Participants consented to data collection and were informed that they could discontinue participation at any time with no penalty.

Research Questions

Army Aviation is the subject of the case-based approach for the specific survey development. The development of a measurement tool to evaluate U.S. Army pilot TIA requires satisfactory answers to the following three research questions.

1. What factors influence TIA for Army pilots using Army Aviation systems?

    a. Hypothesis 1: The Analytic Hierarchy Process (AHP) decision-making tool will be effective in identifying critical TIA factors and establishing face and content validity of the proposed TIA factors for inclusion on an initial aviation systems trust survey.

    b. Hypothesis 2: Many of the key factors identified within the initial literature review will likely be considered relevant factors for Army Aviation pilots, with the exception of personal attachment and demographics.

2. Can a survey instrument developed from identified factors reliably measure pilot TIA perception of Army Aviation systems?

    a. Hypothesis 1: A survey pre-test, subject to factor analysis and reliability testing, will successfully determine construct validity and reliability for the proposed TIA survey.

    b. Hypothesis 2: The proposed survey will contain two overarching factors (human and automation) verified through exploratory factor analysis.

    c. Hypothesis 3: Scenario-based validity testing will result in a single factor trust construct when analyzing anchor scenarios of positive and negative experience automation.

    d. Hypothesis 4: Scenario-based validity testing will identify two overarching factors (human and automation) for alternating imperfect automation scenarios.

3. Can the survey instrument be used effectively in formal design testing to provide actionable information to data analysts (e.g., Human Factors Engineers) and product managers?

    a. Hypothesis 1: The developed survey will be a useful tool for analysts and program managers to identify TIA deficiencies, based on decision-maker and analyst ratings of effectiveness.

b. Hypothesis 2: The recommended actions list will provide appropriate courses of action to correct the deficiencies, based on decision-maker and analyst ratings of effectiveness.

Figure 1 provides an overview of the study order used to address the research questions. Study 1 includes a literature review, evaluation of factors through the AHP and development of a notional survey. Study 2 starts with a pre-test of the notional survey followed by factor analysis and reliability testing in support of a final validated survey. Study 3 interviews Subject Matter Experts (SMEs) for feedback related to the survey tool and mitigation suggestions resulting in a final survey packet for use during Army testing.

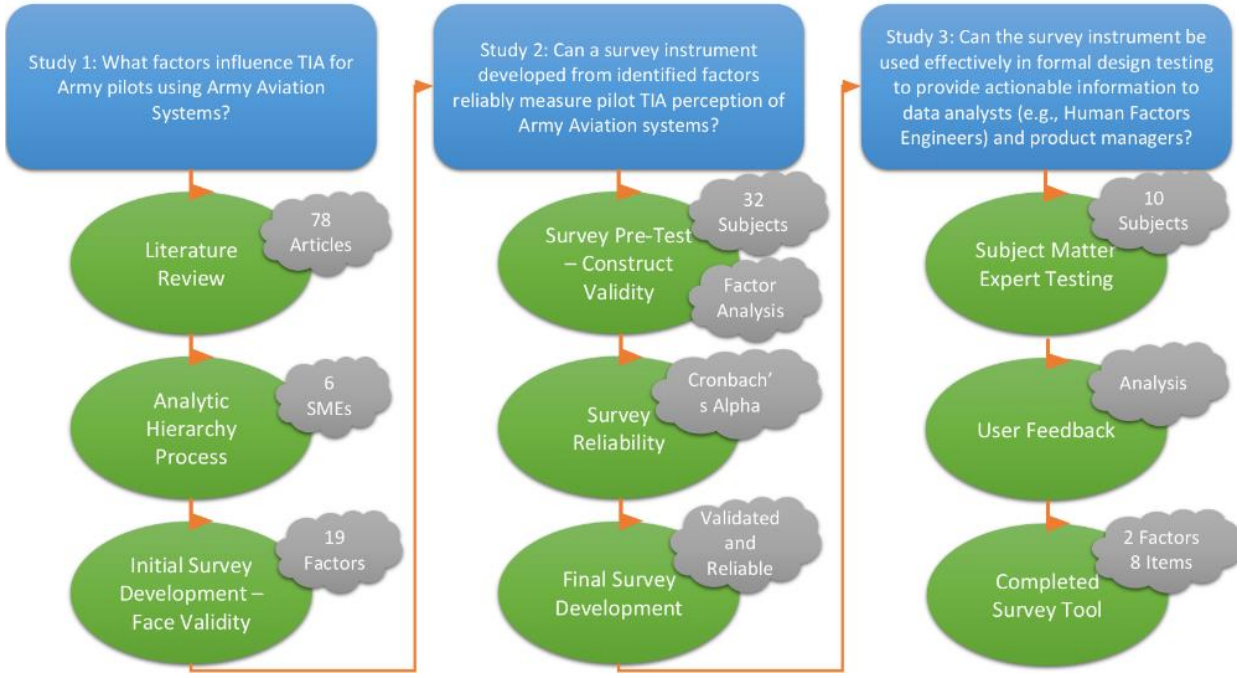

Figure 1 Study order

The results of this research provided a validated survey instrument and associated information on multi-dimensional analysis for identifying potential trust deficiencies related to

the human-automation interactions among aircraft systems and recommended potential mitigations for identified deficiencies.

Assumptions and Limitations

The development of this methodology is initially limited in scope to the identified case population of U.S. Army pilots in Army Aviation. The generalizability of the developed method will likely transfer to other military systems or aviation products but will need further validation to ensure appropriate application.

A primary limitation of the proposed method is the lack of immediate and complete access to line pilots for survey validation. However, Winter, Dodou, & Wieringa (2009) and Peters (2015) make statistical cases for using 30 – 50 participants for data collection inferences based on evaluation of Cronbach's alpha for scale reliability and Exploratory Factor Analysis (EFA) dataset characteristics. A sampling of 30-50 participants was an attainable goal for pilot participation.

CHAPTER II

LITERATURE REVIEW

Introduction

Autonomous systems play a major role in modern aircraft use and development, pilots must trust that the automation is performing to standard to ensure appropriate performance of the aircraft and satisfactory completion of pilot tasks. While there are many examples of automated systems requiring human trust, aviation related human-automation applications have been at the foremost position of the research efforts. Technologies such as autopilot, fly-by-wire, automated route planners, GPS, and advanced IMU's, are all part of the growing multitude of automated systems found in modern aircraft. Military aviation applications often utilize even more complex systems than general or commercial aviation, due to special equipment and stressful mission requirements (Pamplona & Alves, 2020).

The U.S. Army is especially concerned with pilot TIA, as new technologies such as the Army FVL program are rapidly approaching. The FVL program is one of the Army's three major modernization priorities and includes initiatives for both a FARA and a FLRAA (Mayfield, 2021). A significant design consideration for FVL is optimizing the human-automation interactions that will occur during system use. FVL is required to provide multiple levels of supervised autonomy within the aircraft (e.g., autonomous takeoffs/landings, cueing, and adaptive interventions). Korber, Baseler, & Bengler (2018) states that TIA is a key determinant for the adoption of automated systems and their appropriate use. Appropriate levels

of TIA for the pilots will be extremely important for FVL platforms due to a significant focus on automated processes and automated assistance in high-speed and DVE rotorcraft operations (Freedberg, 2020).

Problem Statement

There is currently no standard methodology that is in use for the Army to assess TIA for pilots as a holistic measurement that identifies trust deficiencies and the relationship of trust to user reliance with follow-up actions. Measuring TIA is difficult due to its multi-faceted nature (Brzowski & Nathan-Roberts, 2019). In general, the measurement of TIA is often associated with the subjective user perception of trust between the user and the automated system (Parasuraman & Riley, 1997). Several surveys are currently in use, but only used as a general indicator of system trust with no recommendations regarding follow-up actions to increase user trust or suggestion of mitigations. Similar issues exist across other research areas. McKnight & Chervany (1996) express concern over comparing different types of trust constructs (e.g., personality-based vs. structural) across studies. Much of the TIA research is focused on a specific area of interest, where the research may not be generalizable to other domains (Pak et al., 2016). Not having a standard theoretical and empirical trust construct for system measurement, results in incompatible comparisons across the research applications.

To address the lack of a standardized survey tool for TIA measurement in Army Aviation, a comprehensive literature review was conducted to identify key factors that may influence TIA in aviation systems and answer the research question “What factors influence TIA for Army pilots using Army Aviation systems?” These identified factors were used as a foundation to develop a validated survey tool specialized for use in Army Aviation system assessments.

Human vs. Automation Trust

There are several concepts of *trust* that are used to describe the vulnerable relationship of humans to an agent (e.g., other human, automated system). In order to better understand the relationship of human related TIA, a review of basic trust concepts is necessary to avoid confusion among the different explanations of other trust-related concepts and behaviors. Castaldo, (2008) conducted a meta-analysis that reviewed 72 different definitions of trust to identify common constructs. The three primary elements common throughout the literature were: a subject, an action or behavior, and a future action or expectation. The future action is considered to be a distinctive and critical feature of trust. Trust is typically based on an agent's (i.e., person or system) past behavior and the requirement of anticipating some future action (Borum, 2010).

Hardin (2006) provides three components that are common across trust relationships. First, there must be a trustor-trustee relationship with something at stake. Next, the trustee must have an incentive to perform a task or action. Finally, there is some level of uncertainty or risk of failure related to performance of the action. In these cases, a trust-based relationship is required to facilitate a cooperative, exchange-based, relationship between the trustor-trustee. In a 1996 literature review of "The Meanings of Trust", McKnight & Chevany (1996) described the word *trust* as a homonym – a word with a variety of meanings or origins. With this in mind, a useful conceptualization of trust must be bound by context in discussion, research, and analysis applications. Two common applications of trust relationships fall under interpersonal trust and technology-based (e.g., automation) trust.

Interpersonal trust can be defined as "the perception you have that other people will not do anything that will harm your interest; the individual is giving the willingness to accept

vulnerability or risk based on expectations regarding another person's behavior" (Williams, 2013). Interpersonal trust is comprised primarily of a combination of cognitive (thinking) and affective (emotional) factors (Borum, 2010). A basic model is displayed in figure 2, based on definitions from Washington (2013).

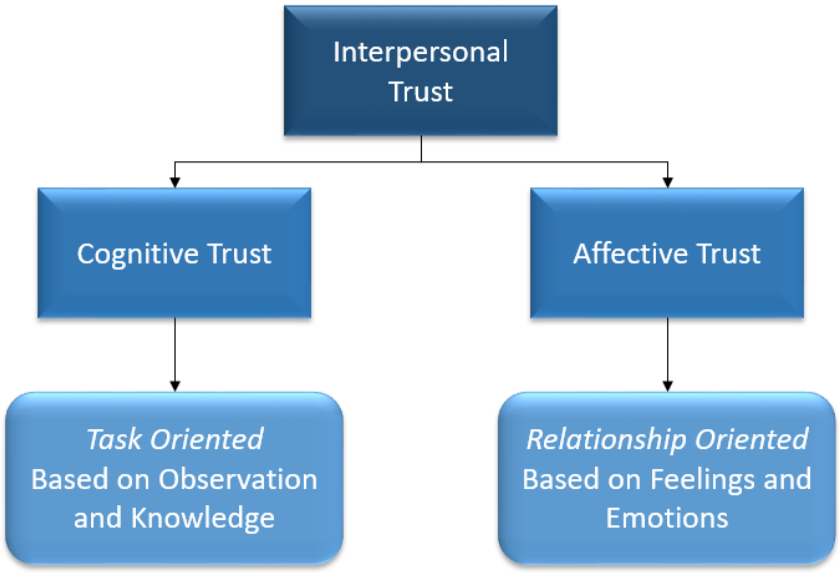

Figure 2 Interpersonal trust model

Washington (2013)

Cognitive trust influences a person's perception of trust from a baseline to either positive or negative expectations, based on utilizing knowledge of another person to forecast their behavior in a transaction, during a task, or in an environment (Borum, 2010). Higher levels of cognitive trust are based on knowing the trustee behavior well enough to feel confident that the trust relationship will not be betrayed. (Borum, 2010). Yang, Mossholder, and Peng (2009) state that cognitive trust is focused on task-oriented aspects of work and is established over numerous observations that establish a reputation.

Affective trust is often based on shared goals, beliefs, values, and identities of the agents involved, creating a connection between them (Borum, 2010). Colquitt, Scott, and LePine (2007) found that trustworthiness, disposition to trust, and emotional responses were the major determinants of trust. In general, affective trust is based on an emotional bond that goes beyond a *professional* relationship or prior knowledge of performance and often based on personal experiences (i.e., trusting because you like someone) (Washington, 2013).

Rotenberg et al. (2005) expanded the interpersonal trust model to include: domains of trust, bases of trust, and dimensions of the target trust. The bases of the framework consist of three fundamentals identified within interpersonal trust.

- Reliability – fulfillment of a promise.
- Emotional – reliance on others to refrain from emotional harm.
- Honesty – telling the truth and engaging in genuine behaviors that are not manipulative.

The domains of trust are split into two areas, cognitive/affective and behavioral. The cognitive/affective areas have been a significant point of interest for interpersonal trust research based on knowledge and emotions, while the behavior domain pertains to the tendency to rely on others. Finally, the two targets of trust that differentiate the bases and domains based on specific qualities of the trustee are specificity with a range from general people to a specific person and familiarity ranging from unfamiliar to very familiar. Figure 3 shows the expanded model of interpersonal trust.

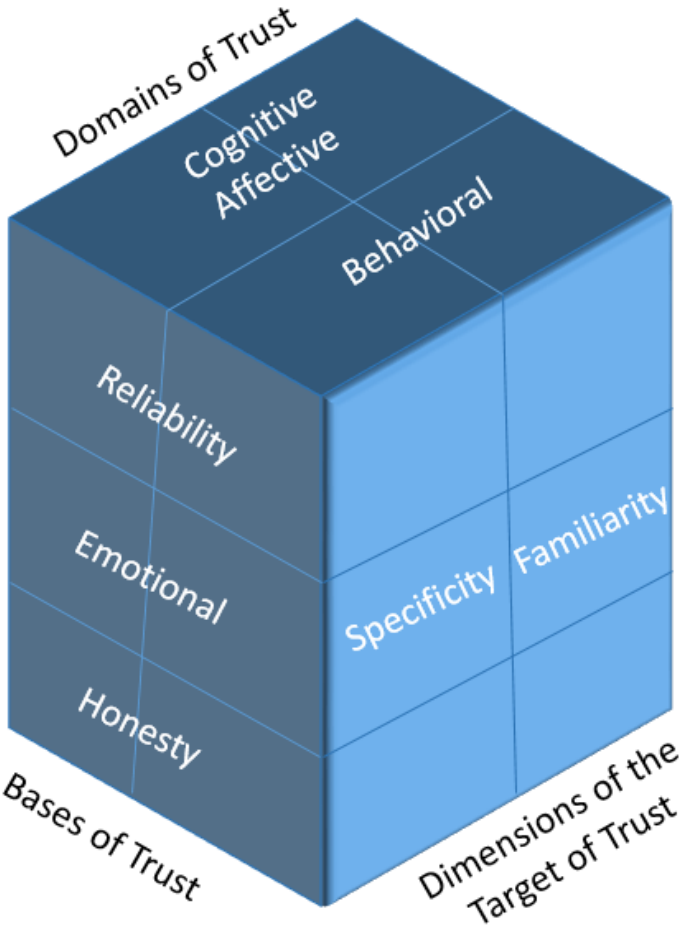

Figure 3 Expanded interpersonal trust model

Rotenberg et al. (2005)

Interpersonal trust is a useful concept to promote the foundational model of a trust-based relationship and understanding of a trust baseline for the ideas that underlie trust concepts. While there are several different usable definitions and descriptions of interpersonal trust, the primary concerns of this research are the trust relationships developed between humans and technology agents or automation.

Trust in Automation

TIA or trust in technology is similar to interpersonal trust but has some unique considerations and context that are more technologically oriented. While most of the interpersonal factors have some influence on the human TIA relationship, additional TIA factors related to automation limitations and behavior include reliability validity, utility, robustness, and false-alarm rate (Hoffman et al., 2013). These additional factors are often included when establishing and assessing user TIA.

TIA is a significant concern for technology developers as safety, performance, and use rate of an automated system can all be influenced by user perceptions of trust (Lee & See, 2004). "No trust, no use" is a principle statement that defines the importance of the impact of trust relationships and projected use between humans and automated systems (Schaeffer et al., 2016). A number of different TIA theoretical models are found in the associated literature and have been iterated over time to reflect the complexities of both general trust and TIA.

Korber, Baseler, & Bengler (2018) provides a model of trust based on dimensions from research conducted by Mayer, Davis, & Schoorman (1995) and Lee and See (2004) (figure 4). The Mayer, Davis, & Schoorman (1995) trust factors deal primarily with interpersonal trust, while Lee and See (2004) adapt the Mayer, Davis, & Schoorman (1995) model to connect the interpersonal factors to TIA considerations.

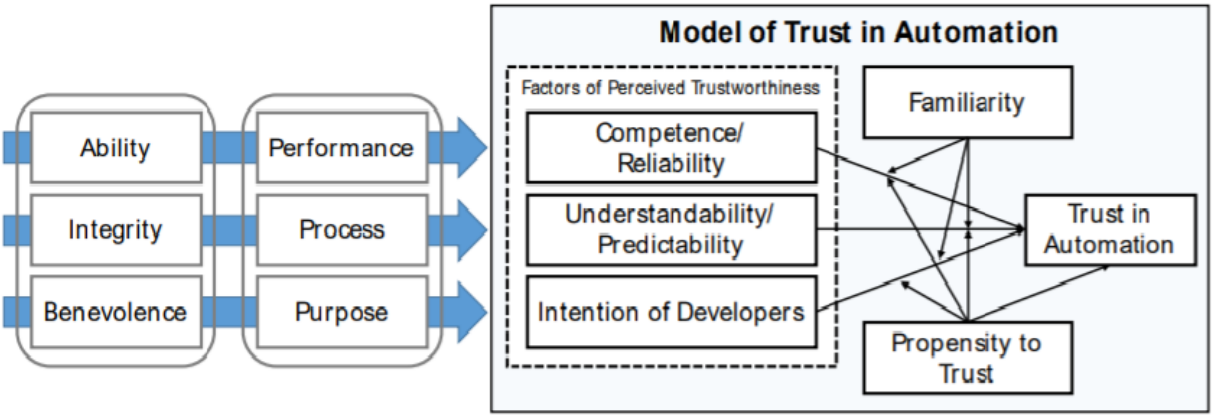

Figure 4 Adaptation model of interpersonal and automation trust

Korber, Baseler, & Bengler (2018)

Korber, Baseler, & Bengler (2018) summarizes the three key dimensions of this model that influence the overall TIA construct as *performance*, *process*, and *purpose. Performance* is an attribute that contains the system characteristics such as reliability, competency, and ability. *Process* refers to how the system operates and its appropriateness to accomplishing operator goals, with the primary characteristic of operator understanding of these operations. *Purpose* relates to the intention of the system design and its defined use cases.

In Hoff & Bashir (2015), the authors provide an analysis of three broad sources of human-automation trust. The human operator, the environment, and the automated system, which closely mirror dispositional, situational, and learned trust defined by Marsh & Dibben (2003). Dispositional trust can contain factors that are relatively stable over time and include culture, age, gender, and personality (Hoff & Bashir, 2015). Situational trust depends on external variability (e.g., workload, environmental setting), internal variability (e.g., self-confidence, expertise, attentional capacity, and affect), complexity of the system, and decisional freedom on how to use the automation (Hoff & Bashir, 2015). Learned trust is affected by preexisting

knowledge of the system, design features (e.g., ease of use, transparency), level of control over automated functions, and actual performance of the system (Hoff & Bashir, 2015). Figure 5 provides a visual representation of the dispositional, situational, and learned trust model.

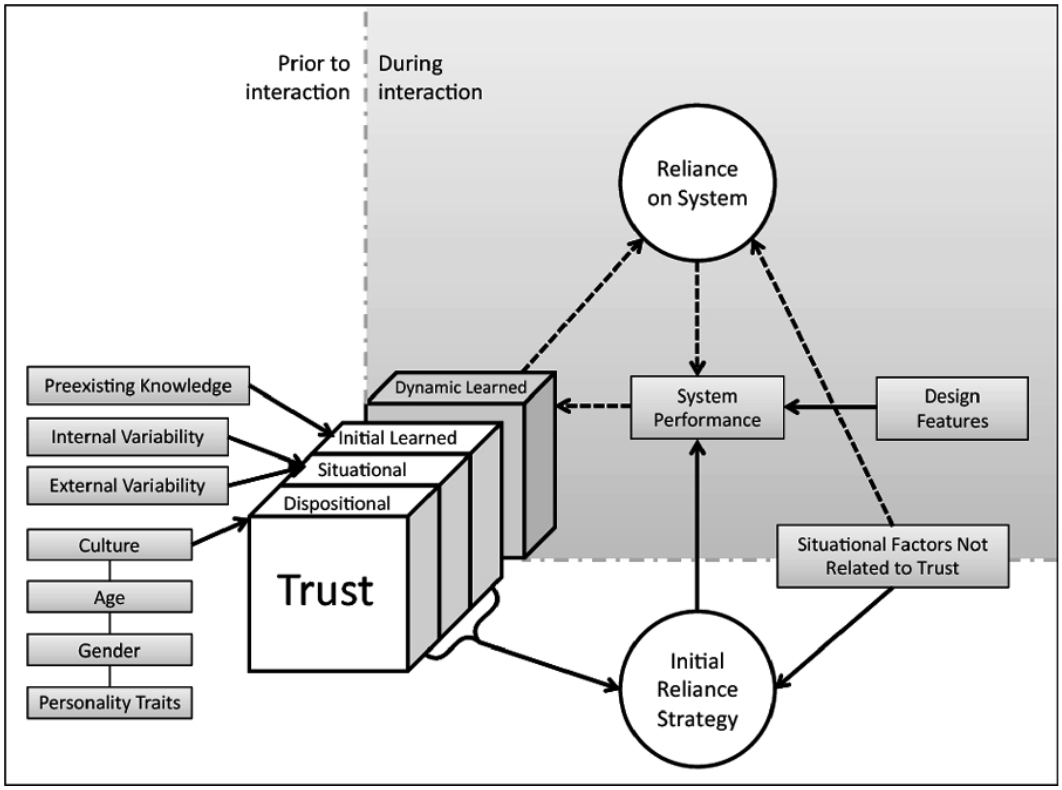

Figure 5 Model of dispositional, situational, and learned trust

Hoff & Bashir (2015)

Kraus (2020) developed a comprehensive model “Three Stages of Trust Framework” (figure 6) that represents the development of trust in automated systems by building trust through three sequential trust layers and three trust stages.

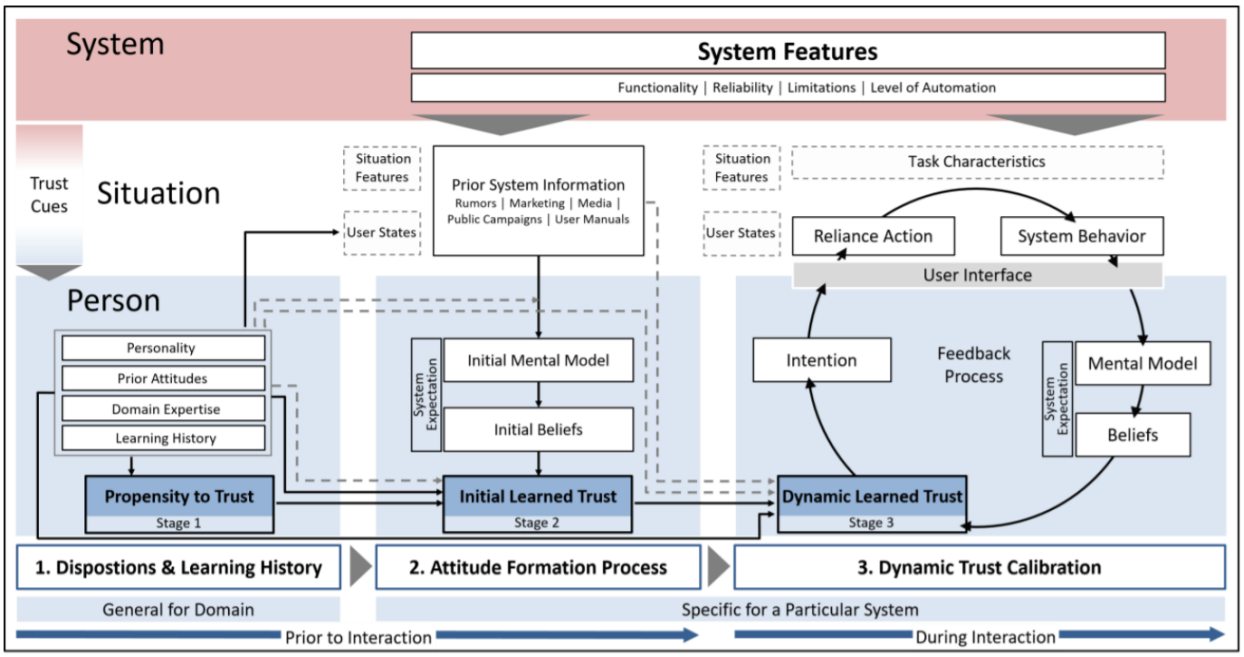

Figure 6 Three stages of trust framework

Kraus (2020)

The Krauss (2020) framework builds on the Hoff and Bashir (2015) model by providing a framework that contains three distinct variable categories – person, situation, and system - that affect trust development during familiarization with the automated system. As users interact with the system, the framework moves temporally from initial disposition to dynamic trust calibration. Both the Krauss (2020) and Hoff and Bashir (2015) models posit that the formation of trust is built iteratively over time and generally based on the following stages:

1. Initial propensity to trust the automation – based on personal characteristics, disposition, and previous history.

2. Initial learned trust – based on initial propensity along with new information about the automated system (e.g., training, marketing) prior to interaction.

3. Dynamic learned trust – based on system interactions and iterative development of trust calibration.

As illustrated in the previous theoretical models, since the Mayer, Davis, & Schoorman (1995) interpersonal trust model, there have been significant research efforts building on further development of theoretical trust models that attempt to incorporate the complexities of the trust construct. These TIA models showcase the difficult task facing human factors researchers and practitioners when attempting to measure TIA for human-automation interactions.

While interpersonal trust and TIA constructs are closely related, Madhavan & Wiegmann (2004) state that research has found critical differences in how people react to automation versus human advice. Initially, people lean towards a bias for trusting an automated system for elaborate information processing, which can easily turn into a bias against the automation as people are more observant of errors made by the automation when compared to humans. TIA is likely to breakdown more quickly than interpersonal trust relationships, due to the sensitivity of automation errors on perceived trust in the systems. These concerns emphasize the requirements for highly reliable automation systems to ensure appropriate trust levels are calibrated and established.

Reliance and Trust Calibration

While trust in a system is often associated with subjective perception from user interactions, there is another construct of objective reliance on the automation (Brzowski & Nathan-Roberts, 2019). Use, misuse, and disuse of automation as well as calibrating trust for appropriate reliance on automation have been the subject of many research studies to further understand and optimize human-automation systems (Parasuraman & Riley, 1997; Lee & See, 2004). Ensuring appropriate reliance on automated systems is vitally important to both system efficiency and safety. The underpinnings of automation reliance are based around mitigating the disuse and misuse of automation (Lee & See, 2004). Parasuraman & Riley (1997) describe

instances of misuse of automation as failure of appropriate monitoring of automated systems, improper use of an automated system, and an overreliance on the automation. While disuse is often associated with underreliance on automation systems, typically occurring when excessive false alarms are present in a system. When disuse occurs the user no longer utilizes the system to its full potential (Fallon et al., 2010). To avoid the negative impacts of under- and overreliance, users must develop an appropriate level of calibrated trust in the system (Johns Hopkins University, 2019).

Trust calibration occurs when users appropriately adjust their level of trust with the actual reliability of the system (Okamura & Yamada, 2020). Fallon et al. (2010) suggests that the calibration of trust between a human and a system can be viewed as a sensemaking process. Sensemaking is a process that helps users improve their awareness of uncertain and ambiguous situations (Fallon et al., 2010). As users interact with a system, they learn the circumstances in which the automation functions are seemingly unreliable and can adjust their trust in the system without misusing the automation (Fallon et al., 2010). Endsley (2015) provides a graphical model of an optimal calibrated trust curve based on system reliability, where correct trust falls between over- and under-trusting with respect to system reliability (figure 7).

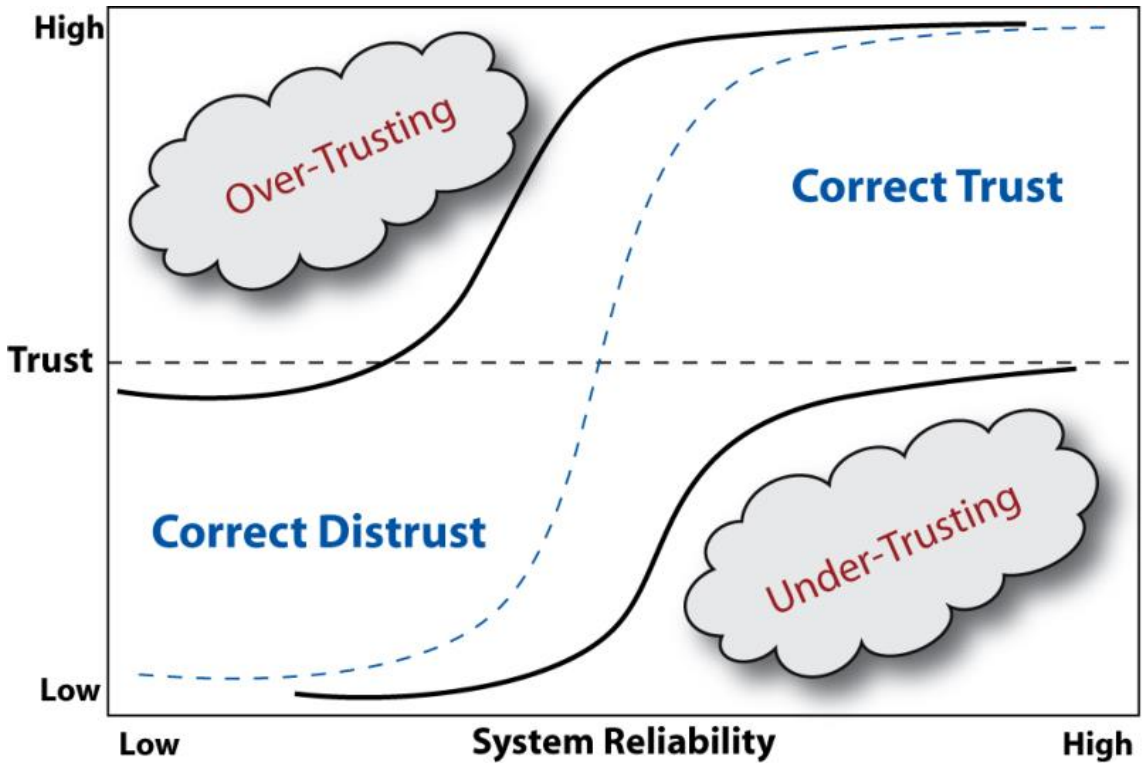

Figure 7 Model of optimal calibrated trust.

Endsley (2015)

Measurement of Trust in Automation

Measurement of TIA based on user perception often involves the use of subjective measurement techniques such as surveys. The process for survey development typically follows a general model that begins with developing a theoretical foundation for what the researchers want to measure (O'brien & Toms, 2010). Once the researchers gain an understanding of what needs to be measured, scale construction can begin, followed by a pre-test to *purify* the scale, and then a final scale evaluation prior to use (O'brien & Toms, 2010).

In general, the procedures for constructing a survey instrument for user perception (e.g., TIA) and usability are similar. Steps commonly found in research for user perception surveys include generating a theoretical framework for what to measure, generating an item pool of

questions or items for initial review, thorough review of those items using experts and existing literature, and finally a pre-test to ensure understandability and identification of any problems with the survey immediately present (O'brien & Toms, 2010). While some of the methods differ (e.g., previous research vs. experts for item pools), the intended outcome for each step is the same.

In most cases, the survey tool must contain and begin with a set of items or questions used to assess specific constructs under investigation. Survey development typically starts with a group of items identified for the area of interest. The initial items are usually obtained from literature review, existing instruments, and/or SME review (Mason et al., 2021; McNamara, 2020; O'brien & Toms, 2010; Simon, 2020). Oftentimes researchers will utilize existing validated scales to pool initial questions and terms to generate the item pool (Mason et al., 2021; McNamara, 2020; Simon, 2020). In some cases, item pools are generated by researcher expertise, existing domain models, rating scale surveys, and SME review (Christophersen & Konradt, 2012; Jian, Bisantz, & Drury, 2000; Schaefer, 2013). In many applications, the author's discretion is used to determine what constructs they are attempting to measure, which establishes the initial item or factor pools. Often, the author's leverage existing research heavily, and/or rely on participant and expert feedback. The initial factors must be reviewed by researchers and experts for face validity to ensure that the factors are appropriate for use in the survey as either uni- or multi-dimensional scale items.

The format of surveys for TIA are often based on a Likert scale. The Likert scale is one of the most common psychometric scales for the measurement of human attitude and often used in the TIA literature to measure TIA (Joshi et al., 2015). Likert scales are typically constructed based on point scales from 5 to 10 and designed to capture the participant perceptions or

opinions around a phenomenon under study (Joshi et al., 2015). Joshi et al. (2015) defines the purpose of the Likert scale as a tool to measure the phenomenon of interest called a 'latent' variable which is expressed through several items on the survey. The items on the Likert scale are mutually exclusive and measure specific dimensions of the latent variable. Ratings are often combined to produce a summated score which measures the phenomenon (Joshi et al., 2015). Figure 8 shows an example 7-point Likert scale survey question and response, with a neutral rating in the middle.

Statement:
Trust in Automation is an important indicator of the effectiveness of human-automation design.

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

Figure 8 Sample statement and Likert scale.

Measurement of Trust Calibration and Reliance

While user perception and perceived trust are commonly measured with subjective rating scales, trust calibration and reliance can be measured by utilizing objective measurement techniques. Objective data such as eye gaze, electrodermal activity, and error rate analysis can help to determine the actual reliance that a user is placing on the automation (Duez, Zuliani, & Jamieson, 2006; Wang, Jamieson, & Hollands, 2008). In Wang, Jamieson, & Hollands (2008) the author's review four methods that can be used to identify user reliance on a system:

- Consistency – related to number of instances users decide to follow automation feedback.
- Performance – differences in user performance when using automation.

- Behavior – visual scans, attention allocations, actions taken, and manual performance of tasks.
- Response Bias – related to signal detection theory and the user perception of response probability.

These methods utilize objective measures that can help to identify when user reliance is inappropriate, resulting in inappropriate trust levels. While a gold standard has not been established in a generalizable manner for the optimal reliance of systems, taken in context with self-reported trust ratings and researcher observation, conclusions can be drawn to identify whether operators trust and rely on the system in the intended manner.

Trust for Military Systems

The survey tool for this research is based on an extensive literature review, analytical processes, and expert interviews to determine the pool of items/questions that were initially formulated. A variety of research related to TIA tools, meta-analyses, and factor identification has already been performed (Jian, Bisantz, & Drury, 2000; Kaltenbach & Dolgov, 2017; Schaefer, 2013; Spain, Bustmante, & Bliss, 2008; Uggirala et al., 2004). This previous research provided a significant resource for a comprehensive literature review to pool items for an initial survey tool.

While some identified factors may be present in other research, TIA surveys or tools with the target population for validation being military users (i.e., pilots, system operators) are scarce. The intent of this research was to follow the general methodology of previously validated tools and adjust the target audience to the appropriate demographic for the specified military case (i.e., Army Aviation). This effort helped to accurately reflect contextual factors related to TIA for the specific military applications.

An example of why the specific tool is important for military applications can be found by examining the Jian, Bisantz, & Drury (2000) TIA survey. The survey was empirically validated, but the statements developed do not translate well to military applications. On a personal anecdote, distributing the survey to military users operating weapons systems has resulted in significant ambiguity in the intent of the statements. For example, the statement "The system provides security" is not relevant to a handheld map device. While cybersecurity concerns may be present, the consensus is that there is no expectation or action required for a handheld map device to provide security. The lack of system provided security doesn't make the handheld device less trustworthy. A similar concern has been echoed during military test events regarding the statement "The system behaves in an underhanded manner". The statement generally implies that the system is purposefully hiding information in a way to destabilize the intended action. This is a far more severe statement than "I'm unsure what the system is doing". While the Jian, Bisantz, & Drury (2000) scale is a validated scale, the item constructs do not necessarily pass face validity with a military demographic of users.

When considering the contextual basis for TIA, concepts such as dispositional, situational, and learned trust can weigh heavily on the influential factors that influence a particular user's trust level in an automated system. For example, when considering trust factors for FVL, some context is required to identify the potential primary factors that would influence the user-FVL relationship. Dispositional trust would likely be the least impactful, due to the target demographic of FVL users. FVL users will be heavily involved with using automation regularly and trained to rigid standards for automation use with the FVL system. It's unlikely that these users would maintain an adverse disposition towards automation.

Miramontes et al., (2015) examined training effects on TIA for Air Traffic Controllers (ATC). The study found that students that utilized more automation features during training were both more efficient and heavier users of the automation during air traffic management tasks. The ATC study (Miramontes et al., 2015) was interested in whether TIA could be trained for the future NextGen ATC environment. Based on the ATC study findings, an emphasis on automated systems and familiarity to promote trust calibration throughout FVL training would likely reduce any negative disposition towards automation systems in FVL, assuming reasonable reliability (Miramontes et al., 2015). While examining personal influences on trust formation in human-agent teaming, Huang & Bashir (2017) found that participants with a rational decision-making style showed higher levels of TIA. Jensen (1995) and the U.S. Federal Aviation Administration (FAA) (1991) examined pilot judgement and consider rational and systematic judgement as part of the decision-making process for pilots. These findings suggest that pilots for FVL will likely utilize rational decision-making styles and be heavily trained in automated system use, indicating that any initial negative disposition to TIA in flight school will likely be altered by extensive experience with the FVL systems. Flight school training will also provide an initial trust calibration for the FVL systems. This research is focused on system improvements identified by experienced aviators post-training and with professional and active knowledge of current aviation systems, implying a definitive trust calibration for baseline systems in which to compare new system upgrades.

Additionally, Borum (2010) suggests that institution-based structures (e.g., military specifications) that provide system oversight can influence the trustor by providing a sense of protection against deliberate harm. Contextual and situational factors may also increase a user's propensity towards either trust or suspicion of the system. These situational factors should be

considered when interpreting TIA ratings. Complex tasks in high workload environments could affect TIA ratings and user reliance, or trust calibration on the system, which would be a potentially common theme across the identified roles. However, the most likely TIA factor influences would be related to system characteristics and interactions (i.e., learned trust), due to the nature of aircraft related tasks being heavily dependent on formal system interactions (e.g., checklists, standard flight procedures). Establishing an initial framework of potential influential trust factors for a given task, can help to focus data collection efforts and data-driven decisions.

Utilizing the past research is extremely helpful in determining the initial factor/item pool and validation procedures, but simply repurposing an existing survey without a significant effort to establish face validity with the proposed population, can result in suboptimal results. The initial research expectation was that many of the factors identified in a systematic literature review would be present in an initial survey tool, with a significant emphasis placed on appropriate demographic terminology for face and content validity.

Initial Review of Trust in Automation Factors

An initial review and categorization of factors that influence TIA was conducted through a comprehensive literature review. The literature review was used to identify the TIA factors that were found throughout TIA research papers with the intent of showing the breadth of factors used in a wide range of TIA research and to utilize these factors as an initial item pool for TIA scale development.

Literature search

The comprehensive literature review was conducted using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist (PRISMA, 2021). Searches

were initiated using the keywords in table 1. Search engines included: Academic Search Premier, University Discovery Service and On-Line Catalog, EBSCOhost, and Google Scholar. No restrictions (e.g., dates, journals) were placed on the search results. Additional searches were conducted by cross-referencing related documents within research articles. Figure 9 shows a graphical representation of the systematic review process.

Table 1 Literature review search terms

Literature Review Search Terms
Trust in Automation
Reliance on Automation
Levels of Automation
Human Trust in Automation
Trust Factors of Automation
Trust in Automation Questionnaire/Survey

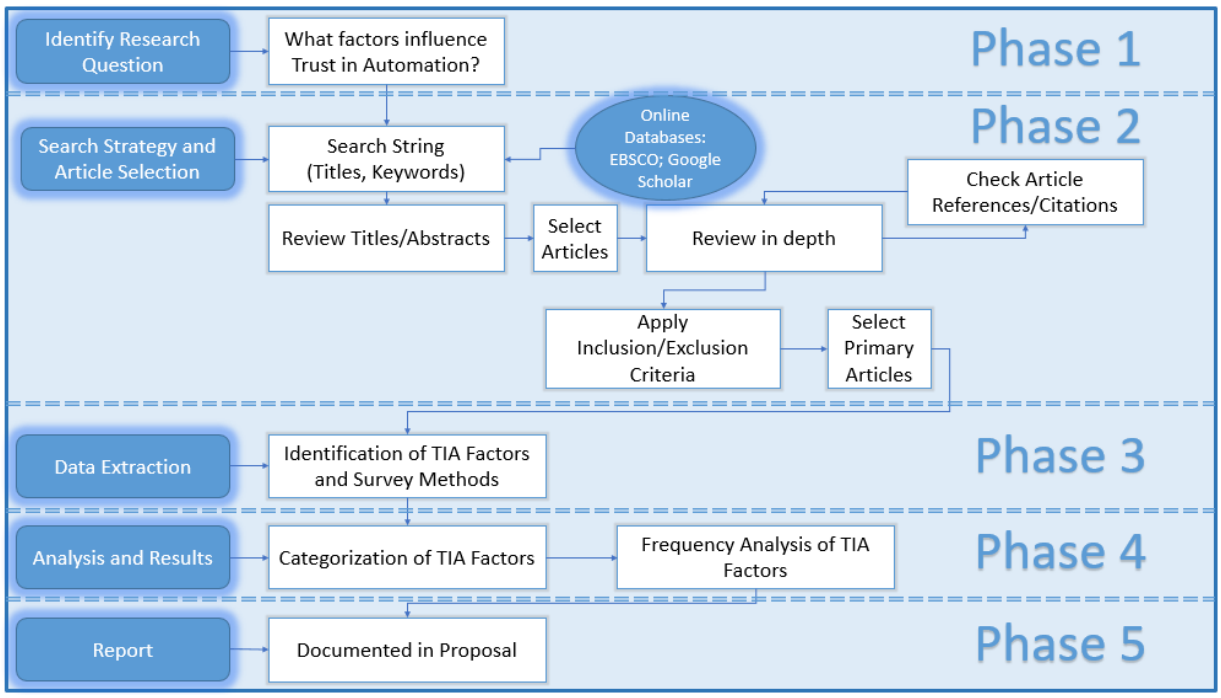

Figure 9 Literature review process

Exclusion and Inclusion Criteria

To be included, studies had to meet the following criteria: (a) present empirical research or a meta-analysis/systematic review of factors that influence TIA; (b) present data collection (e.g., surveys, scales) methods for user responses related to TIA; (c) examine personal traits and/or influencers that affect TIA.

Data Extraction

Data extraction was conducted by identifying factors and/or characteristics in each reviewed research study that contributed to identification or measurement of TIA. Subtleties within the research methods were not heavily considered, as the initial step was to identify gross factors that characterize TIA. Research biases were not considered to influence this analysis, due to the intentional inclusion of a wide variety of methods and resources.

Grouping

TIA characteristics and factors identified throughout the research articles were grouped into descriptive categories representative of the context for each characteristic. Labels were provided for each category and the entire group was designated as “Factors that affect user trust in automation”. Subfactors were identified as human perception factors and automation (performance) factors.

Literature Review Procedure

Microsoft Excel was used to group factors based on the notional TIA categories, followed by a frequency analysis of factor occurrence in literature to determine the most common and frequently researched TIA factors. Research sample sizes and methodology were not heavily considered due to the lack of standardization in experimental research methods in the identified literature. Initial in-depth statistical analysis was not considered applicable, as the intent was to identify any factors that contribute to trust in automation in a variety of experimental scenarios.

Literature Review Sample

The literature review began with 568,151 results, many of which were considered non-applicable due to being presented based on partial key word hits or duplicate results. After conducting an initial abstract and title screening of 824 relevant articles, 708 articles were excluded based on irrelevant information. One hundred and sixteen (116) articles were reviewed for more in-depth information. After a full manuscript review, seventy-eight (78) articles were used in the systematic review to identify factors that affect trust in automation. Most of the articles used were experimental research, with additional articles related to meta-analysis or

systematic reviews. Figure 10 shows a flow-chart diagram of the literature review based on the PRISMA reporting recommendations (PRISMA, 2021).

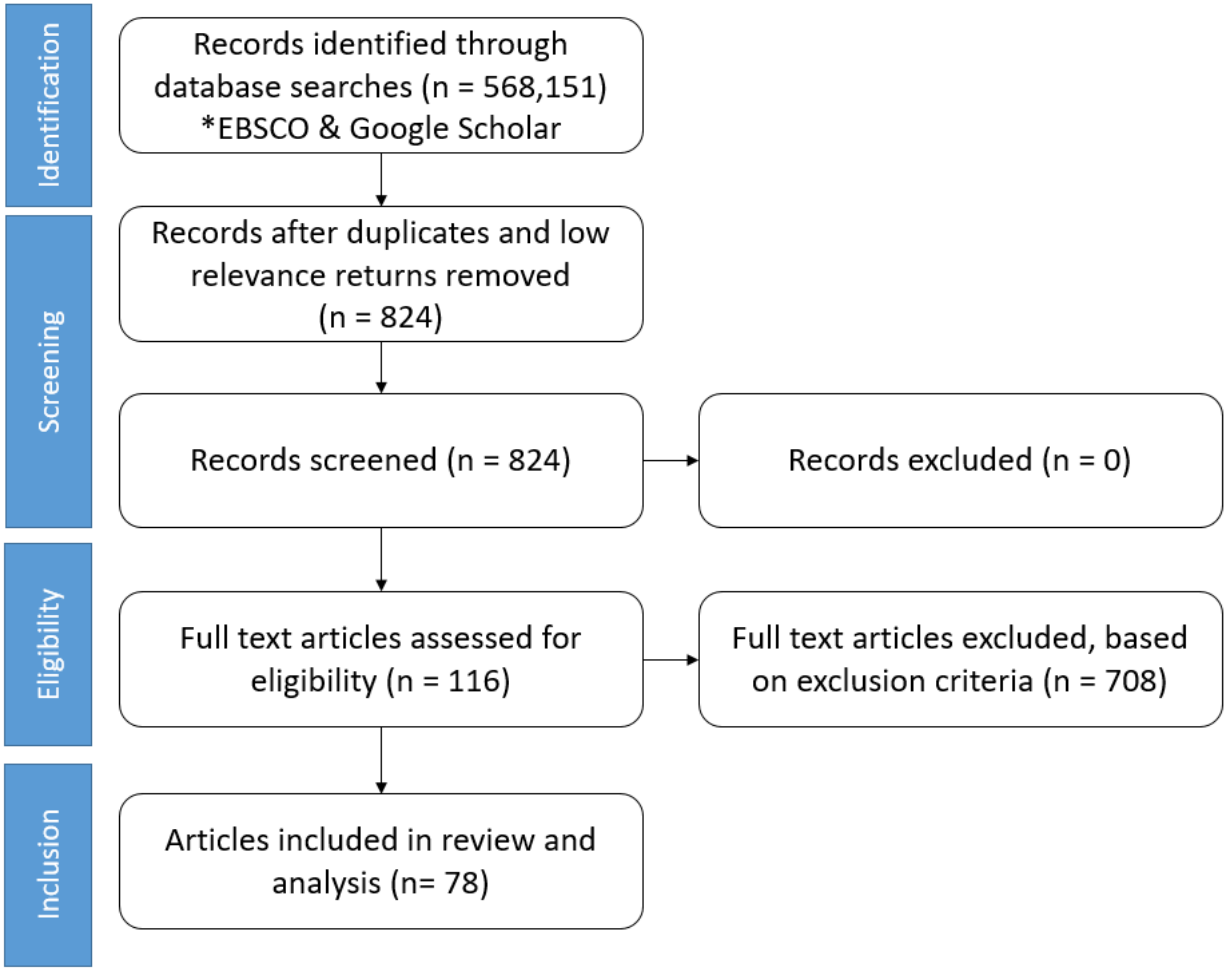

Figure 10 PRISMA Diagram

Results

Research Summary

Table 2 displays a summary of the articles that were examined to determine factors that influence trust in automation. Identified factors were based on experimental research or meta-analysis from studies that examined TIA factors or provided surveys for measuring TIA. Authors and identified factors were included in table 2. The identified factors were used in the summary analysis to determine the frequency of factors within the literature and to provide a framework for survey development.

Table 2 Identified factors and associated authors

Human Factors	
Author	**Identified Factors**
Dolgov et al. (2017) Hoff & Bashir (2015) Johnson et al. (2004) Kaltenbach & Dolgov (2017) Gao & Lee (2006) Niu et al. (2018) Madhavan, Wiegmann, & Lacson (2006) Higham et al. (2013) Jian et al. (2000)	Confidence
Hoff & Bashir (2015) Gao & Lee (2006) Chancey et al. (2017) Johnson et al. (2004) Lee & See (2004) Ho et al. (2017) Korber, Baseler, & Bengler (2018)	Purpose and Intent
Lyons et al. (2016) Gao & Lee (2006) Ho et al. (2017) Cassidy (2009) van Dongen & van Maanen (2013)	Transparency

Table 2 (continued)

Human Factors	
Author	**Identified Factors**
Balfe, Sharples, & Wilson (2018) Dolgov et al. (2017) Chancey et al. (2017) Kaltenbach & Dolgov (2017) Korber, Baseler, & Bengler (2018) Uggirala et al. (2004)	Competence
Dolgov et al. (2017) Kaltenbach & Dolgov (2017) Gao & Lee (2006) Jiang et al. (2004) Uggirala et al. (2004)	Faith
Balfe, Sharples, & Wilson (2018) Dolgov et al. (2017) Kaltenbach & Dolgov (2017) Chien et al. (2016)	Understandability
Dolgov et al. (2017) Kaltenbach & Dolgov (2017) Higham et al. (2013) Jian et al. (2000)	Familiarity
Lyons et al. (2016) Jiang et al. (2004) Korber, Baseler, & Bengler (2018) Uggirala et al. (2004)	Predictability
Shaefer et al. (2016) Sanchez et al. (2011) Hoff & Bashir (2015)	Demographics
Shaefer et al. (2016) Dolgov et al. (2017) Kaltenbach & Dolgov (2017)	Personal Attachment
Parasuraman & Miller (2004)	Risk

Table 2 (continued)

Automation Factors	
Author	**Identified Factors**
Balfe, Sharples, & Wilson (2018) Dolgov et al. (2017) Lyons et al. (2016) Lyons et al. (2017) Kaltenbach & Dolgov (2017) Jeong et al. (2018) Cassidy (2009) Jiang et al. (2004) Madhavan, Wiegmann, & Lacson (2006) Parasuraman & Miller (2004) Wang, Jamieson, & Hollands (2009) Rice (2009) Cafarelli & Hansman (1998) Korber, Baseler, & Bengler (2018)	Reliability
Shaefer et al. (2016) Balfe, Sharples, & Wilson (2018) Mishler et al. (2017) Merritt et al. (2012)	Feedback
Jeong et al. (2018) Higham et al. (2013) Charalambous, Fletcher, & Webb (2015)	Safety
Ho et al. (2017) Jeong et al. (2018) Cafarelli & Hansman (1998)	Usability
Wang, Jamieson, & Hollands (2008) Lee & See (2004) Jeong et al. (2018)	Effectiveness
Dolgov et al. (2017) Jian et al. (2000)	Integrity
Jeong et al. (2018) Parasuraman & Miller (2004)	Accuracy
Jeong et al. (2018)	Suitability

Frequency Analysis

Frequency analysis was conducted to identify the most commonly occurring TIA factors that were examined or used during experimental research and other systematic review/meta-analysis reports. Thirty-eight (38) articles were used to identify trust in automation factors. Table 3 shows the percentage of TIA factor appearances, when compared to other factors in each article based on two notional groupings (human and automation). Essentially, the percentage indicates the ratio of literature occurrences of the TIA factor when compared to the sum of instances of all identified factors for each subgroup. The intent is to show a relative strength of use for each factor when defining TIA characteristics. While identification frequency does not necessarily indicate importance, identification does show that a significant amount of research has been conducted using the particular factor in TIA studies. The identified factor has been listed as recording some influence on TIA in the associated study.

Identified factors were grouped according to two (2) overall subcategories, based on similar attributes. Human factors include perceived personal attributes such as confidence, competency, understandability, and familiarity. Automation factors include characteristics of the automation such as reliability, usability, feedback, and accuracy.

Associated definitions are provided for each term based on previous research studies that considered the item with the definition to have some effect on TIA. An established definition list is important to ensure consistency across terms, especially during survey data collection. Users should be rating a system based on shared context of what each item is expected to measure.

Table 3 Frequency analysis

Human Factors	
Percent of Factor Appearance	**Identified Factors**
17%	Confidence
15%	Purpose and Intent
13%	Transparency
11%	Technology Competence
9%	Faith
7%	Understandability
7%	Familiarity
7%	Predictability
6%	Demographics
6%	Personal Attachment
2%	Risk
Automation Factors	
Percent of Factor Appearance	**Identified Factors**
51%	Reliability
11%	Feedback
9%	Safety
9%	Usability
6%	Effectiveness
6%	Integrity
6%	Accuracy
3%	Suitability

Factor Definitions

Human Factors refer to the inherent individual characteristics of the user and their perceptions of the automation (Henshel et al., 2015).

- *Confidence* refers to the perception of one's ability to effectively interact with the system in a consistent manner (Dolgov, et al., 2017).

- *Purpose and Intent* refers to the user's knowledge of the use case of the automation and its intended actions (Sheridan, 2019).

- *Transparency* can be described as the capability of the automation to provide information to the user on its current state and behavior to assist in user understanding (Westin, Borst, & Hilburn, 2016).

- *Technology Competence* refers to the perceived technical competence of the system to do the task at hand (Miller & Perkins, 2010). Users of the automated system are able to judge the outcome of task related events to determine automation competence and identify appropriate use cases (Miller & Perkins, 2010).

- *Faith* refers to confidence that the automation will perform the intended actions (Miller & Perkins, 2010).

- *Understandability* implies that the user knows how and why the automation is performing specific tasks (Sheridan, 2019).

- *Familiarity* references past experiences of the user with the automation, supposing some historical context for how the automated system works (Sheridan, 2019).

- *Predictability* refers to the matching of the automation performance with the user expectations. When the user is able to predict the automation actions, the user can determine when the automation may fail and adjust their own performance to accommodate (Miller & Perkins, 2010).

- *Demographics* contains the factors of culture, age, gender, and personality and refers to their association to propensity for user trust in automation (Hoff & Bashir, 2015).

- *Personal Attachment* references user agreement that the automated system is agreeable in use and suits personal taste (Chien et al., 2014).

- *Risk* refers to the situational use of the automation in hazardous conditions (Perkins et al., 2010).

Automation Factors refer to the situational characteristics of the automation outside of the user individual characteristics (Henshel et al., 2015).

- *Reliability* implies that the automation maintains consistent performance free of variation or contradiction (Miller & Perkins, 2010).

- *Feedback* refers to the information provided from the system related to the outcome of actions and contextual *future* actions (Schaeffer et al., 2016).

- *Safety* implies that the system outcomes do not create unacceptable hazardous conditions for the user (FAA CHP 8, 1991).

- *Usability* is the extent to which the system can be effectively used to satisfactorily accomplish specified goals (Lecerof & Paterno, 1998).

- *Effectiveness* refers to the extent to which a system can complete its mission under established constraints (Dordick, 1965).
- *Integrity* refers to the degree to which the automated system adheres to a set of established principles (Lee & See, 2004).
- *Accuracy* is how often the automated system makes a correct decision (Nourani, King, & Ragan, 2020).
- *Suitability* refers to the appropriateness of the automation capabilities to carry out the tasks (Smith, Allaham, & Wiese, 2016).

Other words such as *dependability* were also used in the TIA literature to describe factors. In this research, the underlying factors were used in the analysis. For example, dependability is often associated with the system reliability and the occurrences were combined with the *reliability* factor (Rudiger, Wagner, & Badreddin, 2007).

Discussion

A significant amount of research has been conducted for identifying factors that influence trust with respect to both interpersonal and human-automation trust. However, as the literature review confirmed, minimal research has been conducted on both military and pilot specific populations related to TIA. Following the comprehensive literature review of this research, many of the factors identified that influence generalized trust constructs also have the potential to significantly affect military aviation system trust. *Human factors* of trust such as user confidence, understanding, and predictability of automated systems are very likely to play a role in military aviation system trust. Pilots are highly reliant on automated systems to keep them safely in the air and to assist in carrying out required missions (Freedberg, 2020). Influential a*utomation factors* such as reliability, usability, and accuracy are imperative for successful pilot mission performance. Poor automation-system performance and degraded usability can often result in poor mission outcomes (SKYbrary, 2021). However, other factors such as personal

attachment or demographics may have an impact on aviation system trust. Personal attachment would likely be irrelevant for military aviation systems, since pilots are often only provided one option for a system to complete a specified task. Personal taste or preference is considered in early testing, but end users are not often provided with the option to adjust the system preferentially. Demographic factors relate to dispositional trust, and for general aviation, some generational differences among pilot TIA have been found related to initial trust in automated aviation systems (Leadens, 2020). In Leadens (2020), younger pilots considered automation management as a fundamental part of the pilot's skill set. When considering military aviation systems, dispositional trust factors would likely have even less of an influence on the initial perception of new automated systems due to the standardized training and significant experience with automated systems that modern military pilots have attained as part of their fundamental skillset.

The factor identification was an important first step in understanding TIA for Army Aviation systems. Analyzing previous research helped to establish a baseline of TIA terminology and concepts that were further explored for application to pilots. Previous research in TIA has generally focused on either broad concepts or particular applications of automatic processes that may or may not be generalizable to the Army Aviation community. Since no standard currently exists for a TIA perception survey for pilots, the literature review provided an important starting point for development of a targeted survey tool.

While many of the factors of TIA are likely to overlap from more generalizable studies, a more nuanced approach of identifying factors important to Army Aviators is necessary to ensure pilots understand the TIA concepts that they are rating and consistent terminology is being used across Army Aviation testing. Using different surveys or terminology during testing can cause

difficulty when comparing across multiple iterations of system development. By using a standardized survey, previous iterations of design can be appropriately compared throughout the system lifecycle to ensure product changes do not negatively impact user trust when compared to previous systems or iterations. This type of standardization is very helpful for human factors engineers and program managers that are responsible for assessing and meeting system requirements for suitability. A consistent assessment methodology based on the comprehensive literature review of TIA factors can help to establish the survey used for data collection, recommendations for addressing deficiencies based on TIA factor domains (i.e., human vs. automation), and begin to develop baseline trust ratings for comparison to future iterative changes.

The initial factor list provided a baseline for SMEs to review for relevance to the aviation domain and military applications. While the frequency analysis was interesting in providing likely contributors to aviation system trust, further review and systematic analysis were required to ensure relevant factors are utilized for survey tool development. A thorough review through SME interviews and decision-based analysis (i.e., Analytic Hierarchy Process) of the identified factors helped to build and validate a survey instrument that could be used to evaluate the relevant factors during representative mission scenarios that utilize new automation in aviation systems.

Conclusion

In order to identify the key factors that influence TIA for Army Aviation systems, a comprehensive literature review was conducted to establish a historical background of TIA research, identify key concepts in human and automation trust, and to develop an initial list of factors that could potentially impact pilot TIA. One hundred and sixteen (116) articles were

reviewed and the factors that influenced TIA were categorized under key terms as defined by TIA literature. A frequency analysis was used to categorize the factors under *human* and *automation* factors that influence TIA, based on either individual characteristics of the user or situational characteristics of the automation. The frequency analysis found that factors such as automation reliability and user confidence in the system were prevalent throughout the literature. By identifying prominent factors, an initial pool of items can be established for development of the Aviation Systems – Trust Survey (AS-TS).

The factor identification and definitions allowed evaluation by SMEs to begin survey validity testing in order to establish the AS-TS survey for the Army Aviation demographic. The next steps for the research were to utilize SMEs to review the identified factors and provide comparative ratings for each factor through a decision analysis method. The results established face validity for the factors to be included in the initial survey. Validation of the survey was conducted with military pilots using representative scenarios of automation performance. Throughout the validation process the survey was refined and tested for validity and reliability, with the intent of developing a robust survey tool to analyze the trust relationship that pilots establish with automated systems in current and future aircraft. Identifying deficiencies or lower scoring items in the trust relationship can help human factors engineers and program managers focus on improving and calibrating the trust relationships to optimize the human-automation team.

CHAPTER III

SURVEY CONTENT VALIDITY

Introduction

Modern pilots deal extensively with autonomous systems in the aircraft, with many more innovations on the horizon. The U.S. Army is especially concerned with pilot TIA, as new technologies such as the Army FVL program are rapidly approaching. Korber, Baseler, & Bengler (2018) states that TIA is a key determinant for the adoption of automated systems and their appropriate use. The FVL program is one of the Army's three major modernization priorities and includes initiatives for both a FARA and a FLRAA (Mayfield, 2021). A significant design consideration for FVL is optimizing the human-automation interactions that will occur during system use. FVL is required to provide multiple levels of supervised autonomy within the aircraft (e.g., autonomous takeoffs/landings, cueing, and adaptive interventions). Appropriate levels of TIA for the pilots will be extremely important for FVL platforms due to a significant focus on automated processes and automated assistance in high-speed and DVE rotorcraft operations (Freedberg, 2020).

Problem Statement

There is currently no standard methodology that is in use for the Army to assess TIA for pilots as a holistic measurement that identifies trust deficiencies and the relationship of trust to user reliance with follow-up actions. Subjective TIA measurement is typically considered to be perception-based and often utilizes a survey-based measurement tool for establishing user

perception of trust of an automated system (Parasuraman & Riley, 1997). Perception-based survey validation follows a general method of developing a theoretical foundation for what the researchers want to measure, gaining an understanding of what needs to be measured, scale construction, followed by a pre-test to *purify* the scale, and then a final scale evaluation prior to use (O'brien & Toms, 2010).

To address the lack of a standardized survey tool for TIA measurement in Army Aviation, a comprehensive literature review was conducted to identify key factors that may influence TIA in aviation systems (Chapter II). These identified factors were used as a foundation to develop an initial pool of factors for review by SMEs. Three studies were defined to identify the TIA factors that influence pilot trust in automation systems and to validate a data collection survey that measures the influence of the factors when pilots use automated systems.

The first study established content validity of the proposed survey. SMEs are often used to provide content and face validity during the initial development of perception-based surveys through both interview and in this case, use of the AHP decision-making method to refine the pool of factors to establish content validity for a TIA survey tool specialized for use in Army Aviation system assessments. Building on the previous literature review of identified factors, utilizing the AHP and SME interviews to solicit pilot input helped to answer the research question "What factors influence TIA for Army pilots using Army Aviation systems?" and provide initial content validity to a notional TIA survey.

The first hypothesis for this study posits that the AHP decision-making tool will be effective in identifying critical TIA factors and establishing face and content validity of the proposed TIA factors for inclusion on an initial aviation systems trust survey.

The second hypothesis for this study is that many of the key factors identified within the initial literature review (Chapter II) will likely be considered relevant factors for Army Aviation pilots, with the exception of personal attachment and demographics. Personal attachment and demographics focus on the user characteristics, rather than the automation performance or interface. Army pilots are often trained to specific standards and are required to accept automation regardless of their initial disposition, which likely makes them less concerned about their personal attachment to a system or any biases that may be perceived by demographic descriptors such as age or gender.

Literature Review

General Survey Development Method

Perception-based survey development often follows established research methods to identify factors and validate the instrument. Based on a literature review of perception-based survey development, the following paragraphs identify commonalities within research that address the development of perception-based surveys.

Summated ratings scales (e.g., Likert Scale) are typically used in research to examine the attitudes or feelings of participants to a particular stimuli or interaction through a questionnaire (Desselle, 2005). Likert scales are typically constructed based on point scales from 5 to 10 and designed to capture the participant perceptions or opinions around a phenomenon under study (Joshi et al., 2015).

Measurement of TIA based on user perception often involves the use of subjective measurement techniques such as surveys. The process for survey development typically follows a general model that begins with developing a theoretical foundation for what the researchers want to measure (O'brien & Toms, 2010). Once the researchers gain an understanding of what

needs to be measured, scale construction can begin, followed by a pre-test to *purify* the scale, and then a final scale evaluation prior to use (O'brien & Toms, 2010).

In most cases, the survey tool must contain and begin with a set of items or questions used to assess specific constructs under investigation. Survey development typically starts with a group of items identified for the area of interest. The initial items are usually obtained from literature review, existing instruments, and/or SME review (Mason et al., 2021; McNamara, 2020; O'brien & Toms, 2010; Simon, 2020). The initial scale factors must be reviewed by researchers and experts for face validity to ensure that the factors are appropriate for use in the survey as either uni- or multi-dimensional scale items.

A pre-test is conducted to reduce the pool of items that generate the notional scale. Pre-test techniques include presenting the proposed item pool or Likert statements to participants, evaluating the item pool participant ratings for word similarity, examining participant ratings for ease of use and understanding (e.g., cognitive interviews), target audience review, and SME review (Jian, Bisantz, & Drury, 2000; O'brien & Toms, 2010; McNamara, 2020; Ryan, 2009; Salminen et al., 2020; Schaefer, 2013). Once the initial survey is pre-tested, the *purified* scale is ready for validation testing (O'brien & Toms, 2010).

Face and Content Validity

Survey tools are often scrutinized for face and content validity and then tested for construct and/or concurrent validity to ensure the survey captures and measures the required information. The first step in the validation process is to reduce the item pool, if necessary, to reasonable and contextual items that meet the research intent and are applicable to the target audience. This process is often called "face validity", where SMEs and pre-testing help to bound the item pool for a manageable survey instrument (Hermann, Bager-Elsborg, & Parpala, 2017;

Mason et al., 2021; McNamara, 2020). Like face validity, content validity is a more formal process of determining whether the survey represents all relevant aspects that need to be investigated (Mason et al., 2021; Salminen et al., 2020; Witteman et al., 2021). Content validity is established through feedback processes with SMEs and intense literature review (Salminen et al., 2020; Witteman et al., 2021).

Analytic Hierarchy Process

In order to develop an assessment method for the perception-based measurement of trust among U.S. Army pilots, the research question: "What factors influence TIA for Army pilots using Army Aviation systems?" must be explored. A comprehensive literature review of generalizable TIA factors was established in Chapter I of this research and is provided in table 4.

Table 4 TIA identified factors

Identified Factors	
Human Factors	**Automation Factors**
Confidence	Reliability
Purpose and Intent	Feedback
Transparency	Safety
Technology Competence	Usability
Faith	Effectiveness
Understandability	Integrity
Familiarity	Accuracy
Predictability	Suitability
Demographics	
Personal Attachment	
Risk	

Identifying the most relevant factors for Army pilots was the next step in development of a survey instrument. As previously mentioned, two prominent ways of reducing the item list

involve subjective author discretion and SME review. However, a quantitative approach to identifying *critical factors* can be found in the decision-making literature related to the AHP. Utilizing the AHP can help to rank-order the factors based on perceived importance and help to further refine the factor list by ensuring face validity. By using AHP in combination with SME review, a more robust methodology can be used to establish the initial survey items.

AHP was introduced by Saaty (1980), as a multi-criteria decision-making tool that considers both qualitative and quantitative measures. The AHP model is based on a hierarchical structure where the initial goal is identified, followed by criterion for judgement, and then comparison of alternatives (Taherdoost, 2017). The AHP is often used to compare among choices (e.g., Is Car 1 preferred over Car 2?) based on a set of defined criteria that are weighted by respondents. For the car example, criteria could include paint color, passenger space, or cost. The attribute weightings and responses determine which car is preferred. Respondents utilize an "Importance Scale" to indicate preference or relative importance when comparing criteria in a matrix. The Importance Scale provides a two-sided direct comparison between two factors, where "1" is a rating that suggests the factors are equally important for the decision, and "9" in either direction indicates an extreme preference for the factor receiving the "9" over the compared alternative factor. Taherdoost (2017) provides an example Importance Scale (figure 11) and a sample AHP questionnaire for matrix comparison among cybersecurity factors when selecting software (figure 12).

Importance Scale	Definition of Importance Scale
1	Equally Important Preferred
2	Equally to Moderately Important Preferred
3	Moderately Important Preferred
4	Moderately to Strongly Important Preferred
5	Strongly Important Preferred
6	Strongly to Very Strongly Important Preferred
7	Very Strongly Important Preferred
8	Very Strongly to Extremely Important Preferred
9	Extremely Important Preferred

Figure 11 AHP Importance Scale

Taherdoost (2017)

Factor	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Factor
Privacy	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Reliability
Privacy	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Validation
Privacy	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Verification
Privacy	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Integrity
Privacy	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Confidentiality
Privacy	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Availability

Figure 12 Example AHP survey

Taherdoost (2017)

Figure 13 provides a hierarchical layout of the initial proposal of factors for evaluation and comparison by SME's related to the Army Aviation use case. While similar in concept to the traditional AHP method, the comparison of factors among each other is of particular interest in survey development. The proposed AHP method identifies critical factors but stops short of

comparing the factors to a specific technology. Instead, the factors are used to identify the importance level of each factor relative to the hierarchical goal. For the case of TIA, identifying the *critical factors* helps to establish face validity of the final survey, rather than utilizing the AHP method of alternative selection for a specific technology.

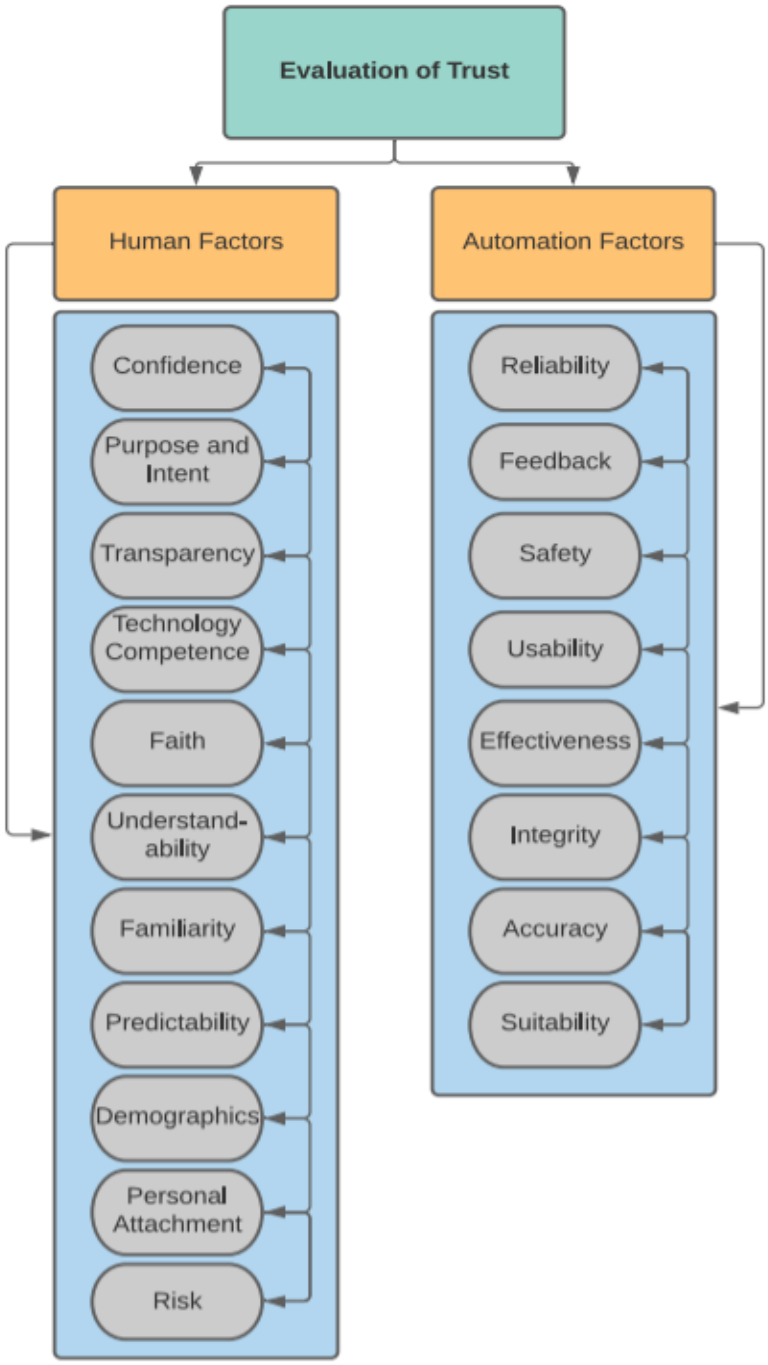

Figure 13 AHP hierarchy of trust factors

This type of application is found in a study conducted by Garg et al., (2012) which utilized the AHP approach to "identify and evaluate the critical success factors which make the

dimensions for measuring customer experience in banking organizations more effective and purposeful". This same type of logic can be applied to identifying TIA influence factors for systems. In the Garg et al. (2012) study, different dimensions of customer experience were identified through literature review and solicitation of a panel of experts to establish the critical factors that influence the banking customer experience. After initial identification, these attributes (dimensions) and sub-attributes (factors) were subject to pairwise comparisons for development of priority matrices and vectors. Final calculations were made, and the priority-levels of critical success factors were generated. In the Garg et al. (2012) research, the highest priorities affecting the banking customer experience were identified as convenience, employee interactions, and online functional elements.

The Garg et al. (2012) study is interesting, in that the AHP approach allows for identification of the *priority-level* of specific factors that influence the measurement. In the case of TIA, the literature provides numerous factors that potentially influence TIA (table 4). An AHP approach can help to reduce the selection pool of the factors, based on priority from the target demographic. For example, personal attachment is a factor that can influence TIA, but may not be relevant for the military demographic (Natarajan & Gombolay, 2020). By subjecting the identified factors to a panel of experts in military technology, the AHP approach can identify low ranking factors and establish an initial *face validity* for a proposed survey tool. The AHP results can also help product designers attempting to identify which TIA factors are most important to the target audience to help promote a positive user experience.

Methods

Subject Matter Expert Participation

A SME can be generally defined as "an individual who, by virtue of position, education, training, or experience, is expected to have a greater-than-normal expertise or insight relative to a particular technical or operational discipline, system, or process" (Pace & Sheehan, 2002). Following the example of AHP research conducted by Garg et al., 2012, where SMEs were used to identify critical factors that affect customer satisfaction when using banks, a similar process was used to ensure the critical factors identified for pilot TIA are appropriate and allow for review of low scoring factors.

Six SMEs were utilized for participation in the AHP factor determination comparison. A study by Polit & Beck, 2006 recommended between three and ten SMEs should be used for establishing content validity of assessment instruments. SMEs consisted of three Army Aviation human factors researchers/engineers with a range of experience from 19-35 years and an average of 26 years in crewstation design and human factors testing, as well as, three Army Aviation research/test pilots with significant knowledge of Army Aviation requirements and capabilities. Pilots had a range of aviation career experience from 16-34 years with an average of 24 years. Pilot flight hours ranged from 1,800 – 5,000 and averaged 3,467 hours. Table 5 provides an overall summary of the collected demographics data for the SME's.

Table 5 SME demographics

SME Demographics	
Gender:	5 – Male; 1 – Female
Age (years):	Range: 40 – 62 Avg: 49
Time in Service/Employment (years):	Range: 16 – 35 Avg: 25
Military Occupation Specialty/Job:	3 – Human Factors Researchers/Engineers 3 - Research/Test Pilots
Primary A/C:	1 – AH-64 Apache 1 – UH-60 Blackhawk 1 – CH-47 Chinook / UH-60 Blackhawk
Total Flight Hours:	Range: 1,800 – 5,000 Avg: 3,467
Combat Flight Hours:	Range: 600 – 1,500 Avg: 950

Recruitment for SMEs was based on direct request through email to persons known to the researcher as having significant knowledge in the Army Aviation domain as either a human factors engineer or pilot. Participants were asked to participate as a SME in the initial down select of TIA critical factors and any follow on reviews of the iterated and final survey.

There is not a consistent standard in the literature related to requirements for SME participation with respect to face validity of surveys. However, by utilizing six SMEs closely associated with Army Aviation standards and future work, an initial face validity through the AHP method can help determine the validity of the proposed factors.

SMEs were interviewed about the context and usefulness of the proposed critical factors, solicited for any additional factors, and used to complete the AHP questionnaires, described below, virtually through Microsoft Teams during a scheduled time period.

Analytic Hierarchy Process Factor Determination

For development of the AS-TS, initial development of the hierarchical dimensions and attributes followed the traditional methods of developing questionnaire item pools through literature and expert review. Figure 14 provides a sample of the AHP survey used for the initial pool of factors for comparison. The full AHP survey is found in Appendix A.

Human vs. Automation Factors																	
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Human	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Automation

Human Factors - Confidence																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Purpose and Intent
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Transparency
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Technology Competence
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Faith
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Understandability
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Familiarity
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Predictability
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Demographics
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Personal Attachment
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk

Automation Factors - Reliability																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Feedback
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Safety
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Usability
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Effectiveness
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Integrity
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Accuracy
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Suitability

Figure 14 Sample AHP survey for TIA factors

Respondents reviewed the questionnaire virtually, and rated the factors using the importance scale, based on pair-wise direct comparison and the provided definition list (Appendix B). SMEs also reviewed the factors and provided feedback on factor characteristics that may be useful in the pilot survey.

Results

Once all the respondents completed the comparison matrices, average responses were used to calculate relative weights of the compared items. In this case, SMEs determined the level of importance of the TIA dimensional factors with respect to each other (i.e., Human vs. Automation), and then compared attribute factors within the overarching categories of Human Factors and Automation Factors. For example, a pilot SME may be more concerned with their confidence in systems compared to their personal attachment. In this example the SME would rate a higher rating towards the confidence factor. Once weights were assigned to the highest-level dimensional factors in a comparative manner, a relative importance table (matrix) was generated for Human versus Automation Factors by averaging the responses and defining the diagonal inverses (table 6).

Table 6 Relative importance matrix

Relative Importance	Human Factors	Automation Factors
Human Factors	1	0.33
Automation Factors	3	1

The priority vector matrix was generated by dividing each entry by the column sums for each attribute, and then averaging across each row to identify a priority vector for each dimensional factor with the result presented in table 7.

Table 7 Priority vectors

Priority Vectors	
Human Factors	0.25
Automation Factors	0.75

This process was repeated for each attribute factor in order to arrive at priority vectors for each attribute. These pairwise weights are the local weights for each factor. Table 8 shows the relative importance table for the Human Factors and table 9 shows the relative importance table for the Automation Factors. Both tables were generated through the pairwise comparison process of calculating and reporting average responses for each comparison and their associated inverse.

Table 8 Relative importance – Human factors

Relative Importance - Human Factors	Confidence	Purpose and Intent	Transparency	Technology Competence	Faith	Understandability	Familiarity	Predictability	Demographics	Personal Attachment	Risk
Confidence	1	2.17	0.46	0.25	0.23	0.32	2.33	0.35	0.67	3.5	1.33
Purpose and Intent	0.46	1	0.67	0.35	0.3	0.5	0.38	0.3	3.12	2.83	1.67
Transparency	2.17	1.5	1	0.67	0.67	0.4	1	0.35	2.83	2.17	2.17
Technology Competence	4	2.83	1.5	1	0.35	0.86	0.67	0.25	2.67	2	1.83
Faith	4.33	3.33	1.5	2.83	1	1	1.33	0.38	2.5	2.5	2.17
Understandability	3.17	2	2.5	1.17	1	1	2.17	0.46	4.17	4.67	3
Familiarity	0.43	2.67	1	1.5	0.75	0.46	1	0.33	2.5	1.67	2.5
Predictability	2.83	3.33	2.83	4	2.67	2.17	3	1	4.83	4.5	3.17
Demographics	1.5	0.32	0.35	0.38	0.4	0.24	0.4	0.21	1	1.83	1.12
Personal Attachment	0.29	0.35	0.46	0.5	0.4	0.21	0.6	0.22	0.55	1	1
Risk	0.75	0.6	0.46	0.55	0.46	0.33	0.4	0.32	0.86	1	1

Table 9 Relative importance – Automation factors

Relative Importance - Automation Factors	Reliability	Feedback	Safety	Usability	Effectiveness	Integrity	Accuracy	Suitability
Reliability	1	3.17	0.38	0.86	0.6	1.83	0.55	0.55
Feedback	0.32	1	0.4	0.35	0.35	0.5	0.26	1
Safety	2.67	2.5	1	2.67	1.83	2.5	1.5	1.5
Usability	1.17	2.83	0.38	1	0.86	2	0.86	0.75
Effectiveness	1.67	2.83	0.55	1.12	1	2.33	1.5	1.12
Integrity	0.55	2	0.4	0.5	0.43	1	0.46	0.5
Accuracy	1.83	3.83	0.67	1.17	0.67	2.17	1	1.5
Suitability	1.83	1	0.67	1.33	0.86	2	0.67	1

Priority vectors were again calculated by dividing each entry by the column sums for each attribute, and then averaging across each row to identify a priority vector for each attribute with the results presented in table 10 for the human factors and table 11 for the automation factors.

Table 10 Priority vectors – Human factors

Priority Vectors - Human Factors	
Confidence	0.07
Purpose and Intent	0.06
Transparency	0.08
Technology Competence	0.10
Faith	0.13
Understandability	0.14
Familiarity	0.08
Predictability	0.22
Demographics	0.04
Personal Attachment	0.04
Risk	0.04

Table 11 Priority vectors – Automation factors

Priority Vectors - Automation Factors	
Reliability	0.10
Feedback	0.06
Safety	0.22
Usability	0.12
Effectiveness	0.15
Integrity	0.07
Accuracy	0.15
Suitability	0.13

The consistency of the ratings was then examined to determine the consistency of the rating logic. For example, by the transitive property if A > B, and B > C, then A > C. However, it's possible that when using subjective ratings, raters could rate A < C and still rate A > B in the provided scenario. The consistency ratio helps to determine whether the rater judgements are consistent enough to be reliable, which could also help identify significant instances that may infer that the factors need to be restructured for clearer meaning (Saaty, 1980).

Mathematically, the following notation (figure 15 and equations 1-3) expresses the AHP method for identifying weights and consistency. Figure 15 shows matrix Z which represents the pairwise comparisons of X1, X2, X3, . . ., Xn elements under a node with numerical weights w1, w2, w3. . .wn, and aij = wi/wj (i, j = 1, 2..., n) ultimately representing the quantified comparative importance matrix elements (Saaty, 1994; Garg et al., 2012).

$$Z = \begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_n \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \dots & X_n \end{array} \\ \begin{bmatrix} w_1/w_1 & w_1/w_2 & \dots & w_1/w_n \\ w_2/w_1 & w_2/w_2 & \dots & w_2/w_n \\ \vdots & \vdots & \vdots & \vdots \\ w_n/w_1 & w_n/w_2 & \dots & w_n/w_n \end{bmatrix} \end{array}$$

$$Z = \begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_n \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \dots & X_n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \end{array}$$

Figure 15 AHP weighted matrix (Z)

Saaty (1994); Garg et al. (2012)

Once the matrices are developed, the eigenvectors and maximum eigenvalues can be identified. The maximum eigenvalues can be calculated using equation 1. The eigenvectors can be computed using equation 2, where W is the eigenvector and λ_{max} is the largest eigenvalue of the Z matrix (Garg et al., 2012).

$$\lambda_{max} = \sum_{j=1}^{n} a_{ij} \frac{W_j}{W_i} \quad (1)$$

$$Z \cdot W = \lambda_{max} \cdot W \tag{2}$$

Table 12 provides the eigenvalues and λ_{max} for the Human Factors and table 13 provides the eigenvalue attributes for the Automation Factors.

Table 12 Eigenvalues – Human factors

Eigenvalues - Human Factors	
Confidence	1.44
Purpose and Intent	1.20
Transparency	1.03
Technology Competence	1.27
Faith	1.08
Understandability	1.05
Familiarity	1.12
Predictability	0.91
Demographics	1.08
Personal Attachment	0.97
Risk	0.89
λ_{max}	12.05

Table 13 Eigenvalues – Automation factors

Eigenvalues - Automation Factors	
Reliability	0.10
Feedback	0.06
Safety	0.22
Usability	0.12
Effectiveness	0.15
Integrity	0.07
Accuracy	0.15
Suitability	0.13
λ_{max}	8.33

The pairwise comparison inconsistency was then measured by a Consistency Index (CI) and coherence was measured by the Consistency Ratio (CR) which can be determined using equation 3, where the Random Index (RI) is predetermined based on n-elements (table14) (Saaty, 1994; Garg et al., 2012). The maximum values of CI and CR are 0.1. Values higher than 0.1 suggest that the pairwise comparison is inconsistent and should be discarded.

$$CI = \frac{\lambda_{max} - n}{n - 1} \quad CR = \frac{CI}{RI} \tag{3}$$

Table 14 AHP Random Index table

n	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
RI	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

Saaty (1994); Garg et al. (2012)

Table 15 provide the CI and CR for the Human and Automation Factors. In both cases, the results met the thresholds of being less than or equal to the CI and CR values of 0.10, required for consistency. The Human Factors resulted in a CI of 0.10 and a CR of 0.07, while the Automation Factors CI and CR were calculated to be 0.05 and 0.03 respectively.

Table 15 Consistency Index and Consistency Ratio table

	Consistency Index	**Consistency Ratio**
Human Factors	0.10	0.07
Automation Factors	0.05	0.03

Since the CI and CR metrics fell within the appropriate AHP parameters, further analysis was conducted by sorting and ranking each factor group by their priority vectors to identify the "most important" factors within each attribute list. Table 16 and 17 show the ranked values of the Human and Automation factors respectively.

Table 16 Human factors - Ranked

Human Factors - Ranked	
Predictability	0.22
Understandability	0.14
Faith	0.13
Technology Competence	0.10
Familiarity	0.08
Transparency	0.08
Confidence	0.07
Purpose and Intent	0.06
Risk	0.04
Demographics	0.04
Personal Attachment	0.04

Table 17 Automation factors - Ranked

Automation Factors - Ranked	
Safety	0.22
Accuracy	0.15
Effectiveness	0.15
Suitability	0.13
Usability	0.12
Reliability	0.10
Integrity	0.07
Feedback	0.06

Finally, global weights were calculated by multiplying the dimensional factor priority vector weights by each attribute local weight. The resulting product provides a global weight that was used to rank the factors by perceived importance compared across all dimensions (table 18). At this point, the AHP process was terminated as the weighting of factors was established for identifying the most important TIA attributes of an automated aviation system.

Table 18 Combined factors - Ranked

Combined Factors - Ranked	
Safety	0.16
Accuracy	0.12
Effectiveness	0.11
Suitability	0.09
Usability	0.09
Reliability	0.08
Predictability	0.05
Integrity	0.05
Feedback	0.04
Understandability	0.03
Faith	0.03
Technology Competence	0.02
Familiarity	0.02
Transparency	0.02
Confidence	0.02
Purpose and Intent	0.01
Risk	0.01
Demographics	0.01
Personal Attachment	0.01

Lower scoring factors were examined for level of perceived importance and the bottom four: Purpose and Intent, Risk, Demographics, and Personal Attachment, received a ranking index of 0.01. While there is not a standard established for removal of items based on the AHP, a subjective determination can be made by examining the scores, context, and SME comments.

Based on the collected data, Purpose and Intent, Risk, Demographics, and Personal Attachment were not included on the notional AS-TS survey for pre-testing as depicted in table 19.

Table 19 Notional TIA survey – AHP results

Trust Questions	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
I am confident in my ability to interact with the system.							
The system provides transparent information.							
The system competently performs the intended task.							
I have faith that the system will perform the intended task.							
I understand what the system is doing.							
I am familiar with the system operation.							
The system operates in a predictable manner.							
The system maintains reliable (consistent) performance.							
The system provides appropriate feedback on current and future actions.							
The system does not present an unacceptable hazardous condition.							
The system effectively accomplishes its tasks.							
The system is easy to use.							
The system operates with integrity to complete the tasks.							
The system is accurate when completing tasks.							
The system is suitable for carrying out the task.							

Additionally, SMEs were asked if there were any factors that were potentially missing that they thought may affect TIA for pilots. No additional significant factors were identified, but "Personal Security" was a concept provided for consideration. As this concept mostly falls under the "Risk" factor, it was decided to not consider "Personal Security" as an additional factor.

Discussion

The results of the study provided sufficient data to answer the research question and evaluate the two hypotheses concerning the effectiveness of the AHP and the relevance of specific factors for inclusion consideration on the notional AS-TS.

What factors influence TIA for Army pilots using Army Aviation systems?

a. Hypothesis 1: The Analytic Hierarchy Process (AHP) decision-making tool will be effective in identifying critical TIA factors and establishing face and content validity of the proposed TIA factors for inclusion on an initial aviation systems trust survey.

b. Hypothesis 2: Many of the key factors identified within the initial literature review will likely be considered relevant factors for Army Aviation pilots, with the exception of personal attachment and demographics.

Hypothesis 1 was accepted by demonstrating that the AHP decision-making tool was useful in identifying SME perceptions of the TIA factors that were presented. SMEs were able to understand the AHP tool and reported that it provided a thorough review of TIA factors through comparative pairwise assessment. The AHP worked similarly to a one-on-one focused interview with the SMEs by providing an outlet for discussion and quantitative data for the qualitative metrics. The use of AHP for this research provided initial face and content validity through consistent AHP results and SME discussion related to the thoroughness of the identified factors with respect to covering the TIA conceptual domain.

Hypothesis 2 was also accepted by verification that “Demographics” and “Personal Attachment” scored the lowest on the associated AHP outcomes. While these scored low, they should still not be discounted as important contributors to the overall assessment of TIA within an automated system. Demographic data should still be collected and examined for statistical relevance to assist in determining whether trends can be correlated using demographic data and the outcomes of the associated AS-TS. While demographics may not be necessary for direct questioning on the AS-TS, it will be recommended to collect and evaluate demographic data directly through other means (e.g., survey), independent of the perceived influence by individual pilots.

Similarly, “Personal Attachment” could cross into the “Demographic” lane when comparing flight hours on a specific system. While SMEs agreed that “Personal Attachment” could play a role in some cases, the most common conversational comment was that Personal Attachment preference was likely to be more prevalent for higher flight hour or older pilots that were particularly proficient on a specific legacy or older system. With this context in mind, a similar handling of Personal Attachment can be associated with collected demographic data. When reviewing collected AS-TS data, examining the pilot demographics for flight hours and time spent on the legacy or previous systems should be considered to make inferences on perceived personal attachment, especially if discrepancies occur between system performance and pilot perception of the system.

Additionally, two other factors, “Purpose and Intent” and “Risk” scored low on the overall ratings. SMEs commented that “Purpose and Intent” was closely related to “Understanding” of the system, as well as, the other associated factors, that when considered collectively would provide a representative mental model of how the system works and its

intended purpose. For example, “Understandability”, “Familiarity”, “Faith”, and “Technology Competence” all relate to general knowledge of the “Purpose and Intent” of the system. Removing “Purpose and Intent” from the notional AS-TS would potentially reduce redundancy without sacrificing sensitivity of the survey.

“Risk” was somewhat unique in that it was considered context dependent for situational assessment in a “risky” environment. During flight testing and simulation, mental workload ratings are often collected to identify tasks or situations that pilots perceive as high workload events. Highly complex situations are often correlated with high workload ratings (Paxion, Galy, & Berthelon, 2014). In the case of system performance in a risky environment, SMEs reported that TIA for the system could be affected, especially when considering use of the system, but that the “Risk” term is more contextual for specific mission sets and technology designs. Since most TIA factors are concerned with system or pilot perception characteristics, the inclusion of “Risk” is somewhat ambiguous for a lower risk situation and the AS-TS would be asking the pilot to both establish a risk criterion for the mission and judge the automation performance. To account for “Risk” while using the AS-TS it will be important to consider the context of the mission-automated system relationship and ensure appropriate reliance on the automation is accomplished by observing mission events during higher risk scenarios. A more objective approach of user reliance (e.g., amount of use, errors made) would likely address the “Risk” factor of system use, better than subjective ratings from the pilot performing the tasks.

The removal of the four factors: Demographics, Personal Attachment, Purpose and Intent, and Risk from the notional AS-TS leaves 15 total factors for survey evaluation, seven Human Factors and eight Automation Factors. A follow-on study was conducted to perform construct validity testing on the proposed survey with a demographic of Army Aviation pilots.

Conclusion

In order to identify the key factors that influence TIA for Army Aviation systems, a comprehensive literature review was conducted to establish a historical background of TIA research, identify key concepts in human and automation trust, and to develop an initial list of factors that could potentially impact pilot TIA. By identifying prominent factors, an initial pool of items was established for development of the AS-TS.

The factor identification and associated definitions were evaluated by SMEs to conduct survey face validity testing in order to establish the initial AS-TS survey for the Army Aviation demographic.

Six SMEs (3 Human Factors Researchers/Engineers and 3 U.S. Army Pilots) were recruited to evaluate the identified TIA factors through pairwise comparison using the AHP decision-making methodology. Additional SME comments were collected during the data collection about their perception of the TIA factors and any additional TIA factors that should be included. SMEs agreed that the factor list was inclusive, and no significant comments were captured related to additional factor requirements.

Once the pairwise comparisons were completed, calculations were made to assign priority and ranking to the TIA factors. The data collected successfully addressed the research question and both research hypotheses.

The AHP was used effectively to identify critical TIA factors for Army pilots and establish initial face validity of a TIA survey. Additionally, four TIA factors (i.e., Purpose and Intent, Risk, Demographics, and Personal Attachment) were removed from consideration of the notional AS-TS due to low rankings and SME comments related to the factors.

Formal construct validation of the AS-TS was conducted as a follow-on study with military pilots using representative scenarios of automation performance. Throughout the validation process the survey was refined and tested for validity and reliability, with the intent of developing a robust survey tool to analyze the trust relationship that pilots establish with automated systems in current and future aircraft. Identifying deficiencies or lower scoring items in the trust relationship can help human factors engineers and program managers focus on improving and calibrating the trust relationships to optimize the human-automation team.

Additionally, the SME review and feedback of these factors can help to provide critical factors of TIA for consideration during requirements development and automation design of future systems. The AHP method can also be generalized across perception-based survey development as a quantitative method for determining face validity of new questionnaires.

CHAPTER IV

SURVEY CONSTRUCT VALIDITY AND RELIABILITY

Introduction

TIA is a key concept in influencing user acceptance of new automated technology (Korber, Baseler, & Bengler, 2018). Both modern and future aircraft contain autonomous systems that play a major role in their use and development. Pilots must trust that the automation is performing to standard to ensure appropriate performance of the aircraft and satisfactory completion of pilot tasks. The U.S. Army is promoting efforts to investigate pilot TIA, as new technologies begin to emerge, such as the Army FVL program. The FVL program is one of the Army's three major modernization priorities and includes initiatives for both a FARA and a FLRAA (Mayfield, 2021). Optimization of the pilot-FVL interactions is a major priority for FVL crewstation development. The FVL systems will be required to provide multiple levels of supervised autonomy within the aircraft (e.g., autonomous takeoffs/landings, cueing, and adaptive interventions). Appropriate levels of TIA for the pilots will be extremely important for FVL platforms due to a significant focus on automated processes and automated assistance in high-speed and DVE rotorcraft operations (Freedberg, 2020).

Problem Statement

There is currently no standard methodology that is in use for the Army to assess TIA for pilots as a holistic measurement that identifies trust deficiencies and the relationship of trust to user reliance with follow-up actions. Several surveys are currently in use, but only used as a

general indicator of system trust with no recommendations to improve the user-automation trust relationship. To address the lack of a standardized survey tool for TIA measurement in Army Aviation, a comprehensive literature review was conducted to identify key factors that may influence TIA in aviation systems (Chapter II). Three studies were defined to identify the TIA factors that influence pilot trust in automation systems and to validate a data collection survey that measures the influence of the factors when pilots use automated systems. The first study established face and content validity of the proposed survey. The identified factors were used as a foundation to develop an initial pool of factors for review by SMEs through both interview and use of the AHP to refine the pool of factors to establish face validity for a TIA survey tool specialized for use in Army Aviation system assessments. Building on the previous literature review of identified factors, the AHP method was used to solicit pilot input to answer the research question "What factors influence TIA for Army pilots using Army Aviation systems?" A notional TIA survey was developed and provided in table 19.

The SME comments and AHP method established face and content validity for the development of the survey (Chapter III). However, further validation and reliability analysis were required to answer the research question: "Can a survey instrument developed from identified TIA factors reliably measure pilot TIA perception of Army Aviation systems?" The second study in this research focused on survey validation and determining appropriate reliability.

The initial hypothesis for this study is that a survey pre-test, subject to factor analysis and reliability testing, will successfully determine construct validity and reliability for the proposed TIA survey. A secondary hypothesis is that the survey will contain two overarching factors (human and automation) verified through factor analysis.

Literature Review

Once the initial factors were identified through SME review and AHP decision-making methods, the survey tool was ready for further validation through statistical processes. Pilot testing is typically conducted to gather initial data for further survey analysis, using the initial scale developed from SME review and early validity testing (Salminen et al., 2020). Following pilot testing, construct validity and scale reliability are examined to complete the survey validation. Consideration for participant sample sizes is also important, especially given the small numbers, limited access, and availability of the Army Aviation pilot population.

Survey Pre-Test

Survey pre-tests are considered a "critical examination" of the tool to ensure it is both valid and reliable (Converse & Presser, 1986). Pre-testing is typically conducted to gather initial data for further survey analysis, using the initial scale developed from SME review and early validity testing (Salminen et al., 2020). Recruitment for respondent-driven survey pre-tests are often conducted on a small subsample of the sample population with emphasis on ensuring the demographic and cultural profile matches the intended population (Ferketich, Phillips, & Verran, 1993). By utilizing a respondent-driven and scenario-based approach, participants can answer survey questions based on various outcomes of the expected situational use cases for the tool. Participant ratings are then used to identify problem areas, reduce measurement error and participant burden, ensure correct interpretation of questions, and minimize any order of question influence on participant ratings (Gillespie, Ruel, & Wagner, 2015).

Construct Validity

After the pre-test data collection for the survey, further validity testing can be conducted. Construct validity is a method of ensuring that questionnaires actually test the theory they are measuring (Ginty, 2013). In perception-based tool development, construct validity is a very common step to ensure the overall validity of the survey tool (Bargas-Avila & Bruhlmann, 2016; Mason et al., 2021; Simon, 2020). While there is no single metric that defines construct validity, several statistical tests help to determine the dimensionality of the survey and to ensure the survey items are reliable (Salminen et al., 2020; Westen & Rosenthal, 2003).

In order to validate proposed multi-dimensional grouping for survey development, correlation analysis (e.g., Pearson and/or Spearman Rho) can be used to generate an initial correlation matrix and a scree plot can be produced to validate factor grouping or adjust factor grouping based on the correlation results (Murray, 2013; Frey, 2018). Additionally, a factor analysis is often used to determine the dimensionality of the survey and generate a final factor loading and correlation matrix (Yong & Pearce, 2018; Salminen et al., 2020; Schaefer, 2013; Simon, 2020). The results from the correlation analysis, a parallel analysis, and scree plot are then used to initially identify the appropriate number of factors for use during the analysis.

The validity analysis typically utilizes EFA or some variation (e.g., Confirmatory Factor Analysis) to identify scale dimensionality (Bargas-Avila & Bruhlmann, 2016; Mason et al., 2021; Salminen et al., 2020; Simon, 2020; Spain, Bustamante, & Bliss, 2008). EFA is used when the intent of the research is to "create a measurement instrument that reflects a meaningful underlying latent dimension(s) or construct(s) represented in observed variables" (Chyung et al., 2017). EFA allows for the identification and labeling of groups of variables that have high

correlations with specific factors, by comparing the relationships of the latent dimensions to scale items (figure 16) (Chyung et al., 2017).

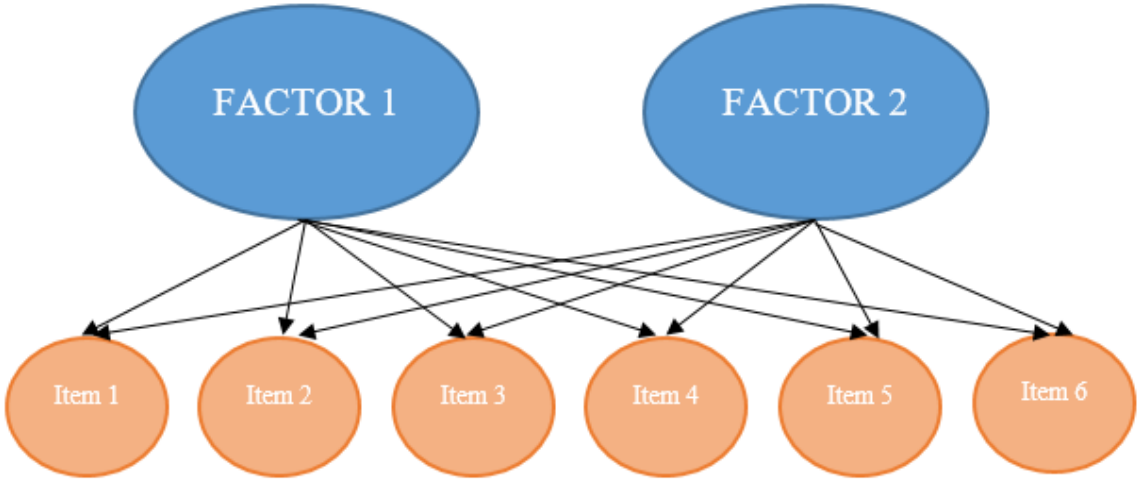

Figure 16 Sample EFA diagram

Ultimately, the factor analysis produces a correlation matrix and provides evidence for the number of related latent factors within the scale items. Factors that do not cleanly load can be removed from the scale or further evaluated to determine necessity (Institute for Defense Analyses, 2018). While the initial hypothesis for the AS-TS is for a two-factor scale, the EFA process will validate whether the survey responses accurately reflect the two-factor model.

Scale Reliability

Once the survey is validated, the reliability of each scale can be measured using Cronbach's alpha to ensure internal consistency (Bargas-Avila & Bruhlmann, 2016; Hermann, Bager-Elsborg, & Parpala, 2017; Salminen et al., 2020; Simon, 2020). Where internal consistency can be defined as the "extent to which all items in a test measure the same concept or construct" (Takavol & Dennick, 2011) and "how well the different items complement each other

in their measurement of different aspects of the same variable or quality" (Litwin, 2003). Cronbach's alpha can be used to help further justify item groupings by identifying close relationships, reduce unnecessary items, and remove items that negatively impact internal scale consistency (Warmbrod, 2014). For example, the researchers in Wojton et al. (2020) used a concurrent validity test of correlation to identify whether a two-factor scale (understanding, performance) positively correlated with the statement "I trust the system". Cronbach's alpha can be used for each subscale (i.e., Human and Automation Factors) to ensure internal consistency. Cronbach's alpha is calculated using equation (4) and comparing the outcome to table 20 (*Cronbach's Alpha*, 2021). Higher alpha numbers indicate better internal consistency and ensure that the scale is consistently measuring the intended construct.

$$\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}} \tag{4}$$

Where: N = number of items

$\bar{c}$ = average covariance between item pairs

$\bar{v}$ = average variance

Table 20 Cronbach's Alpha consistency scale

Cronbach's Alpha	Internal Consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Acceptable/Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

Cronbach's Alpha (2021); *Taber* (2018)

Participant Sample

A significant challenge in this research was collecting an adequate sample size of responses to measure scale reliability. Kline (1986) suggested that reliability analysis should not be performed on samples sizes smaller than 300 participants. However, later research has found that by adhering to a set of specific guidelines for handling smaller sample sizes, inferences on scale reliability can still be valid. Yurdugul (2008) conducted a study to determine the minimum sample size for Cronbach's alpha coefficient. This study is also cited by Samuels (2017) as a method to handle smaller sample sizes when considering scale reliability and analysis. The guidelines identified by these studies suggested a sample size of $N = 30$ should be sufficient to conduct reliability testing. The principles identified in these studies were primarily determined based on Principal Component Analysis (PCA) factor loading and eigenvalue cut-offs.

When considering EFA, a similar study by de Winter, Dodou, & Wieringa (2009) was conducted using a variety of simulated combinations to determine whether smaller sample sizes would be valid for conducting EFA under particular cases. The study found that when factors loaded high ($\lambda > 0.6$), factor numbers were low ($f = 2$), and the number of variables considered high ($p = 24$), a reasonable N for participants to successfully complete an EFA was suggested as $N = 34$. In general, the study found that datasets with a high λ, low f, and high p, allow for significantly less than 50 participants as a sample N for completion of a meaningful EFA. The current proposed AS-TS survey has an expectation of low factor complexity ($f = 2$) and approximately $p = 15$ variables. This proposed design, with appropriate factor loading, should be adequate for smaller sample size data collection while retaining statistical integrity for factor analysis and scale reliability. By adhering to the proposed participant sample guidelines, a

reasonable determination on the reliability of the scale can be made, even with a smaller sample size.

Methods

Pre-Testing

A pre-test of the notional survey was required to begin the validation and reliability analysis. By utilizing a scenario-based approach, participants can answer survey questions based on decisions made by the automation. Figure 17 shows a virtual scenario (i.e., simulation) of an aircraft with an automated guidance system that can avoid obstacles. Participants received a briefing on general trust concepts and the purpose of the automated system and then experienced four scenarios. One where the automation performs very well, a second where the automation performs poorly - completely opposite of the system intention, a third where the pilot is unaware of the system parameters and the correct automation decision is made, and a fourth where the pilot is fully aware of the system parameters, but the wrong automation decision is made.

Each scenario began with similar initial conditions. The scenarios were pre-recorded and presented over MS Teams. No participant input was required. The aircraft began in low level flight in flat terrain with towers as hazards in a DVE (unaided night, rain). The aircraft was equipped with an obstacle detection and avoidance system that *should* assist the pilot in avoiding towers.

Three pieces of significant information were provided to the pilots during the simulation video. If a tower hazard was detected, a red rectangle was overlayed over the tower. A ground track indicating the aircraft route was provided using purple chevrons. The ground track would show whether the aircraft route updated based on hazard detection. Finally, a feedback window

was provided at the top of the screen indicating important information on actions performed by the automation.

Figure 17 Virtual scenario example

Scenario Descriptions

Scenario 1 Description: You are a pilot in an aircraft with automated flight controls and obstacle detection/avoidance. Your automation has been proven to be highly reliable in obstacle detection and you have extensive training on the use of the system.

Automation Interaction: Obstacle detection is activated and soon after avoids an oncoming obscured obstacle by identifying it with an overlay and providing information on the proposed turn for avoidance and final clearance.

Scenario 2 Description: You are a pilot in an aircraft with automated flight controls and obstacle detection/avoidance. You have no additional information on the system, other than the use case, and are unclear on any detailed parameters for operation.

Automation Interaction: Pilot begins flight, automation identifies an obstacle, but not the obstacle in the flight path and the aircraft impacts the object.

Scenario 3 Description: You are a pilot in an aircraft with automated flight controls and obstacle detection/avoidance. Your automation has been proven to be highly effective in obstacle detection and avoidance, but you have little training on the use (e.g., system parameters) of the system and no additional information on how the system works.

Automation Interaction: Pilot begins flight, automation requests setup/activation. Pilot is unaware of setup procedures. Automation identifies an obstacle near impact and makes an evasive maneuver to save the aircraft, with feedback provided to the pilot.

Scenario 4 Description: You are a pilot in an aircraft with automated flight controls and obstacle detection/avoidance. You are trained on the automation and are well aware of its history of poor performance.

Automation Interaction: Pilot begins flight, pilot activates the automation. Automation identifies multiple obstacles with only one in the flight path. Automation fails and requests the pilot to manually control the aircraft.

The scenarios for the pre-test allowed for a range of expected TIA interactions that pilots may experience. By utilizing a scenario-based approach, the AS-TS scale validation could be determined over cases of automation successes and failures.

Pilot Participation

The intent of this research was to collect as much data as possible within the Army Aviation pilot community with an expectation of a minimum of 32 participants. The minimum number is derived from the Latin Square design (table 21) where 32 participants would create an equal number of participants completing each condition (i.e., four conditions completed eight times). It was unlikely that a larger sample (e.g., 300 or more) of pilots would be able to complete the survey due to time constraints and limited accessibility. Notionally, Army pilots have similar training and experiences with very similar types of aircraft technologies. An effort was made to balance the Army pilots based on primary aircraft (e.g., AH-64 Apache, UH-60 Blackhawk), but data collection was subject to pilot availability for participation. The collection of 300+ responses would cover more diverse experience, however due to the limited nature of the proposed survey and small population of Army pilots, it is likely unnecessary for a large sample size to establish survey validity.

Counterbalancing followed a Standard Latin Square for a four-factor study (Mason, Gunst, & Hess, 1989). Table 21 shows the experimental condition order in which each condition appears exactly once in each row and column. Participants were assigned at random to complete each condition by row order (A-D), with an equal number of participants for each row.

Table 21 Counterbalanced study design

Standard Latin Square Design				
Condition Order: A	1	2	3	4
Condition Order: B	2	3	4	1
Condition Order: C	3	4	1	2
Condition Order: D	4	1	2	3

Pilot Recruitment

Recruitment for pilots was based on direct request through email to persons known to the researcher as having qualified in an Army helicopter (e.g., AH-64 Apache, UH-60 Blackhawk, CH-47 Chinook). Additional recruitment opportunities were identified through participant recommendation on additional pilots that were available to participate in the pre-test.

Both pilots and Army acquisition professionals with pilot experience and qualifications were utilized for participation. Best efforts were made to recruit a broad sample of pilots based on availability. All participation was conducted virtually over MS Teams.

Army Aviation organizations for recruitment included: Army Capabilities Manager (ACM), Directorate of Evaluation and Standards (DES), U.S. Army Aeromedical Research Laboratory (USAARL), Redstone Test Center (RTC), Directorate of Simulation (DOS), Special Operations Command (SOCOM), Aviation Missile Command (AMCOM), and other Army Aviation Units.

Pilot Test Procedures

After initial introductions, participant demographics were collected, and each pilot was assigned a unique Personal Identification Number (PIN) for data collection (table 22). Gender, age, time in service, Military Occupational Specialty (MOS), primary aircraft, total flight hours, and combat flight hours respectively provide potential information that could be used to investigate biases (i.e., age-related, aircraft related, flight hours related) and establish overall experience with aviation systems. Primary aircraft for Army helicopter pilots typically include the categories of Attack/Recon (AH-64 Apache) or Cargo/Lift (UH-60 Blackhawk and CH-47 Chinook). Each aircraft has unique autonomous systems that could potentially affect the pilot perception of automated systems within the aircraft. Additionally, high flight hours are often

correlated with time in service or age but may also indicate more experience with aviation related automated systems. Further analysis on demographic correlations to pilot ratings could be investigated later to identify potential pilot rating influences based on the demographic categories.

Table 22 Demographics survey

Demographics PIN:_____	
Gender:	
Age:	
Time in Service:	
MOS:	
Primary A/C:	
Total Flight Hours:	
Combat Flight Hours:	

Participants received a concept briefing of TIA and the associated scenario description with instructions to rate each event by indicating their perception of trust based on the questionnaire items. Participants watched the four pre-recorded videos of the aircraft maneuvers in a counter-balanced order.

After each video, participants used the survey (table 23) to answer the TIA questions about each scenario. A definition list was provided for each item to ensure participants understood the item under review (Appendix B).

Table 23 Notional TIA survey from AHP

Trust Questions	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
I am confident in my ability to interact with the system.							
I understand the purpose and intent of the system.							
The system provides transparent information.							
The system competently performs the intended task.							
I have faith that the system will perform the intended task.							
I understand what the system is doing.							
I am familiar with the system operation.							
The system operates in a predictable manner.							
I have a personal attachment to the system.							
The system takes risks to complete the task.							
The system maintains reliable (consistent) performance.							
The system provides appropriate feedback on current and future actions.							
The system does not present an unacceptable hazardous condition.							
The system effectively accomplishes its tasks.							
The system is easy to use.							
The system operates with integrity to complete the tasks.							
The system is accurate when completing tasks.							
The system is suitable for carrying out the task.							

Data Analysis

The following paragraphs describe analyses that were conducted using a statistical package software (i.e., SPSS) for scale validation.

In order to validate the proposed multi-dimensional grouping, an initial correlation matrix and a scree plot were produced to validate factor grouping and adjust factor grouping based on the correlation results (Murray, 2013; Frey, 2018). Additionally, a factor analysis was used to determine the dimensionality of the survey and generate a final factor loading and rotated correlation matrix (Yong & Pearce, 2018; Salminen et al., 2020; Schaefer, 2013; Simon, 2020).

The results from the correlation analysis, parallel analysis, and scree plot were used to initially identify the appropriate number of factors for use during the analysis. The analysis utilized EFA to identify scale dimensionality (Bargas-Avila & Bruhlmann, 2016; Lim and Jahng, 2019; Mason et al., 2021; Salminen et al., 2020; Simon, 2020; Spain, Bustamante, & Bliss, 2008). For scenarios 1 and 2, descriptive statistics were generated to show correlation of the ratings based on the expectation of the ratings fully correlating with a single (positive or negative) trust construct (Bolarinwa, 2015). After reviewing the initial descriptive results, a PCA was conducted on scenario 1 and further EFA analysis was conducted on scenario 2. An EFA was required for scenario 3 and 4 to validate that the scale multi-dimensionality measures appropriately for the proposed scenarios (Zhuo et al., 2021). The hypotheses for the scenarios are below.

a. Scenarios 1 and 2 will likely result in a single factor of trust measurement due to the scenario descriptions providing information related to increased or decreased trust to both hypothesized subscales of *human* and *automation* factors.

b. Scenarios 3 and 4 will likely identify the separate or multi-dimensional (*human* and *automation*) aspect of the proposed survey.

The factor analysis produced correlation matrices and provided evidence for the number of related latent factors within the scale items. Factors that did not cleanly load were removed from the scale or further evaluated to determine necessity (Institute for Defense Analyses, 2018). While the initial hypothesis was for a two-factor scale, the EFA process validated whether the survey responses accurately reflected the two-factor model.

Scale Reliability

The reliability of each scale, for each scenario, was measured using Cronbach's alpha to ensure internal consistency (Bargas-Avila & Bruhlmann, 2016; Hermann, Bager-Elsborg, & Parpala, 2017; Salminen et al., 2020; Simon, 2020). The analysis was conducted using statistical package software (i.e., SPSS) using equation (5) and comparing the outcome to table 24 (*Cronbach's Alpha*, 2021).

$$\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N - 1) \cdot \bar{c}} \tag{5}$$

Where: N = number of items

$\bar{c}$ = average covariance between item pairs

$\bar{v}$ = average variance

Table 24 Cronbach's Alpha consistency scale

Cronbach's Alpha	Internal Consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Acceptable/Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

Cronbach's Alpha (2021); *Taber* (2018)

Results - Scenarios

Results of pilot demographic data and ratings during the pretest were analyzed using statistical software (e.g., MS Excel, SPSS). Charts and tables that support the factor analysis (e.g., scree plot) are included to show data calculation results for correlation and factor analysis for each scenario. Scale/sub-scale reliability for each scenario was determined and appropriate calculations are provided.

Pilot Demographics

Pilot demographic data were collected prior to participation in the scenario evaluations. Table 25 provides the results of the demographics collected for pilots.

Table 25 Demographic data

Demographics PIN:_____	
Gender:	31 Male; 1 Female
Age (years):	Range: 30-67 Average: 44.5
Time in Service (years):	Range: 8 – 48 Average: 21.6
MOS:	Warrant Officers: 19 Aviation Officers: 6 Flight Test Engineer: 1
Primary A/C:	UH-60: 18 AH-64: 8 CH-47: 3 Other: 3
Total Flight Hours:	Range: 535 – 14,000 Average: 3,237
Combat Flight Hours:	Range: 0 – 3,000 Average: 780

Thirty-two (32) pilots were recruited for the study. Pilots ranged in experience form 535 – 14,000 total flight hours with an average service time of 21 years. UH-60 pilots were the largest pool of participants, followed by AH-64 pilots. The breakdown of participants is also somewhat representative of the Army fleet of aircraft. The Army maintains approximately 2,135 UH-60 aircraft, 800 AH-64 aircraft, and 400 CH-47 aircraft (Chadwick, 2022; Reim, 2021; Airforce Technology, 2021). UH-60 aircraft numbers are nearly triple the other platforms in Army inventory. Participants were primarily Warrant Officers or Aviation Officers and included a variety of jobs such as: Instructor Pilot, MEDEVAC Pilot, Aviation Requirements Development, Aviation Operations Officer, and Aviation Simulation Trainer. One participant was a Flight Test Engineer who primarily performed co-pilot duties during flight testing but had significant experience using advanced automated systems in Army aircraft cockpits.

Summary Results

Data collected during the study were initially analyzed for descriptive statistics (i.e., mean, standard deviation) and provided in bar-charts to show the general distribution of answers across each scenario. Numerical data corresponds to a 7-point Likert scale where 1 = Strongly Disagree; 2 = Disagree; 3 = Somewhat Disagree; 4 = Neither Agree nor Disagree; 5 = Somewhat Agree; 6 = Agree; 7 = Strongly Agree.

Scenario 1

Scenario 1 was considered an "anchor scenario" in which most of the ratings were expected to be positive, corresponding to the Somewhat Agree – Strongly Agree ratings. The descriptive statistics of scenario 1 align with the proposed hypothesis 3 for the second study.

a. Hypothesis 3: Scenario-based validity testing will result in a single factor trust construct when analyzing anchor scenarios of positive and negative experience automation.

Further statistical analysis is provided in the Results – Factor Analysis section to verify a single factor loading. Table 26 and 27 provide the descriptive statistics and figure 18 provides the distribution of the ratings.

Table 26 Scenario 1 Human Factors - results

Human Factors	Confidence	Transparency	Competence	Faith	Understanding	Familiarity	Predictability
Average	6.41	6.16	6.41	6.09	6.06	6.53	6.16
Std. Deviation	0.50	0.77	0.61	0.86	0.88	0.80	0.81

Table 27 Scenario 1 Automation Factors - results

Automation Factors	Reliability	Feedback	Hazardous	Effectiveness	Ease of Use	Integrity	Accuracy	Suitability
Average	6.28	5.81	5.75	6.41	6.53	6.28	6.41	6.34
Std. Deviation	0.89	1.20	0.92	0.67	0.51	0.77	0.50	0.65

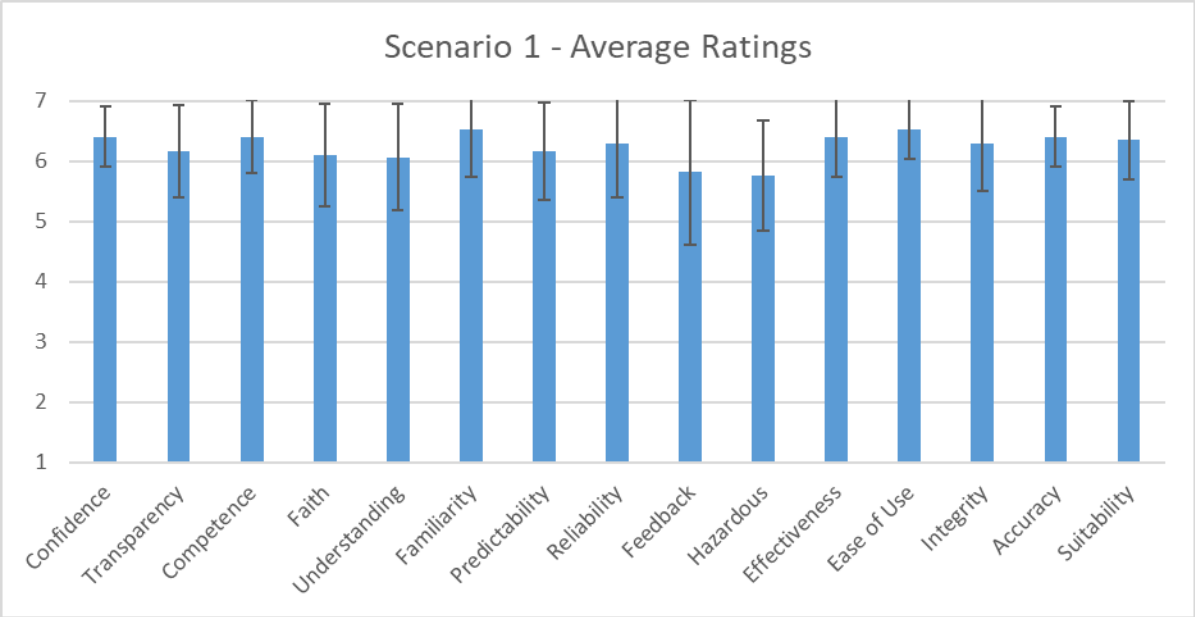

Figure 18 Scenario 1 ratings distribution

The data collected for scenario 1 supports the acceptance of the hypothesis that a positive automation experience will result in positive TIA ratings. The strongest ratings included pilot perception of their confidence and familiarity of the system, as well as the system technical competence and accuracy. The lower performing ratings included feedback and safety (hazardous conditions). In these cases, pilots commented that while feedback was provided in the form of the status window and flight path, additional feedback on distance to obstacle and forecasted decisions (e.g., turn direction) could have improved the experience. With respect to

system safety, pilots commented that the simulation seemed to move quickly to avoid the object and they were unsure if the turn was safely executed, or they would prefer a smoother avoidance maneuver. Overall, pilots rated scenario 1 positively and considered the automated task performance to be successful.

Scenario 2

Scenario 2 was also considered an anchor scenario in which most of the ratings were expected to be negative, corresponding to the Somewhat Disagree – Strongly Disagree ratings. The descriptive statistics of scenario 2 align with the proposed hypothesis 3 for the second study.

a. Hypothesis 3: Scenario-based validity testing will result in a single factor trust construct when analyzing anchor scenarios of positive and negative experience automation.

Further analysis is provided in the Results – Factor Analysis section to factor loading. Table 28 and 29 provide the descriptive statistics and figure 19 provides the distribution of the ratings.

Table 28 Scenario 2 Human Factors - results

Human Factors	Confidence	Transparency	Competence	Faith	Understanding	Familiarity	Predictability
Average	3.31	3.00	1.25	1.22	3.56	2.91	1.94
Std. Deviation	2.33	1.95	0.57	0.49	2.08	1.86	1.37

Table 29 Scenario 2 Automation Factors - results

Automation Factors	Reliability	Feedback	Hazardous	Effectiveness	Ease of Use	Integrity	Accuracy	Suitability
Average	1.78	1.69	1.22	1.31	4.19	1.81	1.34	1.19
Std. Deviation	1.70	1.06	0.94	0.59	1.91	1.42	0.79	0.40

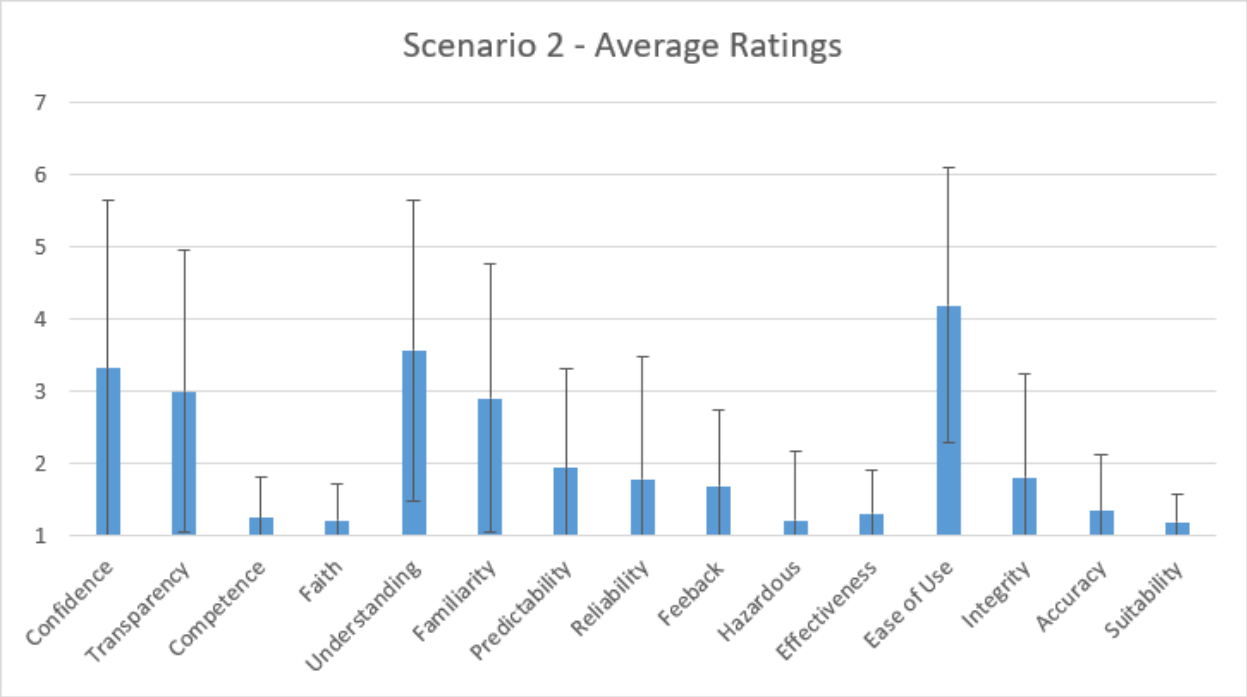

Figure 19 Scenario 2 ratings distribution

The data collected for Scenario 2 provides evidence to support acceptance of the hypothesis that a negative automation experience will result in negative TIA ratings. The strongest negative ratings included pilot perception of their confidence and faith of the system, as well as the safety and suitability of the automation. The highest performing factor was "Ease of Use", an expected result due to the minimal interaction requirements. In general, pilots did not trust the system and did not feel that the system provided a sense of trust or the required performance to be effective. Overall, pilots rated Scenario 2 negatively and considered the automated task performance to be a failure.

Scenario 3

Scenario 3 was a dynamic scenario in which most of the ratings were expected to be alternating between the overarching factors. Initial expectations were that the human factors

would correspond to the Somewhat Disagree – Strongly Disagree ratings, and the automation factors would correspond to the Somewhat Agree – Strongly Agree ratings. The descriptive statistics of scenario 3 somewhat aligned with the proposed hypothesis 4 for the second study. Further analysis, in Results – Factor Analysis, was required to ensure the factors are aligned within the appropriate subscales.

a. Hypothesis 4: Scenario-based validity testing will identify two overarching factors (human and automation) for alternating imperfect automation scenarios.

Table 30 and 31 provide the descriptive statistics and figure 20 provides the distribution of the ratings.

Table 30 Scenario 3 Human Factors - results

Human Factors	Confidence	Transparency	Competence	Faith	Understanding	Familiarity	Predictability
Average	2.31	4.66	6.03	5.22	4.03	2.31	4.66
Std. Deviation	1.35	1.64	0.93	1.43	1.79	1.06	1.68

Table 31 Scenario 3 Automation Factors - results

Automation Factors	Reliability	Feedback	Hazardous	Effectiveness	Ease of Use	Integrity	Accuracy	Suitability
Average	5.81	4.56	4.72	5.72	4.09	5.56	5.84	5.09
Std. Deviation	1.31	1.76	1.49	1.33	1.65	1.24	1.11	1.42

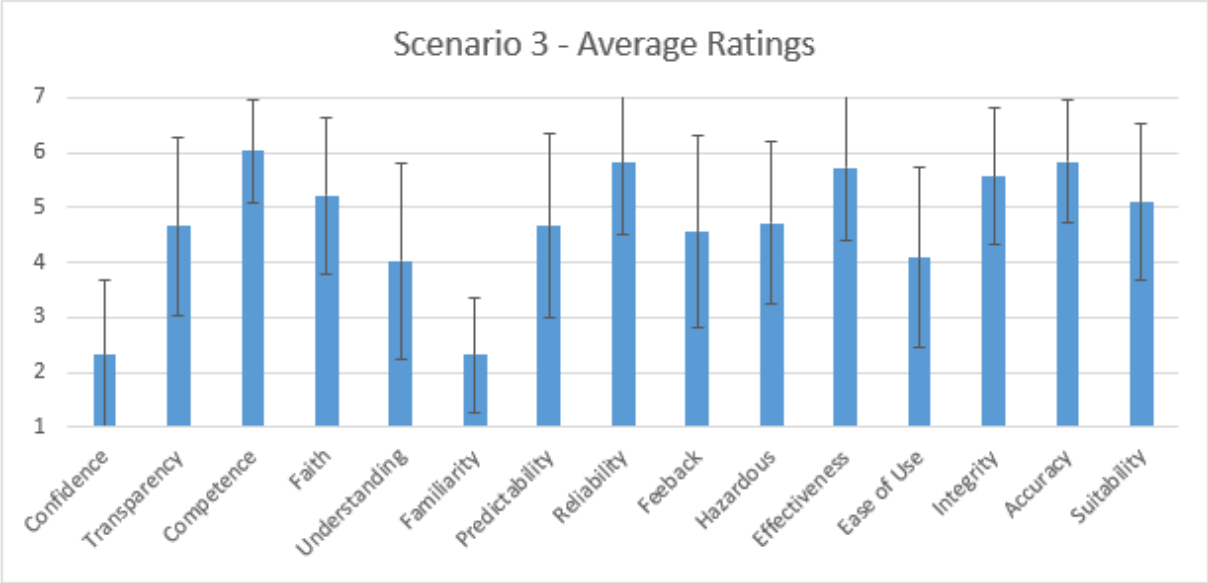

Figure 20 Scenario 3 ratings distribution

The data collected for Scenario 3 did not provide sufficient evidence to accept the hypothesis that a negative automation experience related to training and minimal pre-knowledge will result in negative TIA ratings for human factors, while successful system performance would result in higher automation factors. In this case, human factors such as faith, transparency, and technical competence rated sufficiently high to average as a positive rating. However, the automation factors did generally fall within the positive ratings as expected.

The strongest negative ratings included pilot perception of their confidence and familiarity with the system. Overall, pilots rated scenario 3 positively for its performance on safely avoiding the obstacle and negatively in their general confidence in using the system.

Scenario 4

Scenario 4 was a dynamic scenario in which most of the ratings were expected to be alternating between the overarching factors. Initial expectations were that the human factors

would correspond to the Somewhat Agree – Strongly Agree ratings, and the automation factors would correspond to the Somewhat Disagree – Strongly Disagree ratings. The descriptive statistics of scenario 4 somewhat aligned with the proposed hypothesis 4 for the second study. Further analysis, in Results – Factor Analysis, was required to ensure the factors are aligned within the appropriate subscales.

a. Hypothesis 4: Scenario-based validity testing will identify two overarching factors (human and automation) for alternating imperfect automation scenarios.

Table 32 and 33 provide the descriptive statistics and figure 21 provides the distribution of the ratings.

Table 32 Scenario 4 Human Factors - results

Human Factors	Confidence	Transparency	Competence	Faith	Understanding	Familiarity	Predictability
Average	5.31	4.47	1.94	1.69	4.97	5.97	3.38
Std. Deviation	1.77	1.70	1.16	0.82	1.28	0.69	2.01

Table 33 Scenario 4 Automation Factors - results

Automation Factors	Reliability	Feedback	Hazardous	Effectiveness	Ease of Use	Integrity	Accuracy	Suitability
Average	2.94	3.03	1.59	1.84	5.31	2.53	2.66	1.59
Std. Deviation	2.08	1.87	1.16	0.77	1.40	1.65	1.64	0.80

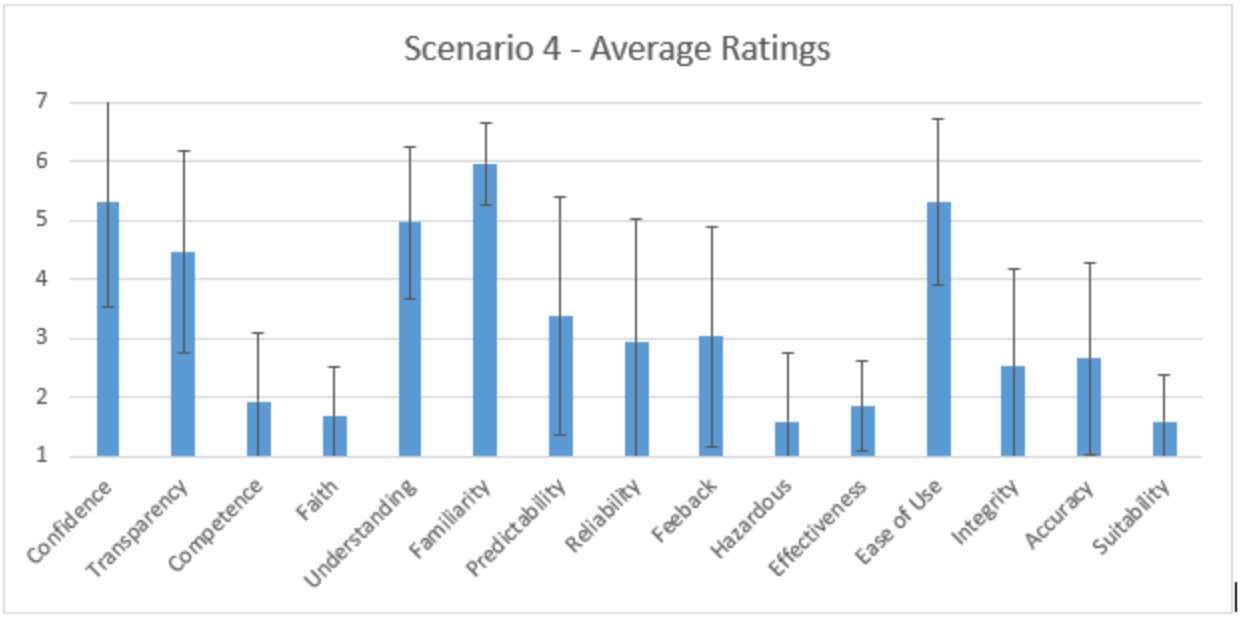

Figure 21 Scenario 4 ratings distribution

The data collected for scenario 4 mostly aligned with the proposed hypothesis and provided evidence for acceptance. Human factors averaged higher scores than automation factors. Faith and technical competence were the primary human factors that scored much lower than the others, due to the system not completing the task. The highest performing factor was familiarity. In general, pilots did not trust the system and did not feel that the system provided an adequate sense of trust or the required performance to be effective. Overall, pilots rated scenario 4 negatively and considered the automated task performance to be a failure.

Discussion - Scenarios

Prior to the factor analysis, some interesting findings deserve discussion regarding the pilot ratings. Scenario 1 ratings were very much expected due to the success of the system, minimal interactions, and thorough pre-knowledge. Nearly all pilot ratings were highly positive of the scenario's trust interaction.

Ratings collected during scenario 2 were somewhat unexpected when confidence and understanding were considered. Although the aircraft crashed in scenario 2, it initially seemed as if the automations were working. Pilots who had a high confidence score reported that even though the system failed, they were still able to take the "good" data of a detection and use that to their advantage. The sentiment was that if there is any data that can be used, pilots will find a way to use it. In a flight situation, pilots may know the system is a poor performer and still be appreciative of the limited "good" data that is available to them. Having confidence in their own ability to adjust or calibrate to the automation resulted in higher ratings. Similarly, understanding received higher average ratings. Pilots felt that they understood what the system was trying to do, whether it accomplished the task or not.

Scenario 3 ratings were also interesting. pre-knowledge provided to the pilots provided them with minimal system training, but an expectation of successful task completion. Confidence and familiarity were rated low, while other factors were rated positively, consistent with the pre-knowledge. Pilots reported that the highly rated factors were due to the success of the system and their "trust" for the system to save the aircraft in the emergency. In contrast to scenario 2, confidence was rated low. This was likely due to the pilots not receiving any data prior to the emergency interaction. They were unable to calibrate their initial trust in the system reducing their confidence of using the system in any manner to preemptively avoid obstacles.

Scenario 4 exhibited a similar trend as scenario 2. Pilots were able to adjust their confidence in using the system based on pre-knowledge of system behavior and their understanding. In this case, knowledge of the system is paramount even if it's a poor performer. Allowing pilots to adjust their expectations and calibrate their trust can provide the pilot

confidence and understanding. Pilots reported that knowing the limitations of the system can help them employ the system in a way to optimize the use cases and calibrate their reliance.

Results – Factor Analysis

Results of pilot ratings during the pretest were analyzed using statistical software (e.g., MS Excel, SPSS) to perform PCA and EFA on the data to check for dimensionality and examine scale reliability using the Cronbach's alpha metric.

A PCA was performed on all scenarios and an EFA was performed for scenarios 2, 3, and 4, which showed a likelihood towards multiple factors, to examine the potential subscale factors, as those scenarios were varied with pre-knowledge and scenario outcomes (i.e., positive and negative) and are useful for determining factor separations. The intent of analyzing the scenarios was to find scale items that commonly load on separate factors. The expectation being that these factors are most likely separate dimensional measurements (i.e., human and automation) when examined over multiple scenarios. Scale items that crossed into both factor groupings over the scenarios were further evaluated for survey inclusion.

Scenario 2, 3, and 4 data were analyzed using SPSS EFA methods of dimension reduction, generating a correlation matrix of coefficients, and examining the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy, and Bartlett's Test of Sphericity. KMO is a statistic used to determine how suited the data is for factor analysis, with a value larger than 0.6 as being acceptable (Nasaireh, 2020). A significant result (i.e., $p < 0.05$) for Bartlett's Test of Sphericity indicates that at least two of the items are correlated and factor analysis can be continued for data reduction (Bartlett, 1951). Eigenvalues were then computed and compared to a parallel analysis, and a scree plot was generated for scenario 2, 3, and 4 to determine factor dimensions. The parallel analysis is used to assist in determining the number of components to keep when

conducting a PCA (Lim & Jahng, 2019). The Oblimin test for oblique data was conducted for scenarios 2, 3, and 4. In both cases the data were determined to be orthogonal, due to neither factor for each scenario loading over 0.50 (correlation) in the Oblimin test. Orthogonal-Varimax rotation was used to generate a final rotated component matrix for the scenarios.

The correlation matrix for scenario 1 is provided in figure 22. The KMO Measure of Sampling Adequacy exceeded the threshold of 0.50, indicating that the sample size was adequate for the correlation analysis (figure 23). Additionally, the Bartlett's Test indicated that the data were significant, meaning that at least one significant correlation between two of the items (figure 23).

Correlation Matrix[a]

		Confidence	Transparency	Competence	Faith	Understanding	Familiarity	Predictability	Reliability	Feedback	Hazardous	Effectiveness	Easy	Integrity	Accuracy	Suitability
Correlation	Confidence	1.000	.082	.181	.512	.087	.491	.238	.534	.077	.371	.264	.267	.531	.611	.251
	Transparency	.082	1.000	.340	.124	.417	.280	.220	.123	.243	.333	.314	.278	.196	-.003	.340
	Competence	.181	.340	1.000	.538	.131	.268	.193	.257	.368	.415	.451	.424	.363	.391	.766
	Faith	.512	.124	.538	1.000	.035	.254	.351	.431	.174	.442	.271	.253	.398	.512	.460
	Understanding	.087	.417	.131	.035	1.000	.226	.668	.059	.470	.542	.176	.285	.021	.161	.130
	Familiarity	.491	.280	.268	.254	.226	1.000	.365	.417	.374	.274	.429	.473	.688	.410	.317
	Predictability	.238	.220	.193	.351	.668	.365	1.000	.251	.330	.622	.118	.027	.341	.238	.201
	Reliability	.534	.123	.257	.431	.059	.417	.251	1.000	.111	.208	.783	.374	.445	.534	.440
	Feedback	.077	.243	.368	.174	.470	.374	.330	.111	1.000	.454	.300	.433	.232	.292	.372
	Hazardous	.371	.333	.415	.442	.542	.274	.622	.208	.454	1.000	.225	.156	.376	.441	.472
	Effectiveness	.264	.314	.451	.271	.176	.429	.118	.783	.300	.225	1.000	.583	.336	.556	.559
	Easy	.267	.278	.424	.253	.285	.473	.027	.374	.433	.156	.583	1.000	.265	.522	.405
	Integrity	.531	.196	.363	.398	.021	.688	.341	.445	.232	.376	.336	.265	1.000	.615	.442
	Accuracy	.611	-.003	.391	.512	.161	.410	.238	.534	.292	.441	.556	.522	.615	1.000	.449
	Suitability	.251	.340	.766	.460	.130	.317	.201	.440	.372	.472	.559	.405	.442	.449	1.000

Figure 22 Scenario 1 correlation matrix

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.648
Bartlett's Test of Sphericity	Approx. Chi-Square	271.750
	df	105
	Sig.	<.001

Figure 23 Scenario 1 KMO and Bartlett's Test

The scenario 2 correlation matrix is provided in figure 24, followed by KMO and Bartlett's Test in figure 25.

Correlation Matrix[a]

		Confidence	Transparency	Competence	Faith	Understanding	Familiarity	Predictability	Reliability	Feedback	Hazardous	Effectiveness	Easy	Integrity	Accuracy	Suitability
Correlation	Confidence	1.000	.248	-.085	.164	.435	.558	.006	.115	.432	-.061	.044	.537	.193	-.043	.004
	Transparency	.248	1.000	-.233	-.067	.374	.321	.169	.000	.374	-.123	-.251	.373	-.093	-.126	-.083
	Competence	-.085	-.233	1.000	.723	.068	.084	.270	.125	.027	.739	.719	-.193	.658	.595	.358
	Faith	.164	-.067	.723	1.000	.160	.130	.310	.214	.136	.521	.645	.024	.753	.717	.445
	Understanding	.435	.374	.068	.160	1.000	.642	.229	-.046	.287	.149	.167	.461	.244	-.142	.064
	Familiarity	.558	.321	.084	.130	.642	1.000	.049	-.109	.165	.123	.057	.324	.298	-.176	-.019
	Predictability	.006	.169	.270	.310	.229	.049	1.000	.453	-.081	.161	.304	.116	.126	.380	.141
	Reliability	.115	.000	.125	.214	-.046	-.109	.453	1.000	.086	.051	.102	.073	.076	.372	.063
	Feedback	.432	.374	.027	.136	.287	.165	-.081	.086	1.000	-.091	.058	.396	.131	.171	.067
	Hazardous	-.061	-.123	.739	.521	.149	.123	.161	.051	-.091	1.000	.394	-.293	.609	.243	.059
	Effectiveness	.044	-.251	.719	.645	.167	.057	.304	.102	.058	.394	1.000	-.054	.454	.661	.567
	Easy	.537	.373	-.193	.024	.461	.324	.116	.073	.396	-.293	-.054	1.000	.013	.149	.336
	Integrity	.193	-.093	.658	.753	.244	.298	.126	.076	.131	.609	.454	.013	1.000	.376	.350
	Accuracy	-.043	-.126	.595	.717	-.142	-.176	.380	.372	.171	.243	.661	.149	.376	1.000	.613
	Suitability	.004	-.083	.358	.445	.064	-.019	.141	.063	.067	.059	.567	.336	.350	.613	1.000

Figure 24 Scenario 2 correlation matrix

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.589
Bartlett's Test of Sphericity	Approx. Chi-Square	252.048
	df	105
	Sig.	<.001

Figure 25 Scenario 2 KMO and Bartlett's Test

For scenario 2, the KMO Measure of Sampling Adequacy exceeded the threshold of 0.50, implying that the sample size was adequate for the correlation analysis. The Bartlett's Test indicated that the data were significant, indicating at least one significant correlation between two of the items.

Scenario 3's correlation matrix is provided in figure 26, followed by KMO and Bartlett's Test in figure 27.

Correlation Matrix

		Confidence	Transparency	Competence	Faith	Understanding	Familiarity	Predictability	Reliability	Feedback	Hazardous	Effectiveness	Easy	Integrity	Accuracy	Suitability
Correlation	Confidence	1.000	.486	.094	.430	.249	.446	.035	.071	.222	.013	.122	.534	.275	.119	.336
	Transparency	.486	1.000	.113	.391	-.018	.008	.132	-.031	.405	.065	.162	.286	.050	.058	.125
	Competence	.094	.113	1.000	.309	-.252	-.173	.213	.243	.028	.286	.242	.082	.179	.347	.119
	Faith	.430	.391	.309	1.000	-.041	.379	.570	.368	.283	.318	.442	.291	.455	.570	.513
	Understanding	.249	-.018	-.252	-.041	1.000	.539	.079	-.191	.138	-.130	.086	.283	.152	.100	.303
	Familiarity	.446	.008	-.173	.379	.539	1.000	.352	-.073	.179	-.024	.087	.351	.156	.125	.343
	Predictability	.035	.132	.213	.570	.079	.352	1.000	.500	.439	.426	.666	-.139	.575	.611	.541
	Reliability	.071	-.031	.243	.368	-.191	-.073	.500	1.000	.146	.454	.490	-.201	.643	.624	.392
	Feedback	.222	.405	.028	.283	.138	.179	.439	.146	1.000	.371	.471	-.085	.426	.245	.326
	Hazardous	.013	.065	.286	.318	-.130	-.024	.426	.454	.371	1.000	.302	-.041	.333	.188	.272
	Effectiveness	.122	.162	.242	.442	.086	.087	.666	.490	.471	.302	1.000	-.091	.882	.758	.562
	Easy	.534	.286	.082	.291	.283	.351	-.139	-.201	-.085	-.041	-.091	1.000	.005	-.009	.133
	Integrity	.275	.050	.179	.455	.152	.156	.575	.643	.426	.333	.882	.005	1.000	.767	.590
	Accuracy	.119	.058	.347	.570	.100	.125	.611	.624	.245	.188	.758	-.009	.767	1.000	.581
	Suitability	.336	.125	.119	.513	.303	.343	.541	.392	.326	.272	.562	.133	.590	.581	1.000

Figure 26 Scenario 3 correlation matrix

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.665
Bartlett's Test of Sphericity	Approx. Chi-Square	257.353
	df	105
	Sig.	<.001

Figure 27 Scenario 3 KMO and Bartlett's Test

For scenario 3, the KMO Measure of Sampling Adequacy exceeded the threshold of 0.50, indicating that the sample size was adequate for the correlation analysis. Additionally, the Bartlett's Test indicated that the data were significant, meaning that at least one significant correlation between two of the items.

The correlation matrix, KMO, and Bartlett's test for scenario 4 are provided in figure's 28 and 29.

Correlation Matrix[a]

		Confidence	Transparency	Competence	Faith	Understanding	Familiarity	Predictability	Reliability	Feedback	Hazardous	Effectiveness	Easy	Integrity	Accuracy	Suitability
Correlation	Confidence	1.000	.357	-.163	-.153	.090	.428	.129	-.065	-.110	.237	.037	.168	.074	.127	.070
	Transparency	.357	1.000	.260	.293	.273	.095	.201	.300	.197	.018	.206	.058	.242	.568	.240
	Competence	-.163	.260	1.000	.655	-.001	-.162	.231	-.068	.534	.196	.640	.151	.220	.242	.389
	Faith	-.153	.293	.655	1.000	-.040	-.074	-.044	.026	.468	.100	.689	.088	.270	.301	.391
	Understanding	.090	.273	-.001	-.040	1.000	.288	.180	.387	.255	.100	.093	.383	.146	.286	.082
	Familiarity	.428	.095	-.162	-.074	.288	1.000	.170	-.091	-.123	.024	-.191	.275	-.041	.019	.093
	Predictability	.129	.201	.231	-.044	.180	.170	1.000	.322	.450	.371	.123	.243	.289	.275	.379
	Reliability	-.065	.300	-.068	.026	.387	-.091	.322	1.000	.133	.096	-.006	.251	.227	.496	-.269
	Feedback	-.110	.197	.534	.468	.255	-.123	.450	.133	1.000	.525	.587	.205	.256	.203	.613
	Hazardous	.237	.018	.196	.100	.100	.024	.371	.096	.525	1.000	.180	.100	-.120	.043	.269
	Effectiveness	.037	.206	.640	.689	.093	-.191	.123	-.006	.587	.180	1.000	.257	.503	.264	.526
	Easy	.168	.058	.151	.088	.383	.275	.243	.251	.205	.100	.257	1.000	.275	.231	.088
	Integrity	.074	.242	.220	.270	.146	-.041	.289	.227	.256	-.120	.503	.275	1.000	.405	.243
	Accuracy	.127	.568	.242	.301	.286	.019	.275	.496	.203	.043	.264	.231	.405	1.000	.161
	Suitability	.070	.240	.389	.391	.082	.093	.379	-.269	.613	.269	.526	.088	.243	.161	1.000

Figure 28 Scenario 4 correlation matrix

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.572
Bartlett's Test of Sphericity	Approx. Chi-Square	193.322
	df	105
	Sig.	<.001

Figure 29 Scenario 4 KMO and Bartlett's Test

Like scenarios 1, 2 and 3, the KMO Measure of Sampling Adequacy exceeded the threshold of 0.50 for correlation analysis and the Bartlett's Test found a significant correlation between at least two of the items.

The eigenvalues for scenario 1 are provided in figure 30. Parallel analysis for scenario 1 and the scree plot are provided in figure 31 and figure 32. The second component eigenvalue was approximately the same as the parallel analysis eigenvalue cutoff (2.0). When viewing the scree plot for scenario 1 (figure 32) and the component matrix for scenario 1 (figure 33), a strong case is presented for a single factor model of the items, where each item loaded onto factor 1 positively with moderately strong correlation (Hair et al., 1998).

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.895	39.300	39.300	5.895	39.300	39.300
2	2.021	13.470	52.770	2.021	13.470	52.770
3	1.515	10.097	62.867	1.515	10.097	62.867
4	1.226	8.172	71.039	1.226	8.172	71.039
5	.889	5.923	76.962			
6	.879	5.863	82.825			
7	.645	4.301	87.126			
8	.493	3.289	90.415			
9	.416	2.773	93.188			
10	.318	2.118	95.305			
11	.220	1.468	96.773			
12	.198	1.317	98.090			
13	.142	.946	99.036			
14	.098	.654	99.690			
15	.046	.310	100.000			

Extraction Method: Principal Component Analysis.

Figure 30 Scenario 1 eigenvalues

Component or Factor	Mean Eigenvalue	Percentile Eigenvalue
1	2.409994	2.794859
2	2.007457	2.261308
3	1.723161	1.950582
4	1.508967	1.668512
5	1.306420	1.455493
6	1.135201	1.251337
7	0.995223	1.088901
8	0.870492	0.968848
9	0.753053	0.861485
10	0.646461	0.728472
11	0.538856	0.654886
12	0.439243	0.539791
13	0.329417	0.452887
14	0.229435	0.369025
15	0.106620	0.249579

Figure 31 Scenario 1 parallel analysis

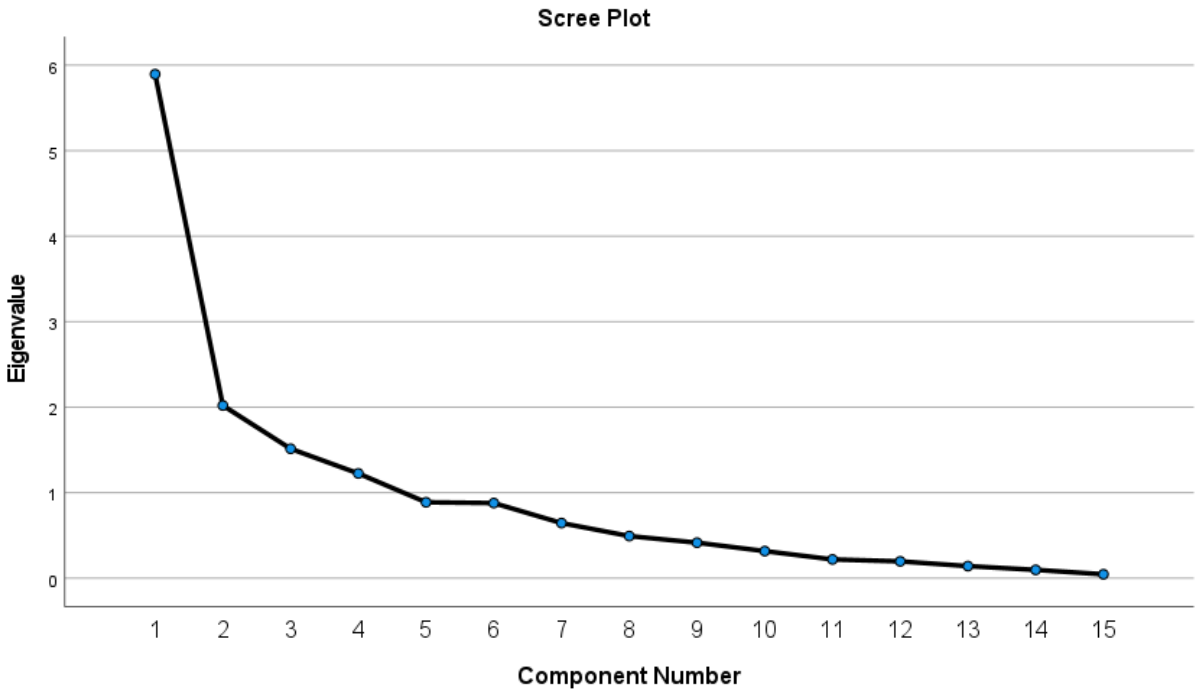

Figure 32 Scenario 1 scree plot

Component Matrix[a]

	Component			
	1	2	3	4
Accuracy	.754	-.310	-.168	.035
Suitability	.727	-.058	.348	-.394
Effectiveness	.702	-.260	.400	.242
Integrity	.695	-.234	-.305	.052
Familiarity	.675	-.061	-.135	.433
Reliability	.667	-.420	-.037	.203
Competence	.664	.008	.384	-.509
Hazardous	.658	.462	-.258	-.275
Faith	.634	-.176	-.230	-.489
Easy	.620	-.103	.444	.336
Confidence	.607	-.323	-.491	.081
Feedback	.536	.434	.231	.118
Transparency	.415	.405	.339	.056
Understanding	.411	.771	-.050	.265
Predictability	.515	.559	-.466	.004

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Figure 33 Scenario 1 component matrix

The eigenvalues for scenario 2 are provided in figure 34. A parallel analysis (figure 35) and scree plot (figure 36) were also generated to identify the appropriate number of components for further analysis. The third component eigenvalue was approximately equal to the parallel analysis eigenvalue cutoff, indicating that at least two factors should be considered for analysis. The scree plot also shows two factors with separation from the remaining components.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.548	30.322	30.322	4.548	30.322	30.322
2	3.061	20.408	50.730	3.061	20.408	50.730
3	1.741	11.607	62.338	1.741	11.607	62.338
4	1.297	8.649	70.987	1.297	8.649	70.987
5	.968	6.450	77.437			
6	.827	5.511	82.948			
7	.571	3.808	86.757			
8	.444	2.961	89.717			
9	.375	2.498	92.215			
10	.343	2.287	94.502			
11	.277	1.847	96.349			
12	.245	1.634	97.983			
13	.144	.961	98.944			
14	.106	.705	99.649			
15	.053	.351	100.000			

Extraction Method: Principal Component Analysis.

Figure 34 Scenario 2 eigenvalues

Component or Factor	Mean Eigenvalue	Percentile Eigenvalue
1	2.409994	2.794859
2	2.007457	2.261308
3	1.723161	1.950582
4	1.508967	1.668512
5	1.306420	1.455493
6	1.135201	1.251337
7	0.995223	1.088901
8	0.870492	0.968848
9	0.753053	0.861485
10	0.646461	0.728472
11	0.538856	0.654886
12	0.439243	0.539791
13	0.329417	0.452887
14	0.229435	0.369025
15	0.106620	0.249579

Figure 35 Scenario 2 parallel analysis

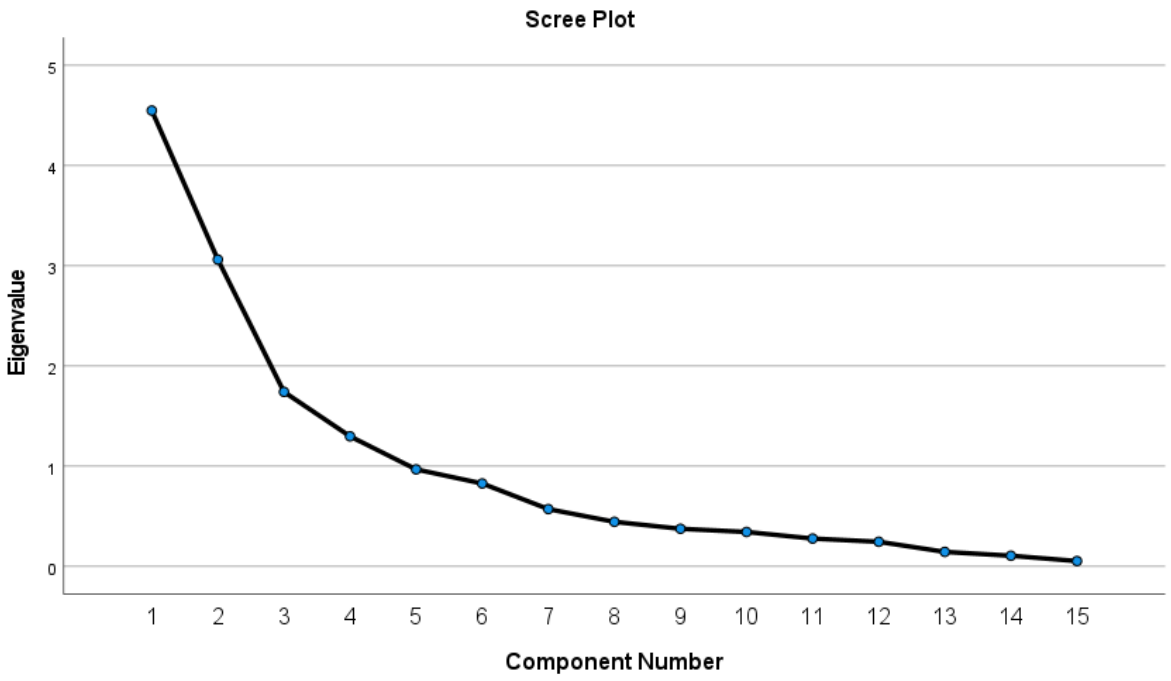

Figure 36 Scenario 2 scree plot

Eigenvalues were calculated for scenario 3 and provided in figure 37. A parallel analysis (figure 38) and scree plot (figure 39) were also generated to identify the appropriate number of components for further analysis. The third component eigenvalue was less than the parallel analysis eigenvalue cutoff, indicating that two factors should be retained for analysis. This calculation was verified by the scree plot.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.262	35.078	35.078	5.262	35.078	35.078
2	2.504	16.691	51.770	2.504	16.691	51.770
3	1.606	10.709	62.479	1.606	10.709	62.479
4	1.199	7.996	70.474	1.199	7.996	70.474
5	.921	6.143	76.618			
6	.727	4.849	81.467			
7	.699	4.662	86.129			
8	.469	3.129	89.258			
9	.449	2.991	92.249			
10	.384	2.560	94.809			
11	.270	1.797	96.606			
12	.231	1.540	98.146			
13	.124	.827	98.973			
14	.110	.730	99.703			
15	.044	.297	100.000			

Extraction Method: Principal Component Analysis.

Figure 37 Scenario 3 eigenvalues

Component or Factor	Mean Eigenvalue	Percentile Eigenvalue
1	2.409994	2.794859
2	2.007457	2.261308
3	1.723161	1.950582
4	1.508967	1.668512
5	1.306420	1.455493
6	1.135201	1.251337
7	0.995223	1.088901
8	0.870492	0.968848
9	0.753053	0.861485
10	0.646461	0.728472
11	0.538856	0.654886
12	0.439243	0.539791
13	0.329417	0.452887
14	0.229435	0.369025
15	0.106620	0.249579

Figure 38 Scenario 3 parallel analysis

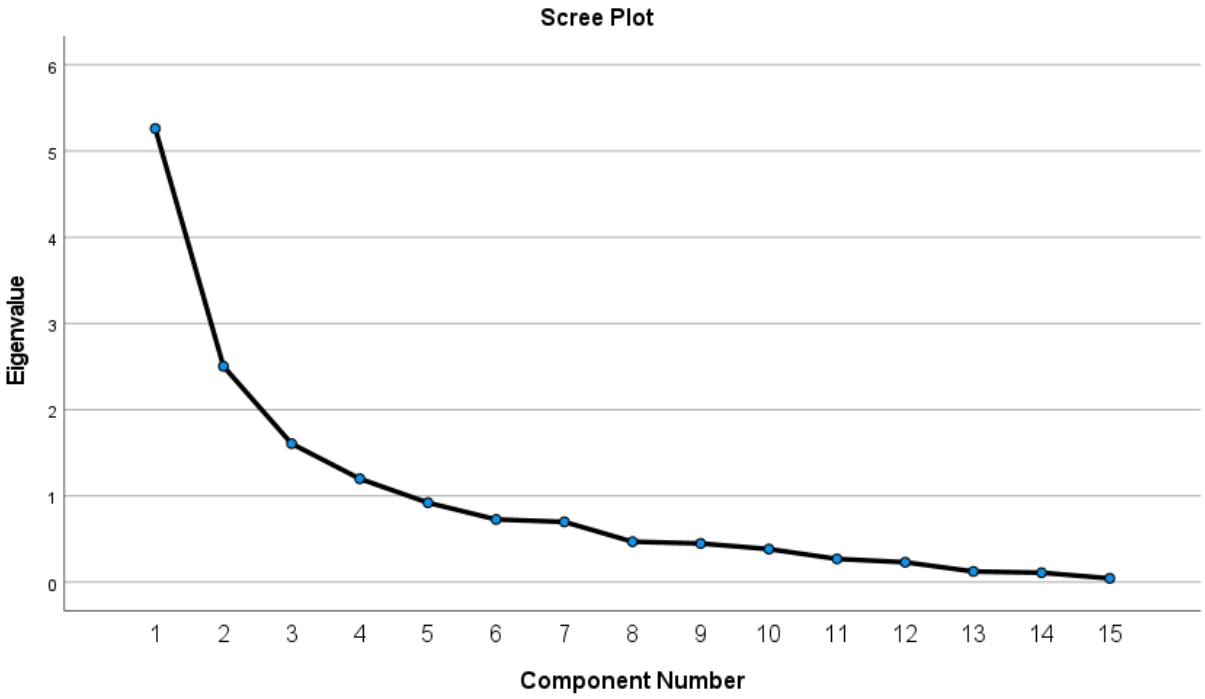

Figure 39 Scenario 3 scree plot

Scenario 4 had similar results to scenario 3. The eigenvalues were calculated (figure 40) and compared to the parallel analysis in figure 41. The third component was less than the parallel analysis and two factors were recommended by the analysis. The scenario 4 scree plot (figure 42) also showed a strong recommendation for retaining two factors.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.203	28.017	28.017	4.203	28.017	28.017
2	2.339	15.591	43.608	2.339	15.591	43.608
3	1.694	11.293	54.901	1.694	11.293	54.901
4	1.384	9.225	64.126	1.384	9.225	64.126
5	1.124	7.491	71.617	1.124	7.491	71.617
6	.920	6.134	77.751			
7	.788	5.255	83.007			
8	.656	4.375	87.382			
9	.480	3.197	90.578			
10	.424	2.824	93.402			
11	.345	2.298	95.700			
12	.248	1.654	97.355			
13	.173	1.151	98.506			
14	.121	.808	99.314			
15	.103	.686	100.000			

Extraction Method: Principal Component Analysis.

Figure 40 Scenario 4 eigenvalues

Component or Factor	Mean Eigenvalue	Percentile Eigenvalue
1	2.409994	2.794859
2	2.007457	2.261308
3	1.723161	1.950582
4	1.508967	1.668512
5	1.306420	1.455493
6	1.135201	1.251337
7	0.995223	1.088901
8	0.870492	0.968848
9	0.753053	0.861485
10	0.646461	0.728472
11	0.538856	0.654886
12	0.439243	0.539791
13	0.329417	0.452887
14	0.229435	0.369025
15	0.106620	0.249579

Figure 41 Scenario 4 parallel analysis

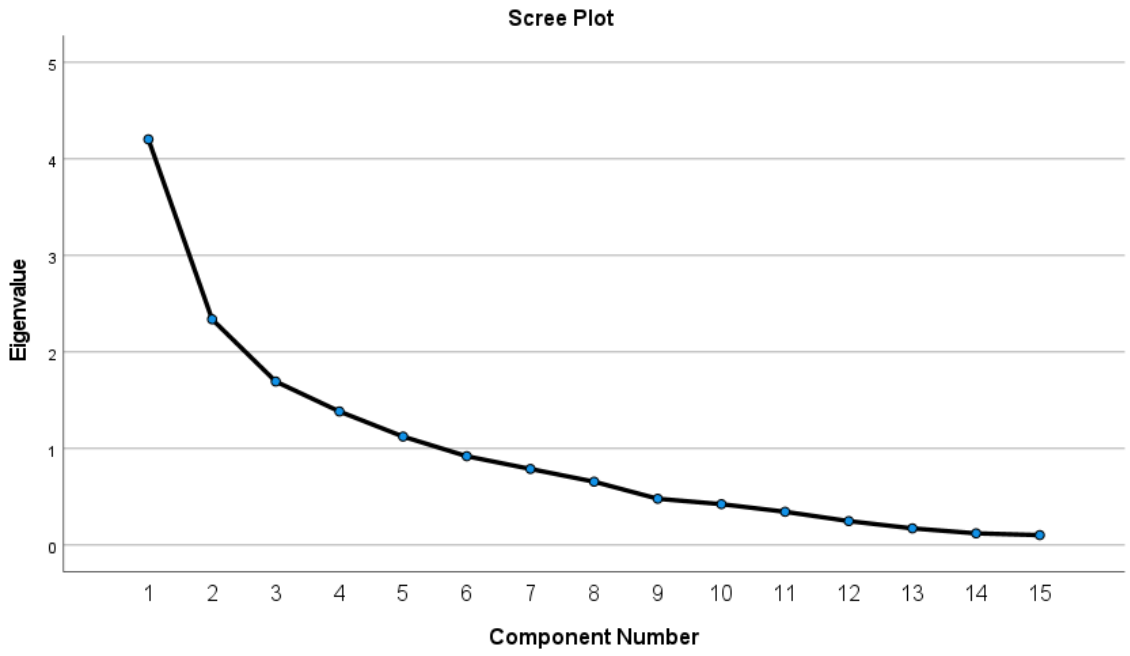

Figure 42 Scenario 4 scree plot

Initially, a component correlation matrix was computed, with the number of extracted factors set to 2, to check for oblique factors for scenarios 2, 3, and 4. In scenarios 2, 3, and 4, the correlation-relationship was less than 0.50, indicating that the factors are orthogonal. The component correlation matrix for scenario 2 is provided in figure 43, scenario 3 is provided in figure 44, and scenario 4 is shown in figure 45.

Component Correlation Matrix

Component	1	2
1	1.000	.375
2	.375	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Figure 43 Scenario 2 component correlation matrix

Component Correlation Matrix

Component	1	2
1	1.000	.122
2	.122	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Figure 44 Scenario 3 component correlation matrix

Component Correlation Matrix

Component	1	2
1	1.000	.166
2	.166	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Figure 45 Scenario 4 component correlation matrix

The rotation was then switched to varimax for orthogonal analysis, resulting in a final rotated component matrix. The Varimax rotated component matrix for scenario 2 is shown in figure 46, scenario 3 in figure 47 and scenario 4 in figure 48.

Rotated Component Matrix[a]

	Component	
	1	2
Faith	.890	
Competence	.886	
Effectiveness	.824	
Accuracy	.788	
Integrity	.755	
Hazardous	.647	
Suitability	.582	
Predictability	.410	
Reliability		
Confidence		.771
Understanding		.760
Easy		.752
Familiarity		.716
Transparency		.622
Feedback		.582

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Figure 46 Scenario 2 rotated component matrix

Rotated Component Matrix[a]

	Component	
	1	2
Effectiveness	.859	
Integrity	.854	
Accuracy	.836	
Predictability	.797	
Reliability	.763	
Suitability	.648	.414
Faith	.620	.448
Hazardous	.549	
Feedback	.484	
Competence		
Confidence		.777
Familiarity		.744
Easy		.710
Understanding		.617
Transparency		.452

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Figure 47 Scenario 3 rotated component matrix

Rotated Component Matrix[a]

	Component	
	1	2
Effectiveness	.859	
Competence	.813	
Feedback	.793	
Faith	.785	
Suitability	.686	
Integrity	.436	
Hazardous		
Understanding		.659
Reliability		.614
Accuracy		.611
Transparency		.547
Easy		.544
Predictability		.525
Confidence		.490
Familiarity		.471

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Figure 48 Scenario 4 rotated component matrix

Discussion – Factor Analysis

The rotated component matrices were compared among scenarios 2, 3, and 4 to determine which scale items loaded independently of the others. Having independent scale items when examining the component matrix for each scenario allows for higher confidence that the items are indeed independent under different conditions. Common scale items among scenarios 2, 3, and 4 are presented in table 34 under their associated factor.

Table 34 Common scale items

Common Scale Items	
Factor 1	**Factor 2**
Understanding	Effectiveness
Confidence	Integrity
Familiarity	Faith
Transparency	Suitability

An examination of the remaining factors was conducted to determine whether any should be included as additional scale items, regardless of independence over the two scenarios.

- Technical Competence – while technical competence rated high in scenario 4, it did not meet the threshold requirement of 0.4 for scenario 3 to load on either factor. Technical competence is also closely related to "effectiveness" concerning the outcome of tasks. Therefore, technical competence was not included in the scale items proposed for further evaluation.
- Hazardous Conditions (Safety) – Safety loaded onto one factor moderately in scenario 3 but was below the threshold for factor loading in scenario 4. The safety metric was chosen not to move forward, as safety should be evaluated throughout experimentation or test events to ensure products meet all safety standards. More robust methods of objective and subjective safety measures are often in place during testing to ensure pilot safety. A safety failure likely indicates a significant issue with the system beyond the scope of the "Trust" metric. It is recommended that safety be always a consideration and "Trust" should be evaluated on systems

with approved safety standards and mitigations. The safety factor was not included in the further evaluation of scale items.

- Reliability – reliability is often measured objectively during aviation related test events to ensure robust system design. While pilot perception of reliability can potentially contribute to their system trust, the direct question of whether the system maintains reliable performance was somewhat confusing in negative scenarios. Pilots reported in some cases that the system was "reliably bad" and rated that they agree with "consistently negative" performance. Additionally, reliability is closely related to the concept of an effective system, where the system completes a task under established constraints. Objective reliability is a significantly more robust measurement for system characteristics, and like safety, systems under pilot test should maintain acceptable reliability for use. If objective reliability is low and pilot comments are generally negative towards the system performance, then it's likely "Trust" is of lesser concern until the system operates appropriately. The "Faith" scale item also covers the context of reliability through the definition of a belief that the system will perform the intended action. Reliability was not included in the further evaluation of scale items.
- Accuracy – accuracy loaded alternately onto both factors during scenarios 3 and 4. While the perception of correct decision making is important for trust, the concept also crosses into effectiveness of task completion and faith that the task will be completed. Additionally, accuracy can be measured objectively for most automated tasks. Tasks that provide inaccurate information or do not perform to standard should be corrected prior to production of the automated system. While

user observation can help to identify inaccurate tasks, objective measures should be in place to ensure adherence to the appropriate standards. Accuracy was not included in the further evaluation of scale items.

- Ease of Use – easy to use interfaces are essential in a general Human Factors Engineering approach to design. There are extensive surveys and measurement techniques in existence to examine ease of use and interface design beyond a single trust scale question. A combination of survey tools and techniques should be employed to examine whether ease of use responses on data collection tools correlate with trust scores on the TIA survey. In general, users of the TIA survey should consider how low and high ratings of an "ease of use" survey may impact data collected on the TIA survey. Higher ratings on an ease-of-use survey may indicate a positive trust experience and vice versa. Ease of use was not included in the further evaluation of scale items. Additionally, "ease of use" in the case of the scenarios was not examined in a way that required significant user interaction. Participants were only provided pre-knowledge on the system characteristics and relied on the pre-recorded video for interaction representations. This likely caused moderate ratings on "ease of use" due to the lack of pilot interaction. More complex scenarios with user interactions would likely have altered the ratings for "ease of use", however it is expected that usability responses could be adequately captured with an alternative survey method.
- Predictability – predictability is a useful concept for trust measurement, especially for trust calibration. However, it also covers a variety of other factors, including understanding, effectiveness, and familiarity. If users are familiar with the system,

understand the system, and deem the system effective, then the predictability of the system is inherently known. This correlation is present on the scenario 3 and 4 results, where predictability loaded on alternating factors. Predictability was not included in the further evaluation of scale items.

- Integrity – integrity was removed from moving forward for analysis due to its close association and correlation with effectiveness during positive outcome scenarios indicated by the correlation matrix. In both definitions of integrity and effectiveness, it's assumed that the automation operates under a guiding set of constraints or principles. In many cases, pilots commented that the two definitions seemed similar.
- Feedback - feedback loaded cleanly on scenarios 3 and 4 on a similar factor as other "automation" scale items and had mild correlation with scale items under the "human" designation on scenario 2. A decision was made to add feedback on the automation factor as a scale item. In scenario 2, pilots reported that the system provided partial feedback, it was just too late, likely contributing to the correlation with higher rated human factor scale items. Feedback is also a useful item, not necessarily captured by the remaining common items. Appropriate feedback allows for mental model adjustment (in real time) of a system and is an important trust factor.

Factor naming

Based on the independent scale items, the factor names were chosen to remain the same. Factor 1 considers the Human Factor items of Understanding, Confidence, Familiarity, and Transparency while the Automation Factor contains Effectiveness, Integrity, Feedback, Faith,

and Suitability. One noticeable difference from the hypothesized scale item and factor relationships is that Faith is now considered an automation factor. This change makes intuitive sense, where faith is adjusted based on system or automation outcomes more strongly than pre-knowledge of the intended system operation.

Results – Reliability Analysis

Cronbach's alpha was used to analyze the reliability of the independent scale items under factors 1 and 2. Table 35 shows the items considered for the reliability analysis.

Table 35 Scale items for reliability analysis

Factor Scale Items	
Human Factor	**Automation Factor**
Understanding	Effectiveness
Confidence	Feedback
Familiarity	Faith
Transparency	Suitability

A cutoff value of 0.6 was used as an initial reference for acceptable reliability. Scenario 1 was used in addition to the expected positive responses of scenarios 3 and 4 to determine the scale reliability. For example, scenario 1 responses were expected to be mostly positive when rated by all participants. Descriptive data from scenario 1 indicated that all scale items were rated positively. In scenario 3, the "automation factors" were expected to have a more positive outcome, since the automation successfully accomplished the task, while the participant was unclear on the "human factors". Automation factor ratings for scenarios 1 and 3 were combined

to determine whether positive outcomes for automation factors across two scenarios could be reliably measured. Similarly, data from scenarios 1 and scenario 4 were combined for the human factors, where the expectation was that positive ratings would be provided for those factors in both scenarios 1 and scenario 4.

Based on the independent item reliability analysis, the “human factors” resulted in a 0.698 Cronbach’s alpha (figure 49) and the “automation factors” resulted in a 0.781 Cronbach’s alpha (figure 50).

Reliability Statistics

Cronbach's Alpha	N of Items
.698	4

Figure 49 Human factors reliability

Reliability Statistics

Cronbach's Alpha	N of Items
.781	4

Figure 50 Automation factors reliability

The Cronbach’s alpha for the human factors (0.698) was very close to the 0.7 universal acceptability metric. However, 0.6 – 0.7 is often also considered an acceptable reliability range (Taber, 2018). Evaluation of the data suggests that in some cases, where the participant was well trained and understood the tasks, participants still rated the automation low for human factors when the task was not completed, especially in scenario 4 where the aircraft crashed at the end of the scenario. This evaluation was also confirmed by participant comments at the end of the

scenario. The evaluation suggests that had the aircraft not suffered a catastrophic failure, higher ratings would likely have occurred for the lower rated items, especially for the confidence and transparency factors. These higher ratings would increase scale correlation and overall reliability.

Cronbach's alpha for the automation factors (0.781) exceeded the 0.7 threshold of acceptability and fell within the acceptable 0.7 – 0.8 scale consistency range. In general, pilots seemed to more easily rate the automation factors when the automation outcome was clearly positive or negative, when compared to human factors where self-reflection and mental model adjustment were required to determine the level of trust.

Hypothesis testing

The results of the factor analysis provided an outcome of a one factor model for scenario 1 and two factors in scenarios 2, 3, and 4. When considering the anchor scenarios (1 and 2), data to support hypothesis 3 was sufficient to reject the statement that a single factor trust construct would result for both positive and negative experiences in anchor scenarios.

a. Hypothesis 3: Scenario-based validity testing will result in a single factor trust construct when analyzing anchor scenarios of positive and negative experience automation.

Scenario 1 was a positive experience anchor scenario. Data analysis for scenario 1 showed that hypothesis 3 was satisfied, where a single factor resulted based on PCA. However, data for scenario 2 showed a two-factor model utilizing PCA and EFA. The two-factor model rejected the totality of hypothesis 3, but still provided meaningful information and more fidelity into the item identification process for inclusion on the final survey. The two-factor model for scenario 2 likely resulted due to the approach to automation use for pilots when dealing with aviation systems. Pilots reported that even when the automation fails, if some usable information

is available, they could feel fairly confident in their interactions with the system and still maintain a level of understanding of the system operation.

Data collected for scenarios 3 and 4, were sufficient to accept hypothesis 4. Two overarching factors related to human and automation characteristics were identified through PCA and EFA for the scenarios.

b. Hypothesis 4: Scenario-based validity testing will identify two overarching factors (human and automation) for alternating imperfect automation scenarios.

Discussion – Reliability Analysis

The reliability analysis results confirmed that the scales for the human and automation factors were both sufficiently reliable to be considered moderately or highly correlated respectively. Reliability analysis is important to "measure whether questions that are intended to measure the same phenomenon are 'pulling in the same direction'" (Stensen & Lydersen, 2022). In the case of the human factor scale, the Cronbach's alpha outcome indicated that the scale was moderately acceptable due to the very close outcome statistic (0.698) to the universally accepted 0.70 for acceptable scale reliability, where some authors argue that 0.60 is still an acceptable statistic. For this survey, it's likely that less extreme scenarios would correlate more readily for human factors characteristics at moderate trust levels. A total lack of confidence, transparency, and understanding of the system (e.g., Strongly Disagree) would be less likely to occur if the aircraft had not crashed. The decision to keep these items moving forward was bolstered by the moderate acceptability from the reliability analysis, and the numerous occurrences of these scale items in other trust surveys, where analyses showed positive correlation with the trust construct.

Care should be taken when interpreting the human factors results to investigate outlying data to determine causes and collect user feedback for further interpretation.

Cronbach's alpha for the automation factors provided an acceptable scale consistency and reliability. All items from the automation factors were chosen to move forward for inclusion on the validated survey.

Results – Validated Survey

The final validated survey is provided in table 36 and is based on pilot ratings and the completed validity and reliability testing. Four scale items from the human factors and four scale items from the automation factors were included on the final survey.

Table 36 Validated AS-TS survey

Trust Questions	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
I am confident in my ability to utilize the system.							
The system provides transparent information.							
I understand what the system is doing.							
I am familiar with the system operation.							
I have faith that the system will perform the intended task.							
The system provides appropriate feedback on current and future actions.							
The system effectively accomplishes its tasks.							
The system is suitable for carrying out the task.							

"The system provides appropriate feedback on current and future actions" is a unique question in the survey. In some cases, automated systems may not be designed to provide "future action" information. When considering survey administration, consideration should be given to whether "current and future" actions should be listed, or only "current" actions. The expected level of autonomy and outcomes of the system should drive the wording of the survey. Autonomous systems with predictive capabilities should be investigated for current and future feedback and their appropriateness. Systems with only status messaging should be investigated for current feedback appropriateness. Pilot comments reflected this consideration, where concern was expressed on how to rate systems not designed for predictive "future" information.

If a system is not required to provide feedback, removing the feedback item from the survey would still provide acceptable reliability to the remaining "Automation Factor" scale. The remaining items would maintain a Cronbach's alpha reliability statistic of 0.781, greater than the acceptable 0.7. Further analysis would also recommend splitting the feedback question into two questions. One question that focuses on current actions and another question that focuses on future actions.

Conclusion

Data were collected from 32 pilot participants to validate the notional TIA survey scale items and their subscale reliability. Pilots were interviewed and surveyed after receiving an introductory briefing and participating in four virtual and scripted scenarios. At the conclusion of the data collection, PCA and EFA were conducted on the data to identify the total number of factors and eliminate poorly performing scale items. Eight commonly loaded scale items were retained after the EFA. Four under human factors: Confidence, Transparency, Understanding, and Familiarity and four under automation factors: Faith, Effectiveness, Suitability, and

Integrity. Each item that was removed was investigated further to ensure appropriate removal and adequate reasoning. An additional item (integrity) was removed under automation factors due to its close relationship in definition and results to the effectiveness item. Feedback was added as a scale item under automation factors as it cleanly loaded on two of the three scenarios and only mildly under the human factor scale for one scenario. Feedback is considered a valuable scale item, that does not significantly overlap with the other items, useful for calibrating and determining pilot mental models of systems and their trust relationship.

Once the factors were reduced and validation verified, Cronbach's alpha reliability testing was conducted on the two subscales (human and automation). The human factor scale items resulted in a 0.698 alpha and the automation factor scale items resulted in a 0.781. In both cases, the scale was considered to be moderately to strongly acceptable for reliability. The results of the EFA and reliability testing produced a validated survey tool for analyzing trust when using automated systems in the cockpit (table 36).

The survey underwent a third study to investigate the ease of distribution and acceptance of use by the Army test community, as well as determination of survey result comprehension and the relationship of trust to other measures such as reliability and usability.

CHAPTER V

SURVEY IMPLEMENTATION AND ANALYSIS

Introduction

TIA is a key concept in influencing user acceptance of new automated technology (Korber, Baseler, & Bengler, 2018). Both modern and future aircraft contain autonomous systems that play a major role in their use and development. Pilots must trust that the automation is performing to standard to ensure appropriate performance of the aircraft and satisfactory completion of pilot tasks. The U.S. Army is especially concerned with pilot TIA, as new technologies such as the Army FVL program are rapidly approaching. The FVL program is one of the Army's three major modernization priorities and includes initiatives for both a FARA and a FLRAA (Mayfield, 2021). A significant design consideration for FVL is optimizing the human-automation interactions that will occur during system use. FVL is required to provide multiple levels of supervised autonomy within the aircraft (e.g., autonomous takeoffs/landings, cueing, and adaptive interventions). Appropriate levels of TIA for the pilots will be extremely important for FVL platforms due to a significant focus on automated processes and automated assistance in high-speed and DVE rotorcraft operations (Freedberg, 2020). A series of studies are included in this research in support of a development and validation effort for a TIA survey and assessment methodology to assist in identifying pilot-automation trust deficiencies and provide potential mitigations.

Problem Statement

There is currently no standard methodology that is in use for the Army to assess TIA for pilots as a holistic measurement that identifies trust deficiencies and the relationship of trust to user reliance with follow-up actions. Several surveys are currently in use, but only used as a general indicator of system trust with no recommendations to improve the user-automation trust relationship. Similar issues exist across other research areas.

To address the lack of a standardized survey tool for TIA measurement in Army Aviation, a comprehensive literature review was conducted to identify key factors that may influence TIA in aviation systems (Chapter II). These identified factors were used in an initial study as a foundation to develop a pool of factors for review by SMEs through both interview and use of the AHP to refine the pool of factors to establish face validity for a notional TIA survey tool specialized for use in Army Aviation system assessments. Building on the previous literature review of identified factors and utilizing the AHP to solicit pilot input answered the research question "What factors influence TIA for Army pilots using Army Aviation systems?". Research and AHP analysis conducted in Chapter III of this dissertation were used as an initial study to establish content and face validity of the factors included in the notional survey.

Further validation and reliability analysis were required to answer the research question: "Can a survey instrument developed from identified TIA factors reliably measure pilot TIA perception of Army Aviation systems?". In Chapter IV of this research, as a follow-on study, 32 participants experienced a combination of four virtual automation-related scenarios to collect pilot ratings using the notional TIA survey. Pilot ratings were analyzed using factor analysis and scale reliability methods to establish construct validity and reliability of the survey.

Once the survey was validated, additional research was required to answer the following question and associated hypotheses, “Can the survey instrument be used effectively in formal design testing to provide actionable information to data analysts (e.g., Human Factors Engineers) and product managers?”.

a. Hypothesis 1: The developed survey will be a useful tool for analysts and program managers to identify TIA deficiencies, based on decision-maker and analyst ratings of effectiveness.

b. Hypothesis 2: The recommended actions list will provide appropriate courses of action to correct the deficiencies, based on decision-maker and analyst ratings of effectiveness.

A final study was conducted that showcases a use case of the survey to operational test experts to verify the ease of data collection and usefulness for decision-making. The primary difference between this study and the previous pre-test is that the survey tool was used with context for data collection and analysis by Army test evaluators. Evaluators examined the efficacy of the survey to help determine whether the survey would be an asset to the operational test data collection and decision-making process. Using the survey and understanding the context provides a use case for Army community acceptance of the survey tool and examines the practicality of distribution and data collection.

The first hypothesis for this study postulates that the developed TIA survey will be a useful tool for analysts and program managers to identify TIA deficiencies. Example data collection and analysis will determine the ease of use. Participant comments and ‘ease of use’ ratings will determine the utility of the responses. Once deficiencies are identified, an additional hypothesis posits that the recommended actions list provides appropriate courses of action to correct the deficiencies. Participant agreeance on recommended courses of action will verify or provide feedback to adjust the action recommendations.

Background Information

Operational Testing

Operational Testing (OT) is designed to test a system in an actual or simulated environment under realistic operational conditions with the target population (DoD, 2019). Operational test events often occur to evaluate new military system equipment to determine whether the system is operationally effective and operational suitability for mission use (DoD, 2019). As part of OT, data is collected to evaluate critical Human Systems Integration (HSI) domains (i.e., Manpower, Personnel, Training, Safety and Health Hazards, Human Factors Engineering, Force Protection and Survivability, and Habitability) relevant to the system under test (DoD, 2019). As part of this evaluation surveys are often used to collect data such as mental workload, situational awareness, usability, TIA, and objective data (e.g., biometrics, reliability, task timing). Operational test plans are put in place to meticulously collect data during these events for program management decision-making on both the strengths and weaknesses of the system under test and to shape the path forward and timelines for system acquisition.

Data Collection

Typically, survey responses and pilot comments are collected after a completed operational mission. Multiple crews perform missions and data is aggregated to identify trends and outliers in pilot responses. Human factors engineers and test evaluators collect the data and interview the pilots to verify context and investigate ratings.

Data Analysis

Data analysis for ratings scales like the AS-TS are often analyzed and examined using visual tools (e.g., frequency charts) for identification of high and low scoring items. Non-

parametric statistical testing (i.e., Mann-Whitney U) appropriate for ordinal data comparison is recommended for comparison (e.g., before and after automation improvements) test cases that assess automation changes (*Mann-Whitney U*, 2021). An overall composite score is not considered a relevant metric for the AS-TS as a multi-dimensional survey tool; however, subscale and item scores will be relevant and useful for identifying trust levels over the identified factors. AS-TS outcomes can be analyzed by a combination of descriptive statistics and hypothesis testing statistics.

Deficiencies

In the context of automation, a trust deficiency can be defined as an inadequate trust relationship between the user and the automated system, where the user perceives the system as untrustworthy and/or the user is unable to reasonably calibrate their trust in the system (Brzowski & Nathan-Roberts, 2019; Hoff & Bashir, 2015). When deficiencies are identified with products under test, the program management office must decide how to incorporate upgrades, enhancements, or additions to the current and future system (DoD, 2019). Utilizing a specific survey tool like the proposed AS-TS to identify trust deficiencies in system design has the potential to greatly assist program managers in developing mitigation plans to address issues.

Mitigations

Depending on the outcomes of the AS-TS, data analysts can review the survey ratings and any pilot comments for context of lower scoring factors. By considering the *human* and *automation* factors, a recommendation for a plan for improvements to enhance system TIA can be developed to address the deficiencies either with direct knowledge transfer to the pilot, system improvements, or both. The AS-TS will allow Program Managers to effectively identify

deficiencies and establish a plan to improve the system, based on ratings, pilot comments, and analyst recommendations.

Mitigation recommendations can then be provided to the program manager, test director, or included in the evaluation reports. A sample of mitigation techniques are provided for each factor in table 37. While mitigation can require a combination of methods (e.g., ratings, interview, and performance data) or further investigation for root cause, the intent of table 15 is to provide a starting point for addressing lower ratings that may be useful for evaluators and iterative automation developments. The mitigation recommendations were developed based on researcher experience and direct relationship to the provided definition list (Appendix B) for the AS-TS rating tool. The mitigations were also reviewed and rated for context and use case by participants in this study.

Table 37 Mitigation recommendations

Human Factors	Mitigation Recommendation
Confidence	Improve system training and knowledge of system limitations.
Transparency	Improve delivery of useful information to the user.
Understandability	Inform users on how and why the automation performs specific tasks to allow for accurate mental models of system processes.
Familiarity	Provide more training or opportunity to work with the system and historical context of use.
Automation Factors	**Mitigation Recommendation**
Feedback	Ensure the system provides appropriate user information and contextual future actions of the automation.
Effectiveness	Ensure the system can accomplish required mission tasks to standard.
Faith	Ensure automation consistently performs the intended actions.
Suitability	Ensure the system is being used for the appropriate tasks.

Methods

SME Assessment

Ten (10) SMEs (e.g., Army test evaluators, human factors engineers, and statisticians) provided feedback on the usefulness of the TIA survey in operational testing. Representation was provided from the following agencies: DEVCOM Analysis Center (DAC), Army Evaluation Center (AEC), USAARL, and Boeing. Two of the ten participants were not associated with the aviation field and were HFE experts in artillery, air and missile defense, and long-range precision

fires. Their feedback was solicited to determine the perception from non-aviation related personnel related to their interest levels for continuing evaluation for the AS-TS for potential applications across Army domains. Table 38 provides the demographic data collected from the participants.

Table 38 Demographic data

Demographics PIN:_____	
Gender:	7 Male; 3 Female
Age (years):	Range: 34-68 Average: 46.3
Time in HSI Analysis Work (years):	Range: 5 – 30 Average: 16.9
Job Title:	Human Factors Engineer (6) Suitability/Evaluator (2) Statistician/Evaluator (1) Research Psychologist (1)
Organizations:	DEVCOM Analysis Center Army Evaluation Center U.S. Army Aeromedical Research Laboratory Boeing

Participant SMEs were provided the survey tool, associated definition list, and mitigation list prior to meeting with the researcher. At a pre-determined time, SMEs met virtually over MS Teams with the researcher and received a concept briefing and review of the methods and results used to develop the proposed AS-TS tool. Following the concept briefing, SMEs were then presented scenarios 1 and 4 from Chapter IV of this research. Where scenario 1 was an anchor

scenario where mostly positive results were expected and scenario 4 provided useful training and familiarity to the participant but reported poor reliability of the proposed automation system. Scenarios 1 and 4 are described below.

Scenario 1 Description: You are a pilot in an aircraft with automated flight controls and obstacle detection/avoidance. Your automation has been proven to be highly reliable in obstacle detection and you have extensive training on the use of the system.

Automation Interaction: Obstacle detection is activated and soon after avoids an oncoming obscured obstacle by identifying it with an overlay and providing information on the proposed turn for avoidance and final clearance.

Scenario 4 Description: You are a pilot in an aircraft with automated flight controls and obstacle detection/avoidance. You are trained on the automation and are well aware of its history of poor performance.

Automation Interaction: Pilot begins flight, pilot activates the automation. Automation identifies multiple obstacles with only one in the flight path. Automation fails and requests the pilot to manually control the aircraft.

After reviewing the context presentation, use case description, and scenario video, SMEs completed the validated TIA survey (table 39) at the end of each scenario to familiarize themselves with how the survey could be used to collect TIA data under similar circumstances.

Table 39 Validated AS-TS survey

Trust Questions	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
I am confident in my ability to utilize the system.							
The system provides transparent information.							
I understand what the system is doing.							
I am familiar with the system operation.							
I have faith that the system will perform the intended task.							
The system provides appropriate feedback on current and future actions.							
The system effectively accomplishes its tasks.							
The system is suitable for carrying out the task.							

Following completion of both scenario reviews and surveys, SMEs were provided information related to data analysis techniques (e.g., descriptive and non-parametric statistics) and recommendations for use of the AS-TS as part of a battery of data collection tools (e.g., workload, situational awareness, usability, reliability, and objective performance data).

SMEs were then presented with a table of mitigation recommendations for each item of the AS-TS (table 40). SMEs were asked to review the mitigation recommendations for context related to TIA and relevancy to iterative system testing and development.

Table 40 Mitigation recommendations

Human Factors	Mitigation Recommendation
Confidence	Improve system training and knowledge of system limitations.
Transparency	Improve delivery of useful information to the user.
Understandability	Inform users on how and why the automation performs specific tasks to allow for accurate mental models of system processes.
Familiarity	Provide more training or opportunity to work with the system and historical context of use.
Automation Factors	**Mitigation Recommendation**
Feedback	Ensure the system provides appropriate user information and contextual future actions of the automation.
Effectiveness	Ensure the system can accomplish required mission tasks to standard.
Faith	Ensure automation consistently performs the intended actions.
Suitability	Ensure the system is being used for the appropriate tasks.

Following the mitigation review, a "usefulness survey" was provided to capture the perception of the usefulness of the TIA survey as a data collection tool (table 41). SME comments were collected related to the integration of the TIA survey into operational test procedures. The purpose of this study was to ensure the AS-TS can be used effectively in operational assessment and that the survey package provides data useful to decision-makers on automated system deficiencies and potential mitigations.

Table 41 AS-TS Usefulness questionnaire

AS-TS Usefulness Questionnaire PIN:_____ JOB:_____							
Questions	**Strongly Disagree**	**Disagree**	**Somewhat Disagree**	**Neither Agree nor Disagree**	**Somewhat Agree**	**Agree**	**Strongly Agree**
The AS-TS provides detailed definitions on TIA factors of interest.							
The AS-TS is a useful tool for evaluating subjective TIA.							
I would recommend using the AS-TS to collect subjective TIA data during Army testing.							
The recommended mitigations are appropriate for the TIA context.							
The recommended mitigations are consistent with OT issue mitigation strategies.							

Data collected from the questionnaire helps to ensure that the AS-TS is perceived as a useful tool among the test community and assist in developing a case for the acceptance of AS-TS as a formal data collection tool during Army Aviation test events.

Results

Results of the AS-TS from participants in study 3 completing scenarios 1 and 4 respectively, are found in figures 51 and 52.

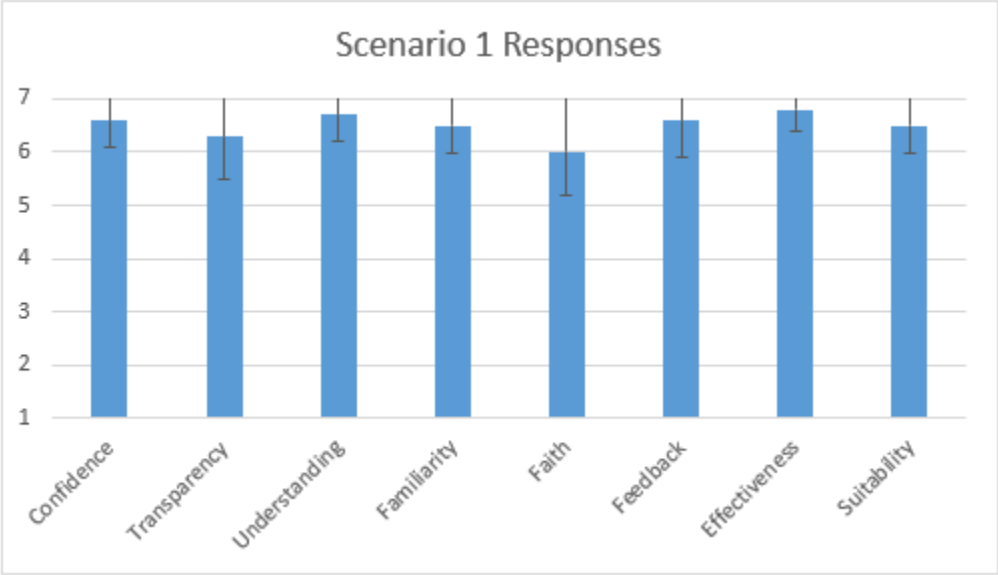

Figure 51 Scenario 1 responses

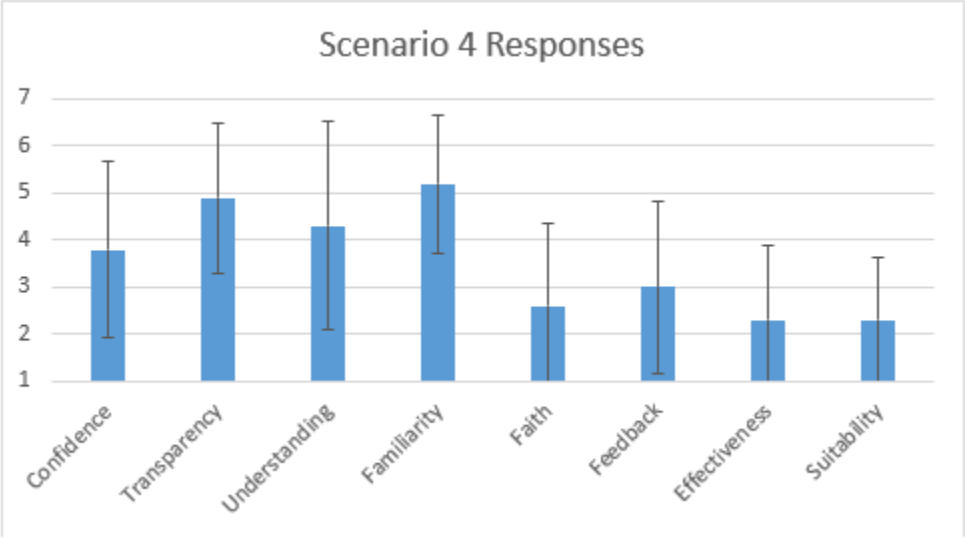

Figure 52 Scenario 4 responses

Positive ratings were associated with scenario 1 and mixed factor ratings were associated with scenario 4. These ratings are similar to the pilot ratings from study 2 with an interesting finding related to “confidence”. Study 3 participants averaged a 3.5 rating for confidence, while pilots averaged a 5.5 rating. While both sample sizes are small for statistical comparison, anecdotally, these ratings are consistent with previous pilot feedback that even in perceived poor trust scenarios, understanding of the system capabilities and limitations can improve user confidence in their interaction with the system. This finding might be different for non-pilots in more general trust scenarios. More research is required to understand user confidence in systems across different demographic representations.

After reviewing the mitigation strategies provided and discussing potential mitigations to improve scenario 4 outcomes, participants completed the usefulness questionnaire. Average participant ratings (n = 10) and standard deviations are provided in table 42.

Table 42 AS-TS Usefulness Questionnaire - results

AS-TS Usefulness Questionnaire PIN: _____ JOB: _____							
Questions	**Strongly Disagree**	**Disagree**	**Somewhat Disagree**	**Neither Agree nor Disagree**	**Somewhat Agree**	**Agree**	**Strongly Agree**
The AS-TS provides detailed definitions on TIA factors of interest.						**Avg 6.5 Std 0.53**	
The AS-TS is a useful tool for evaluating subjective TIA.						**Avg 6.6 Std 0.70**	
I would recommend using the AS-TS to collect subjective TIA data during Army testing.						**Avg 6.6 Std 0.70**	
The recommended mitigations are appropriate for the TIA context.						**Avg 6.3 Std 0.67**	
The recommended mitigations are consistent with OT issue mitigation strategies.						**Avg 6.5 Std 0.71**	

All ratings were positive for the usefulness questionnaire. Participants then provided additional comments to consider for general improvements to the AS-TS package (i.e., instructions for use, item definitions, survey, and mitigations). Participant comments and researcher responses are found in table 43.

Table 43 Participant comments and responses

Participant Comment	Researcher Response
I am interested in the results of this survey in operational testing compared to the TOAST (Trust of Automated System Technologies). I want to see this in action to see how we can use the data.	Intent is to use the AS-TS in parallel with TOAST and other subjective measures to investigate correlations and any divergent findings.
Split the feedback question into two questions. One for current actions and one for future actions. Add/Remove items as necessary depending on automation requirements for feedback.	Feedback question will be split to avoid confounds and instructions updated to adjust the feedback question, based on system requirements.
Is “Faith” a good word? Is it too personal?	There is some debate on “affective” wording in military contexts. Faith touches on system reliability subjectively and associates with confidence in the system. Depending on the advancement of AI, affective rapport may be desirable.
Evaluators are assessors and mitigation recommendations are not typically part of that process.	Comment was made to stress caution with adding mitigations into evaluations reports.
Recommendation to change the item: “The system provides transparent information.” to “The system provides transparent system information.”	Agreed - This change allows for a better description from confusion related to “feedback” information and focuses on transparent information of the specific system processes.
Recommendation to change the item: “The system provides feedback on actions.” to “The system provides feedback on system actions.”	Agreed - This change allows for a better description to focus on the system under review and not add confusion from other general indicators.
Recommend adding the word “user” in front of the confidence definition to indicate confidence within oneself rather than confidence of the automation actions.	Agreed – This change can help to set context for “user” confidence vs. “system” confidence.
Recommendation to change the item: “The system is suitable for carrying out the task.” to “The system is suitable for carrying out this type of task.”	Agreed - Clarifies that the system could be used for similar tasks under a use case, not only the observed task.

Discussion

In general, HFE experts and test evaluators rated the perceived usefulness of the AS-TS as worthy of consideration for Army Aviation OT type events. One primary recommendation was to break apart the item "The system provides appropriate feedback on current and future actions" to accommodate scenarios where the automation is not providing one or the other type of feedback with respect to current and future actions. In this case, it was recommended that the feedback item be measured separately for "current" actions and "future" actions to avoid confounding and ensure that pilot ratings appropriately reflect the system characteristics. Automated systems that do not project future actions can be assessed without the "future actions" feedback item. Table 44 shows the recommended AS-TS tool with the feedback item separated.

Table 44 Updated AS-TS tool

Trust Questions	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
I am confident in my ability to utilize the system.							
The system provides transparent system information.							
I understand what the system is doing.							
I am familiar with the system operation.							
I have faith that the system will perform the intended task.							
The system provides appropriate feedback on *current* system actions.							
The system provides appropriate feedback on *future* system actions.							
The system effectively accomplishes its tasks.							
The system is suitable for carrying out this type of task.							

The comment “Evaluators are assessors and mitigation recommendations are not typically part of that process.” is unique to the job position of Army test evaluators where objective data reporting is required, independent of mitigation or product improvement suggestions. Army evaluators report on the data outcomes and allow other processes to address potential solutions. These processes (i.e., crewstation working groups, design reviews, and after-action reviews) allow for decision-makers and stakeholders to review the objective data and make informed decisions on future improvements or mitigations to correct deficiencies identified in the system performance. The mitigation strategies in this study provide a starting point for discussion of mitigation strategies within the context of the product improvement processes,

Participants agreed that using the AS-TS as part of the established battery of tests currently used in OT is appropriate and that correlations among the additional surveys (e.g., usability, workload, and situational awareness) as well as objective data (e.g., reliability and performance) should be evaluated after OT events to provide a holistic approach to the assessment of trust for the system.

When analyzing the AS-TS subjective data, statistical analysis techniques such as the T-test and Mann-Whitney U can be used for comparative assessment of each scale item when investigating differences between systems or system characteristics. Descriptive statistics (i.e., mean, standard deviation) and visual display graphics can show the relationship among scale items for assessment of items that score lower or seem to be abnormal for the context of the automation use case. The AS-TS question structure is affirmative in nature, where generally, higher ratings correspond to more positive experiences related to the human-automation trust relationship.

The collected data may be able to pinpoint a specific automation feature or stimuli that scores consistently positively or negatively with participants. In these cases, it can be helpful to understand why the scores were provided. For example, in the case of negative perception scoring (i.e., strongly disagree that a product is good), negative participant comments and objective data (e.g., product failure) likely correlate to the issues that provoke negative attitudes. While a specific *action level* is not set on when to intervene for improvements to the product or system based solely on ratings, the individual item ratings and summative ratings taken within context, can help researchers evaluate the potential trust deficiencies within automation factors, contextual circumstances, and causal factors for the ratings. Additionally, interview is highly recommended as a practice to understand the context and thought processes for participant

ratings on both positive and negative ratings. Participants in study 3 agreed that surveys are the gateway to conversation between researchers and subjects. These interviews can greatly enhance the fidelity of the collected data by understanding the intent of the subject and their perception of a system.

The following research question and hypotheses for study 3 were accepted and support the research question, that the AS-TS is a useful tool for identifying TIA deficiencies and the recommended mitigation list is appropriate in context for addressing deficiencies identified using the AS-TS. “Can the survey instrument be used effectively in formal design testing to provide actionable information to data analysts (e.g., Human Factors Engineers) and product managers?”

a. Hypothesis 1: The developed survey will be a useful tool for analysts and program managers to identify TIA deficiencies, based on decision-maker and analyst ratings of effectiveness.

b. Hypothesis 2: The recommended actions list will provide appropriate courses of action to correct the deficiencies, based on decision-maker and analyst ratings of effectiveness.

Overall, participants reported high levels of interest and eagerness to use the AS-TS in evaluations. In the case of the two participants not related to aviation, interest was expressed in follow up data collection to examine the consistency of the AS-TS trust items across military domains (e.g., air and missile defense, long-range precision fires). The two participants rated the survey positively on all accounts and would like to investigate its use cases for non-aviation related systems.

Limitations

Several limitations are present within this research. While the number of participants used in construct validity calculations resulted in the minimum requirements for statistical analysis, a larger sample size would provide more confidence in the determination of scale items.

Care was taken to investigate each scale item to ensure appropriate factor placement and removal/inclusion of each item was critically analyzed.

The scenarios, while representative of an aircraft cockpit and mission set, were not conducted in an actual aircraft with significant consequences. Placing pilots in potentially dangerous situations for research is unacceptable and simulation was required to investigate poorly performing automation. During operational testing, pilots will be using actual aircraft with automated systems. As data are collected, further Confirmatory Factor Analysis (CFA) can be conducted on the survey to ensure that the included items are appropriate for the actual environment.

The sensitivity of the AS-TS has yet to be determined in actual testing environments. In general, the large distribution of responses and "extremes" of the automation performance (e.g., very good, very poor) across the tested scenarios may indicate a potential for reduced sensitivity of the AS-TS when applied to automation instances with minor changes in performance or outcomes among the tested alternatives. The AS-TS is targeted at specific systems and tasks in a manner that when comparative analysis is conducted, ideally, the AS-TS will be sensitive enough to compare one system to another with respect to the trust construct. However, more data collection and analysis opportunities will be required to ensure that the AS-TS is appropriate for less extreme cases of automation differences through comparative testing in representative contexts. Additionally, larger operational questions (i.e., "Do you trust the aircraft in this mission?") have not been investigated and will need supportive data during operational testing to determine whether the AS-TS is appropriate for contexts beyond system related task performance.

Finally, the AS-TS is only a tool for providing subjective information on pilots trust relationship to the automated system under test. Many other factors are required to analyze the human-machine interface, including objective measurements of use, reliability, eye tracking, physiological monitoring and subjective measurements of workload, situational awareness, and usability. The sum of the data collected can help to illustrate the relationship between the pilot and the aircraft. Each metric works together to provide information to researchers and analysts for evaluation of the interactions.

Future Work

The completed research and methodology for the AS-TS will help to assess perceived trust in Army Aviation systems and allow for researcher to assess and improve the human-machine interface to optimize trust-based interactions.

This research has been submitted as two journal articles for peer review and potential publication. The first submission focused on the literature review and identification of the factors that influence TIA, including the decision-making process and method of the AHP and its use in determining key factors for the TIA survey. The second submission focused on the survey validation methods and outcomes of the factor analysis and results for the finalized survey rating scale and assessment method.

U.S. Army organizations and personnel have been in contact with the researcher to consider use cases of the AS-TS during operational test events and simulations to determine efficacy of the survey and compare analysis to current trust data collection methods. The U.S. Army Aeromedical Research Laboratory will use the AS-TS as part of a battery of assessment tools under critical review for research consideration of trust measurement for aviators.

Additionally, there is interest from the Army evaluation community to determine the generalizability of the survey across Army domains (e.g., air and missile defense, artillery). Investigation will continue into the generalizability or use of the research methodology to develop improved TIA surveys for other Army demographics.

Finally, the dissertation document will be sent to interested Army organizations and contractors for consideration in further testing for generalizable use to other Army domains and to investigate the possibility of repeating the methodology in this research to determine significant trust factors for different Army demographic representation.

Conclusion

A TIA rating scale for Army Aviation was developed using an analytic approach over three studies. Study 1 used aviation SMEs to examine and rate key factors and potential survey item definitions found in TIA literature to establish face validity for a notional TIA survey. By using the AHP decision-making process, significant items were identified and included on the notional survey consisting of two factors and fifteen items for use in further construct validity testing using the Army Aviation pilot demographic. Thirty-two (32) Army Aviation participants were used during study 2 to evaluate the construct validity of the survey over four representative aircraft automation scenarios. Following data collection, exploratory factor analysis and reliability testing were used to reduce the survey items and ensure validity. The validated survey was then explained and provided to ten research and evaluation experts for review during study 3. The experts rated the survey positively using a usefulness scale and provided comments to clarify questions and improve the data collection methodology.

Participant survey results and comments confirmed that the AS-TS is an appropriate tool for TIA evaluation to assist analysts and project managers in data collection analysis and

decision making. Additionally, participants confirmed that the mitigation recommendations were appropriate for use and can provide initial starting points for issue correction and management.

After evaluator input during study 3, the final TIA survey was completed with instructions, definitions, and mitigations for addressing lower scoring items when analyzing data collected during testing. When using the AS-TS researchers and evaluators should refer to Appendix C for the instructions, definitions, survey, and mitigation strategies as part of the survey methodology.

CHAPTER VI

RESEARCH SUMMARY

Trust in Automation is considered one of the primary challenges for successful integration of automation, AI, and humans (Beer, Fisk, & Rogers, 2014). Measuring TIA is very difficult due to its multi-faceted nature (Brzowski & Nathan-Roberts, 2019). In general, the measurement of TIA is often associated with the subjective user perception of trust between the user and the automated system (Parasuraman & Riley, 1997).

There are numerous factors that can influence the TIA relationship. Factors ranging from human perception of automation characteristics, system performance, mental model expectations, personality, cultural influences, and mental workload have all been found to play a potential role in TIA (Brzowski & Nathan-Roberts, 2019; Hoff & Bashir, 2015).

The U.S. Army Future Vertical Lift program is an advanced aircraft development program that will consist of both long rage air assault and reconnaissance aircraft. These systems will need robust analysis tools to analyze the human-machine interface and variety of automated system interactions that take place on the aircraft. U.S. Army rotorcraft (helicopter) operations are particularly complex when considering the mission demands and environmental conditions where these operations occur. Helicopter pilots often operate in lower altitudes, complex terrain, hostile environments, and within DVE (Helfrich, 2020). These unique circumstances drive the need for robust human-automation designs that enable the pilots to effectively utilize the aircraft systems for mission accomplishment.

As autonomy continues to play a major role in aircraft system use and development, users must trust that the automation is performing to standard. There is currently no standard methodology that is in use for the Army to assess TIA for pilots as a holistic measurement that identifies trust deficiencies and the relationship of trust to user reliance with follow-up actions. The purpose of this research was to produce a measurement tool and assessment methodology for TIA assessment using an Army Aviation helicopter system use case that can help to identify *trust deficiencies*.

Trust Factor Identification

In order to identify the key factors that influence TIA for Army Aviation systems, a comprehensive literature review was conducted to establish a historical background of TIA research, identify key concepts in human and automation trust, and to develop an initial list of factors that could potentially impact pilot TIA. One hundred and sixteen (116) articles were reviewed and the factors that influenced TIA were categorized under key terms as defined by TIA literature. A frequency analysis was used to categorize the factors under *human* and *automation* factors that influence TIA, based on either individual characteristics of the user or situational characteristics of the automation. The frequency analysis found that factors such as automation reliability and user confidence in the system were prevalent throughout the literature. By identifying prominent factors, an initial pool of items can be established for development of the Aviation Systems – Trust Survey (AS-TS).

Trust Survey Development

The factor identification and associated definitions were evaluated by SMEs to conduct survey face validity testing in order to establish the initial AS-TS survey for the Army Aviation demographic.

Six SMEs (3 Human Factors Researchers/Engineers and 3 U.S. Army Pilots) were recruited to evaluate the identified TIA factors through pairwise comparison using the AHP decision-making methodology. Additional SME comments were collected during the data collection about their perception of the TIA factors and any additional TIA factors that should be included. SMEs agreed that the factor list was inclusive, and no significant comments were captured related to additional factor requirements.

The AHP was used effectively to identify critical TIA factors for Army pilots and establish initial face validity of a TIA survey. Additionally, four TIA factors (i.e., Purpose and Intent, Risk, Demographics, and Personal Attachment) were removed from consideration of the notional AS-TS due to low rankings and SME comments related to the factors.

In a follow on study, data were collected from 32 pilot participants to validate the notional TIA survey scale items and their subscale reliability. Pilots were interviewed and surveyed after receiving an introductory briefing and participating in four virtual and scripted scenarios. At the conclusion of the data collection, PCA and EFA were conducted on the data to identify the total number of factors and eliminate poorly performing scale items. Eight commonly loaded scale items were retained after the EFA. Four under human factors: Confidence, Transparency, Understanding, and Familiarity and four under automation factors: Faith, Effectiveness, Suitability, and Integrity. Each item that was removed was investigated further to ensure appropriate removal and adequate reasoning. An additional item (integrity) was

removed under automation factors due to its close relationship in definition and results to the effectiveness item. Feedback was added as a scale item under automation factors as it cleanly loaded on two of the three scenarios and only mildly under the human factor scale for one scenario. Feedback is considered a valuable scale item, that does not significantly overlap with the other items, useful for calibrating and determining pilot mental models of systems and their trust relationship.

The validated survey was then explained and provided to ten research and evaluation experts for review during study 3. The experts rated the survey positively using a usefulness scale and provided comments to clarify questions and improve the data collection methodology.

Participant survey results and comments confirmed that the AS-TS is an appropriate tool for TIA evaluation to assist analysts and project managers in data collection analysis and decision making. Additionally, participants confirmed that the mitigation recommendations were appropriate for use and can provide initial starting points for issue correction and management.

After evaluator input during study 3, the final TIA survey was completed with instructions, definitions, and mitigations for addressing lower scoring items when analyzing data collected during testing. When using the AS-TS researchers and evaluators should refer to Appendix C for the instructions, definitions, survey, and mitigation strategies as part of the survey methodology.

General Conclusions

The robust literature review and studies for development of the AS-TS were successful in establishing a validated survey for use during Army Aviation testing. Statistical metrics were met for appropriate survey development, SMEs were used to ensure validity, and Army analysts were interviewed to verify use cases and appropriateness of the survey for testing.

The AS-TS will be used in future Army Aviation testing and critically evaluated to ensure appropriate measurement through follow-on testing and further statistical analysis. The methodology used to develop the AS-TS is repeatable and can be applied to other survey developments or replicated for other demographics to identify and evaluate trust factors important to the target audience.

The AS-TS survey and evaluation methodology are poised to be a significant tool in the evaluation of advanced aircraft design. Along with other subjective and objective data collection, the AS-TS can help analysts evaluate the human-machine interface and ensure that automation implementations provide optimal support to pilots for mission accomplishment.

REFERENCES

Alicia, T. J., Hall, B. T., & Terman, M. (2020). *Synergistic Unmanned Manned Intelligent Teaming (SUMIT) Final Report*. (Publication No. FCDD-AMT-20-09). U.S. Army Combat Capabilities Development Command.

Air Force Technology. (2021, September 1). CH-47D/F / MH-47E Chinook Transport Helicopter. Airforce Technology. Retrieved November 16, 2022, from https://www.airforce-technology.com/projects/ch47d-chinook-helicopter/

Antonides J. R. (2014) *Air Mission Commanders.* Army Aviation Magazine. http://www.armyaviationmagazine.com/index.php/archive/not-so-current/1112-air-mission-commanders

Balfe, N., Sharples, S., & Wilson, J. R. (2018). Understanding is key: An analysis of factors pertaining to trust in a real-world automation system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *60*(4), 477–495. https://doi.org/10.1177/0018720818761256

Bargas-Avila, J. A., & Bruhlmann, F. (2016). Measuring user rated language quality: Development and validation of the user interface Language Quality Survey (LQS). *International Journal Human-Computer Studies. 86,* 1-10. http://dx.doi.org/10.1016/j.ijhcs.2015.08.010

Bartlett, M. S. (1951). The effect of standardization on a Chi-square approximation in factor analysis. *Biometrika, 38,* 337-344.

Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a framework for levels of robot autonomy in human-robot interaction. Journal of human-robot interaction, *3*(2), 74–99. https://doi.org/10.5898/JHRI.3.2.Beer

Bolarinwa, O. A. (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science research. *Niger Postgraduate Medical Journal, 22,* 195-201.

Borum, R. (2010). The Science of Interpersonal Trust. *Mental Health Law & Policy Faculty Publications*. 574. https://digitalcommons.usf.edu/mhlp_facpub/574

Britannica, Editors of Encyclopedia (2021, July 21). *Industrial Revolution*. *Encyclopedia Britannica*. https://www.britannica.com/event/Industrial-Revolution

Brzowski, M., & Nathan-Roberts, D. (2019). Trust Measurement in Human-Automation Interaction: A Systematic Review. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 1595-1599, USA. https://doi.org/10.1177%2F1071181319631462

Cafarelli, D. A., & Hansman, R. J. (1998). (rep.). *Effect of False Alarm Rate on Pilot Use and Trust of Automation Under Conditions of Simulated High Risk* (pp. 1–62). Cambridge, MA: Massachusetts Institute of Technology.

Cassidy, A. (2009). Mental Models, Trust, and Reliance: Exploring the Effect of Human Perceptions on Automation Use.

Castaldo, S. (2007). *Trust in market relationships.* Edward Eigar Publishing: UK.

Chadwick, M. (2022, June 28). Sikorsky Black Hawk Helicopter. Lockheed Martin. Retrieved November 16, 2022, from https://www.lockheedmartin.com/en-us/products/sikorsky-black-hawk-helicopter.html

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. (2016). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(3), 333–345. https://doi.org/10.1177/0018720816682648

Charalambous, G., Fletcher, S., & Webb, P. (2015). The development of a scale to evaluate trust in industrial human-robot collaboration. *International Journal of Social Robotics*, *8*(2), 193–209. https://doi.org/10.1007/s12369-015-0333-8

Chien, S.-Y., Lewis, M., Sycara, K., Liu, J.-S., & Kumru, A. (2016). Influence of cultural factors in Dynamic Trust in automation. *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. https://doi.org/10.1109/smc.2016.7844677

Chien, S.-Y., Semnani-Azad, Z., Lewis, M., & Sycara, K. (2014). Towards the development of an inter-cultural scale to measure trust in automation. *Cross-Cultural Design*, 35–46. https://doi.org/10.1007/978-3-319-07308-8_4

Christophersen, T., & Kondradt, U. (2012). Development and validation of a formative and reflective measure for the assessment of online store usability. *Behaviour & Information Technology, 31*(9), 839-857. https://doi.org/10.1080/0144929X.2010.529165

Chyung, S. Y., Winiecki, D. J., Hunt, G., & Sevier, C. M. (2017). Measuring Learners' Attitudes Toward Team Projects: Scale Development Through Exploratory and Confirmatory Factor Analyses. *American Journal of Engineering Education, 8*(2). 61-82.

Converse, J. M., & Presser, S. (1986). Survey questions: Handcrafting the standardized questionnaire. Sage Publications, Inc.

Colquitt, J., Scott, B., & LePine, J. (2007). Trust, Trustworthiness, and Trust Propensity: A Meta-Analytic Test of Their Unique Relationships With Risk Taking and Job Performance. *The Journal of Applied Psychology*. *92*. 909-27. https://doi.org/10.1037/0021-9010.92.4.909.

Cronbach's Alpha: Simple definition, use and Interpretation. Statistics How To. (2021, July 2). https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/cronbachs-alpha-spss/.

Danner, G. E. (2019). The executive's how-to guide to automation. Cham: Springer

de Winter, J. C., Dodou, D., & Wieringa, P. A. (2009). Exploratory Factor Analysis With Small Sample Sizes. *Multivariate behavioral research, 44*(2), 147–181. https://doi.org/10.1080/00273170902794206

Department of Defense (DoD). (2019). Test & Evaluation Management Guide: August 2016. Independent Publication.

Desselle, S. P. (2005). Construction, implementation, and analysis of Summated Rating Attitude Scales. *American Journal of Pharmaceutical Education, 69*(5), 97. https://doi.org/10.5688/aj690597

Dolgov, I., & Kaltenbach, E. K. (2017). Trust in Automation Inventories: An Investigation and Comparison of the Human-Computer Trust and Trust in Automated Systems Scales. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*(1), 1271–1275. https://doi.org/10.1177/1541931213601799

Dordick, H. S. (1965). (rep.). *An Introduction to System Effectiveness* (pp. 1–9). Santa Monica, CA: Rand Corp.

Duez, P. P., Zuliani, M. J., & Jamieson, G. A. (2006). Trust by Design: Information Requirements for Appropriate Trust in Automation. *Proceedings of the 2006 conference of the Centre for Advanced Studies on Collaborative Research, Canada.* http://dx.doi.org/10.1145/1188966.1188978

Endsley, Mica. (2015). Autonomous Horizons: System Autonomy in the Air Force – A Path to the Future. Volume I: Human-Autonomy Teaming. https://doi.org/10.13140/RG.2.1.1164.2003

Fallon, C. K., Murphy, A. K. G., Zimmerman, L., & Mueller, S. T. (2010). The calibration of trust in an automated system: A sensemaking process, *International Symposium on Collaborative Technologies and Systems*, 390-395, https://doi.org/10.1109/CTS.2010.5478488

Freedberg, S. J. (2020). *FVL: Robotic Co-Pilots Will Help Fly Black Hawks in 2021*. Breaking Defense. https://breakingdefense.com/2020/11/fvl-robotic-co-pilots-will-help-fly-black-hawks-in-2021/

Ferketich, S., Phillips, L., & Verran, J. (1993). Development and administration of a survey instrument for cross-cultural research. *Research in Nursing & Health*, *16*(3), 227–230. https://doi.org/10.1002/nur.4770160310

Frey, B. (2018). Scree Plot. *The SAGE encyclopedia of educational research, measurement, and evaluation. 1-4.* Thousand Oaks, CA: Sage Publications, Inc.

Garg, R., Rahman, Z., Qureshi, M. N., & Kumar, I. (2012). Identifying and ranking critical success factors of customer experience in banks: An analytic hierarchy process (AHP) approach. Journal of Modelling in Management, 7(2), 201-220. http://dx.doi.org/10.1108/17465661211242813

Ginty, A. T. (2013). Construct Validity. In: Gellman M.D., Turner J. R. (eds) Encyclopedia of Behavioral Medicine. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-1005-9_851

Gillespie, B. J., Ruel, E., & Wagner III, W. E. (2015). The Practice of Survey Research: Theory and Applications. Sage Publications, Inc.

Ji Gao, & Lee, J. D. (2006). Extending the decision field theory to model operators' reliance on automation in supervisory control situations. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 36(5), 943–959. https://doi.org/10.1109/tsmca.2005.855783

Hair, Joseph F. Jr., Anderson, Rolph E., Tatham, Ronald L., & Black, William C. (1998). Multivariate Data Analysis, 5th ed. Upper Saddle River, NJ: Prentice Hall

Hardin, R. (2006). Trust. Cambridge, UK: Polity

Helfrich, E. (2020, March 12). Eyes up and out: Advancing situational awareness in helicopter avionics. Military Embedded Systems. https://militaryembedded.com/avionics/computers/eyes-up-and-out-advancing-situational-awareness-in-helicopter-avionics.

Henshel, D., Cains, M. G., Hoffman, B., & Kelley, T. (2015). Trust as a Human Factor in Holistic Cyber Security Risk Assessment. Procedia Manufacturing, 3, 1117-1124. https://doi.org/10.1016.j.promfg.2015.07.186

Hermann, K. J., Bager-Elsborg, A., & Parpala, A. (2017). Measuring perceptions of the learning environment and approaches to learning: validation of the learn questionnaire. *Scandinavian Journal of Educational Research, 61*(5), 526-539. http://dx.doi.org/10.1080/00313831.2016.1172497

Higham, T. M., Vu, K.-P. L., Miles, J., Strybel, T. Z., & Battiste, V. (2013). Training Air Traffic Controller Trust in automation within a nextgen environment. *Human Interface and the Management of Information. Information and Interaction for Health, Safety, Mobility and Complex Environments*, 76–84. https://doi.org/10.1007/978-3-642-39215-3_9

Ho, N., Sadler, G. G., Hoffmann, L. C., Zemlicka, K., Lyons, J., Fergueson, W., Richardson, C., Cacanindin, A., Cals, S., & Wilkins, M. (2017). A longitudinal field study of Auto-GCAS Acceptance and Trust: First-year results and implications. *Journal of Cognitive Engineering and Decision Making, 11*(3), 239–251. https://doi.org/10.1177/1555343417701019

Hoff, K., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors, 57*(3), 407-434. http://dx.doi.org/10.1177/0018720814547570

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in Automation. *IEEE Intelligent Systems, 28*(1), 84–88. https://doi.org/10.1109/MIS.2013.24

Hsiao-Ying, H. & Masooda, B. (2017). Personal Influences on Dynamic Trust Formation in Human-Agent Interaction. *In Proceedings of the 5th International Conference on Human Agent Interaction (HAI '17)*. Association for Computing Machinery, New York, NY, USA, 233–243. https://doi.org/10.1145/3125739.3125749

Institute for Defense Analyses (IDA). (2018). Evaluating Human-System Interaction. Presentation: 30 May 2018, Alexandria, VA.

Jensen, R. S. (1995). Pilot Judgement and Crew Resource Management. Avebury Aviation: University of Michigan.

Jeong, H., Park, J., Park, J., Pham, T., & Lee, B. C. (2018). Analysis of trust in Automation Survey instruments using semantic network analysis. *Advances in Intelligent Systems and Computing*, 9–18. https://doi.org/10.1007/978-3-319-94334-3_2

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Engineering, 4*(1), 53-71. https://doi.org/10.1207/S15327566IJCE0401_04

Jiang, X., Khasawneh, M. T., Master, R., Bowling, S. R., Gramopadhye, A. K., Melloy, B. J., & Grimes, L. (2004). Measurement of Human Trust in a hybrid inspection system based on signal detection theory measures. *International Journal of Industrial Ergonomics*, *34*(5), 407–419. https://doi.org/10.1016/j.ergon.2004.05.003

Johns Hopkins University (2019). *Trust in Automation Calibration Insights*. Integrated Adaptive Cyber Defense - Applied Physics Laboratory LLC.

Johnson, J. D., Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Type of automation failure: The effects on trust and reliance in automation. *PsycEXTRA Dataset*. https://doi.org/10.1037/e577212012-007

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology, 7*(4), 396–403. https://doi.org/10.9734/bjast/2015/14975

Kaltenbach, E. & Dolgov, I. (2017). On the Dual Nature of Transparency and Reliability: Rethinking Factors that Shape Trust in Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*(1), 308-312. https://doi.org/10.1177/1541931213601558

Kline, P. (1986). A handbook of test construction: Introduction to psychometric design. London: Methuen.

Korber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics, 66*, 18-31. http://dx.doi.org/10.1016/j.apergo.2017.07.006

Kraus, J. (2020). *Psychological Processes in the Formation and Calibration of Trust in Automation* [Doctoral dissertation, Ulm University - Institute for Psychology and Education].

Leadens, R. (2020). Pilot perception of automation use: A generational assessment (3106) [Thesis, University of North Dakota]. https://commons.und.edu/theses/3106

Lecerof, A., & Paterno, F. (1998). Automatic support for usability evaluation. *IEEE Transactions on Software Engineering*, *24*(10), 863–888. https://doi.org/10.1109/32.729686

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lim, S. & Jahng, S. (2019). Determining the Number of Factors Using Parallel Analysis and Its Recent Variants. *Psychological Methods*, *24*(4). https://doi.org/10.1038/met0000230

Litwin, M. S. (2003). How to assess and interpret survey psychometrics (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Stensen K & Lydersen S. (2022) Internal consistency: from alpha to omega? *Tidsskr Nor Laegeforen. 142*(12). https://doi.org/10.4045/tidsskr.22.0112.Lyons, J. B., Ho, N. T., Van Abel, A. L., Hoffmann, L. C., Sadler, G. G., Fergueson, W. E., Grigsby, M. A., & Wilkins, M. (2017). Comparing trust in auto-GCAS between experienced and Novice Air Force pilots. *Ergonomics in Design: The Quarterly of Human Factors Applications*, *25*(4), 4–9. https://doi.org/10.1177/1064804617716612

Lyons, J. B., Koltai, K. S., Ho, N. T., Johnson, W. B., Smith, D. E., & Shively, R. J. (2016). Engineering trust in complex automated systems. Ergonomics in Design: The Quarterly of Human Factors Applications, 24(1), 13–17. https://doi.org/10.1177/1064804615611272

Madhavan, P., & Wiegmann, D. A. (2004). A New Look at the Dynamics of Human-Automation Trust: Is Trust in Humans Comparable to Trust in Machines? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 48*(3), 581–585. https://doi.org/10.1177/154193120404800365

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in Automated Aids. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 48*(2), 241–256. https://doi.org/10.1518/001872006777724408

Mann-Whitney U test using SPSS Statistics. Mann-Whitney U Test in SPSS Statistics | Setup, Procedure & Interpretation | Laerd Statistics. (2021). https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php.

Marsh, S. & Dibben, M. (2005). The Role of Trust in Information Science and Technology. *ARIST, 37,* 465-498. https://doi.org/10.1002/aris.1440370111.

Mason, J., Classen, S., Wersal, J., & Sisiopiku, V. (2021). Construct Validity and Test-Retest Reliability of the Automated Vehicle User Perception Survey. *Frontiers in Psychology, 25*. https://doi.org/10.3389/fpsyg.2021.626791

Mason, R. L., Gunst, R. F., & Hess, J. C. (1989). *Statistical design and analysis of experiments*. New York: NY: Wiley.

Mayfield, M. (2021, April, 2) *Army Powering Through with Future Vertical Lift Programs*. National Defense Magazine. https://www.nationaldefensemagazine.org/articles/2021/4/2/army-powering-through-with-future-vertical-lift-programs

Mayer, R., Davis, J., & Schoorman, F. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review, 20(*3), 709-734. doi:10.2307/258792

McKnight, D. H., & Chervany, N. L. (1996). *The meanings of trust*. Minneapolis, Minn: Carlson School of Management, Univ. of Minnesota.

McNamara, J., Olfert, M. D., Sowers, M., Colby, S., White, A., Byrd-Bredbenner, C., Kattelmann, K., Franzen-Castle, L. D., Brown, O., Kidd, T., Shelnutt, K. P., Horacek, T., & Greene, G. W. (2020). Development of an Instrument Measuring Perceived Environmental Healthfulness: Behavior Environment Perception Survey (BEPS). *Journal of Nutrition Education and Behavior*, *52*(2), 152-161. https://doi.org/10.1016/j.jneb.2019.09.003

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2012). I trust it, but I don't know why. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *55*(3), 520–534. https://doi.org/10.1177/0018720812465081

Merritt, S. M. (2011). Affective Processes in Human-Automation Interactions. *Human Factors, 53*(4), 356-370. http://dx.doi.org/10.1177/0018720811411912

Miller, D. J. E., & Perkins, L. (2010). Development of Metrics for Trust in Automation (p. 18). AIR FORCE RESEARCH LAB WRIGHT-PATTERSON AFB OH SENSORS DIRECTORATE. https://apps.dtic.mil/docs/citations/ADA525259

Miramontes, A., Tesoro, A., Trujillo, Y., Barraza, E., Keeler, J., Boudreau, A., Strybel, T., & Vu, K-P. (2015). Training Student Air Traffic Controllers to Trust Automation. *Procedia Manufacturing, 3,* 3005-3010. https://doi.org/10.1016/j.promfg.2015.07.844

Mishler, S., Chen, J., Sabic, E., Hu, B., Li, N., & Proctor, R. W. (2017). Description-experience gap: The role of feedback and description in Human Trust in Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*(1). https://doi.org/10.1177/1541931213601559

Murray, J. (2013). Likert Data: What to Use, Parametric or Non-Parametric? *International Journal of Business and Social Science, 21*.

Nasaireh, M. A. (2020). Developing and Validating Instruments for Measurement of Organizational Culture Dimensions for Organizational Development Achievement. *International Journal of Multidisciplinary and Current Educational Research*, *2*(5), 168-174.

Natarajan, M. & Gombolay, M. (2020). Effects of anthropomorphism and accountability on trust in human robot interaction. *ACM/IEEE International Conference on Human-Robot Interaction*, 33-42. https://doi.org/10.1145/3319502.3374839

Niu, J., Geng, H., Zhang, Y., & Du, X. (2018). Relationship between Automation Trust and operator performance for the Novice and expert in spacecraft rendezvous and docking (RVD). *Applied Ergonomics*, *71*, 1–8. https://doi.org/10.1016/j.apergo.2018.03.014

Nourani, M., King, J., & Ragan, E. (2020). The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 8*(1), 112-121. https://ojs.aaai.org/index.php/HCOMP/article/view/7469

O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement in e-commerce environments. Journal of the American Society for Information Science and Technology, 61(1), 50-69. https://doi.org/10.1002/asi.21229

Okamura, K. & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLoS ONE, 15*(2): e0229132. https://doi.org/10.1371/journal.pone.0229132

Pace, D. K. & Sheehan, J. (2002). Subject Matter Expert (SME)/Peer Use in M&S V&V. Foundations for V&V in the 21st Century Workshop. Johns Hopkins University Applied Physics Laboratory, October 22-24, 2002.

Pak, R., Rovira, E., McLaughlin, A. C., & Baldwin, N. (2016). Does the domain of technology impact user trust? Investigating trust in automation across different consumer-oriented domains in young adults, military, and older adults. Theoretical Issues in Ergonomics Science, 18(3), 199–220. https://doi.org/10.1080/1463922x.2016.1175523

Pamplona, D. A., & Alves, C. J. (2020). Does a fighter pilot live in the danger zone? A risk assessment applied to military aviation. Transportation Research Interdisciplinary Perspectives, 5, 100114. https://doi.org/10.1016/j.trip.2020.100114

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality Automated Systems. *Communications of the ACM*, *47*(4), 51–55. https://doi.org/10.1145/975817.975844

Parasuraman, R. & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors, 39*(2), 230-253. https://doi.org/10.1518/001872097778543886

Paxion J, Galy E, & Berthelon C. (2014). Mental workload and driving. *Frontiers in Psychology, 2*(5). https://doi.org/10.3389/fpsyg.2014.01344.

Perkins, L. A., Miller, J. E., Hashemi, A., & Burns, G. (2010). Designing for human-centered systems: Situational Risk as a factor of trust in automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(25), 2130–2134. https://doi.org/10.1177/154193121005402502

Polit, D., & Beck, C. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 47*, 264-276.

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). (2021). http://www.prisma-statement.org/

Reim, G. (2021, October 14). Boeing expects US Army order for at least 100 AH-64E apaches. Flight Global. Retrieved November 16, 2022, from https://www.flightglobal.com/helicopters/boeing-expects-us-army-order-for-at-least-100-ah-64e-apaches/145910.article

Rice, S. (2009). Examining single- and multiple-process theories of trust in automation. *The Journal of General Psychology*, *136*(3), 303–322. https://doi.org/10.3200/genp.136.3.303-322

Rotenberg, K., Fox, C., Green, S., Ruderman, L., Slater, K., Stevens, K., & Carlo, G. (2005). Construction and validation of a children's interpersonal trust belief scale. *British Journal of Developmental Psychology. 23*. 271 - 293. https://doi.org/10.1348/026151005X26192

Rudiger, J., Wagner, A., & Badreddin, E. (2007). Behavior based description of dependability - defining a minium set of attributes for a behavioral description of dependability. *Proceedings of the Fourth International Conference on Informatics in Control, Automation and Robotics*. https://doi.org/10.5220/0001650203410346

Saaty, T. L. (1994). How to Make a Decision: The Analytic Hierarchy Process. *Interfaces, 24*, 19-43. https://doi.org/10.1287/inte.24.6.19

Saaty, T. L. (1980). The Analytic Hierarchy Process. McGraw-Hill, New York.

Salminen, J., Santos, J. M., Kwak, H., An, J., Juny, S., & Jansen, B. J. (2020). Persona Perception Scale: Development and Exploratory Validation of an Instrument for Evaluating Individuals' Perceptions of Personas. *International Journal of Human-Computer Studies, 141*. https://doi.org/10.1016/j.ijhcs.2020.102437

Samuels, P. (2017) Advice on Reliability Analysis with Small Samples - Revised Version. Technical Report. ResearchGate, Birmingham, UK.

Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2011). Understanding reliance on automation: Effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science, 15*(2), 134–160. https://doi.org/10.1080/1463922x.2011.611269

Schaefer, K. (2013). *The Perception and Measurement of Human-robot Trust*. [Doctoral dissertation, University of Central Florida]. http://stars.library.ucf.edu/etd/2688

Schaefer, K., Chen, J., Szalma, J., & Hancock, P. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. *58*. https://doi.org/10.1177/0018720816634228.

Scholl, K., & Hanson, R. (2020). Testing the automation revolution hypothesis. Economics Letters, 193, 109287. https://doi.org/10.1016/j.econlet.2020.109287

Shahrdar, S., Menezes, L., & Nojoumian, M. (2018). A survey on trust in autonomous systems. Advances in Intelligent Systems and Computing, 368–386. https://doi.org/10.1007/978-3-030-01177-2_27

Sheridan, T. B. (2019). Individual differences in attributes of trust in automation: Measurement and application to system design. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01117

Siau, K. & Wang, W. (2018). Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technology Journal, 31*. 47-53.

Simon, A. (2020). Usability of electronic patient record systems: Instrument validation study conducted for hospitals in Germany. *Health Informatics Journal, 26*(3), 1969-1982. https://doi.org/10.1177%2F1460458219895910

SKYbrary (2021). *Cockpit Automation - Advantages and Safety Challenges - SKYbrary Aviation Safety*. SKYbrary. Retrieved October 5, 2021, from https://www.skybrary.aero/index.php/Cockpit_Automation_-_Advantages_and_Safety_Challenges#Automation_Dependency

Smith, M. A., Allaham, M. M., & Wiese, E. (2016). Trust in automated agents is modulated by the combined influence of agent and Task Type. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 206–210. https://doi.org/10.1177/1541931213601046

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an Empirically Developed Scale for System Trust: Take Two. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 52*(19). 1335-1339. https://doi.org/10.1177/154193120805201907

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research. Instruments in Science Education. *Research in Science Education, 48, 1273-1296.*

Taherdoost, H. (2017). Decision Making Using the Analytic Hierarchy Process (AHP); A Step by Step Approach. *International Journal of Economics and Management Systems*, *2*.

Tavakol, M. & Dennick, R. (2011) Making sense of Cronbach's alpha. *Internation Journal of Medical Education,* 27(2), 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Taylor, G., & Turpin, T. (2015). Army Aviation Manned-Unmanned Teaming (MUM-T): Past, Present, and Future. *18th International Symposium on Aviation Psychology*, 560-565. https://corescholar.libraries.wright.edu/isap_2015/12

Uggirala, A., Gramopadhye, A. K., Melloy, B. J., & Toler, J. E. (2004). Measurement of trust in complex and dynamic systems using a quantitative approach. *International Journal of Industrial Ergonomics*, *34*(3), 175-186. https://doi.org/10.1016/j.ergon.2004.03.005

U. S. Federal Aviation Administration. (1991). Aeronautical Decision Making (Advisory Circular 60-22). Washington, DC.

van Dongen, K., & van Maanen, P.-P. (2013). A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies*, *71*(4), 410–424. https://doi.org/10.1016/j.ijhcs.2012.10.018

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*(3), 281–291. https://doi.org/10.1177/0018720809338842

Wang, L., Jamieson, G. A., & Hollands, J. G. (2008). Selecting Methods for the Analysis of Reliance on Automation. *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting, USA.* https://doi.org/10.1177%2F154193120805200419

Warmbrod, J. R. (2014). Reporting and interpreting scores derived from likert-type scales. *Journal of Agricultural Education, 55*(5), 30-47. https://doi.org/10.5032/jae.2014.05030

Washington, M. G., (2013). Trust and Project Performance: The Effects of Cognitive-Based and Affective Based Trust on Client-Project Manager Engagements. [Master of Science in Organizational Dynamics Theses]. University of Pennsylvania. https://repository.upenn.edu/od_theses_msod/67

Westen, D. & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology, 84*(3), 608-618. http://dx.doi.org/10.1037/0022-3514.84.3.608

Westin, C., Borst, C., & Hilburn, B. (2016). Automation transparency and personalized decision support: Air Traffic Controller Interaction with A resolution advisory system. *IFAC-PapersOnLine*, *49*(19), 201–206. https://doi.org/10.1016/j.ifacol.2016.10.520

Williams, T. (2013). The psychology of interpersonal trust. How people feel when it comes to trusting someone. McKendree University, Lebanon, IL.

Witteman, H. O., Vaisson, G., Provencher, T., Dansokho, S. C., Colquhoun, H., Dugas, M., Fagerlin, A., Giguere, A. M. C., Haslett, L., Hoffman, A., Ivers, N. M., Legare, G., Trottier, M. E., Stacey, D., Volk, R. J., & Renaud, J. S. (2021). An 11-Item Measure of User- and Human-Centered Design for Personal Health Tools (UCD-11): Development and Validation. *Journal of Medical Internet Research, 23*(3). https://www.jmir.org/2021/3/e15032

Wojton, H. M., Porter, D., Lane, S. T., Bieber, C., & Madhavan, P. (2020). Initial Validation of the Trust of Automated Systems Test (TOAST). *Journal of Social Psychology, 160*(6), 735-750. https://doi.org/10.1080/00224545.2020.1749020

Yang, J., Mossholder, K., & Peng, T. (2009). Supervisory procedural justice effects: The mediating roles of cognitive and affective trust. *The Leadership Quarterly*. *20*. 143-154. https://doi.org/10.1016/j.leaqua.2009.01.009.

Yong, A. G. & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods Psychology. 9*(2). 79-94.

Yurdugul, H. (2008). Minimum Sample Size for Cronbach's Coefficient Alpha: A Monte-Carlo Study. Hacettepe University Journal of Education, 35. 397-405.

Zhuo, Q., Cui, C., Liang, H. *et al.* (2021) Cross-cultural adaptation, validity and reliability of the Chinese Version of Miller Behavioral Style Scale. *Health Qual Life Outcomes*. *19*(86). https://doi.org/10.1186/s12955-021-01717-9

APPENDIX A

ANALYTIC HIERARCHY PROCESS – FACTOR COMPARISON SURVEYS

Human vs. Automation Factors																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Human	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Automation

Human Factors - Confidence																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Purpose and Intent
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Transparency
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Technology Competence
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Faith
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Understandability
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Familiarity
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Predictability
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Demographics
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Personal Attachment
Confidence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk

Human Factors - Purpose and Intent																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Purpose and Intent	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Transparency
Purpose and Intent	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Technology Competence
Purpose and Intent	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Faith
Purpose and Intent	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Understandability
Purpose and Intent	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Familiarity
Purpose and Intent	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Predictability
Purpose and Intent	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Demographics
Purpose and Intent	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Personal Attachment
Purpose and Intent	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk

Human Factors - Transparency																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Transparency	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Technology Competence
Transparency	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Faith
Transparency	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Understandability
Transparency	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Familiarity
Transparency	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Predictability
Transparency	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Demographics
Transparency	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Personal Attachment
Transparency	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk

Human Factors - Technology Competence																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Technology Competence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Faith
Technology Competence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Understandability
Technology Competence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Familiarity
Technology Competence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Predictability
Technology Competence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Demographics
Technology Competence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Personal Attachment
Technology Competence	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk

Human Factors - Faith																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Faith	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Understandability
Faith	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Familiarity
Faith	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Predictability
Faith	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Demographics
Faith	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Personal Attachment
Faith	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk

Human Factors - Understandability																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Understandability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Familiarity
Understandability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Predictability
Understandability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Demographics
Understandability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Personal Attachment
Understandability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk

Human Factors - Familiarity																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Familiarity	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Predictability
Familiarity	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Demographics
Familiarity	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Personal Attachment
Familiarity	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk

Human Factors - Predictability																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Predictability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Demographics
Predictability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Personal Attachment
Predictability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk

Human Factors - Demographics																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Demographics	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Personal Attachment
Demographics	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk

Human Factors -Personal Attachment																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Personal Attachment	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Risk

Automation Factors - Reliability																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Feedback
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Safety
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Usability
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Effectiveness
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Integrity
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Accuracy
Reliability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Suitability

Automation Factors - Feedback																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Feedback	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Safety
Feedback	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Usability
Feedback	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Effectiveness
Feedback	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Integrity
Feedback	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Accuracy
Feedback	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Suitability

Automation Factors - Safety																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Safety	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Usability
Safety	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Effectiveness
Safety	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Integrity
Safety	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Accuracy
Safety	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Suitability

Automation Factors - Usability																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Usability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Effectiveness
Usability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Integrity
Usability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Accuracy
Usability	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Suitability

Automation Factors - Effectiveness																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Effectiveness	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Integrity
Effectiveness	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Accuracy
Effectiveness	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Suitability

Automation Factors - Integrity																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Integrity	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Accuracy
Integrity	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Suitability

Automation Factors - Accuracy																		
Factor	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Factor**
Accuracy	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Suitability

APPENDIX B

FACTOR DEFNITION LIST

Human Factors refer to the inherent individual characteristics of the user and their perceptions of the automation.

- *(User) Confidence* refers to the perception of one's ability to effectively interact with the system in a consistent manner.

- *Purpose and Intent* refers to the user's knowledge of the use case of the automation and its intended actions.

- *Transparency* can be described as the capability of the automation to provide information to the user on its current state and behavior to assist in user understanding.

- *Technology Competence* refers to the perceived technical competence of the system to do the task at hand (Miller & Perkins, 2010). Users of the automated system are able to judge the outcome of task related events to determine automation competence and identify appropriate use cases.

- *Faith* refers to confidence that the automation will perform the intended actions.

- *Understandability* implies that the user knows how and why the automation is performing specific tasks.

- *Familiarity* references past experiences of the user with the automation, supposing some historical context for how the automated system works.

- *Predictability* refers to the matching of the automation performance with the user expectations. When the user is able to predict the automation actions, the user can determine when the automation may fail and adjust their own performance to accommodate.

- *Demographics* contains the factors of culture, age, gender, and personality and refers to their association to propensity for user trust in automation.

- *Personal Attachment* references user agreement that the automated system is agreeable in use and suits personal taste.

- *Risk* refers to the situational use of the automation in hazardous conditions.

Automation Factors refer to the situational characteristics of the automation outside of the user individual characteristics.

- *Reliability* implies that the automation maintains consistent performance free of variation or contradiction.

- *Feedback* refers to the information provided from the system related to the outcome of actions and contextual *future* actions.

- *Safety* implies that the system outcomes do not create unacceptable hazardous conditions for the user.

- *Usability* is the extent to which the system can be effectively used to satisfactorily accomplish specified goals.

- *Effectiveness* refers to the extent to which a system can complete its mission under established constraints.

- *Integrity* refers to the degree to which the automated system adheres to a set of established principles.

- *Accuracy* is how often the automated system makes a correct decision.

- *Suitability* refers to the appropriateness of the automation capabilities to carry out the tasks.

APPENDIX C

AVIATION SYSTEMS – TRUST SURVEY (AS-TS) INSTRUCTION PACKET

Aviation Systems – Trust Survey (AS-TS)

The Aviation Systems – Trust Survey (AS-TS) is a Trust in Automation (TIA) survey that uses a 7-point Likert scale to identify trust deficiencies. Where a trust deficiency is an inadequate trust relationship between the user and the automated system, where the user perceives the system as untrustworthy and/or the user is unable to reasonably calibrate their trust in the system. The AS-TS was validated through Subject Matter Expert (SME) review and pilot participation by rating various automation related scenarios for helicopter flights.

Two overarching factors influence the AS-TS Trust construct. Human Factors and Automation Factors both play a role in pilot trust. Under each factor are four scale items that reflect pilot perception of the automated system under test.

Administration: The AS-TS should be administered after pilots complete an interaction (e.g., test run) with the system under review. It is recommended that the AS-TS be administered individually for each system under test, rather than a "total" measurement of trust within a complex system. While, the survey may provide usable information as a total system tool, the mitigation strategies will need to focus on specific system deficiencies. Participants should be provided the definition list when completing the survey to ensure consistent context.

Consideration: "The system provides appropriate feedback on current actions." and "The system provides appropriate feedback on future actions." should be tailored for the system characteristics. Some autonomous systems do not provide predictive features. In this case, "The system provides appropriate feedback on future actions" should be removed.

Analysis: Descriptive statistics (i.e., Mean, Standard Deviation) are useful for identifying trends in participant responses. When comparing systems, the T-Test, Mann-Whitney U, and ANOVA can be used for comparing statistical significance. Additionally, AS-TS responses can be compared to usability questionnaires and objective reliability data. An ideal outcome is positive correlation among the data. Deficiencies in any of the three areas are likely to cause a ripple effect through other survey responses and potentially user performance.

Actions: Negative or unexpected ratings should be investigated through participant interview to determine root cause, user reasoning, and identify potential mitigation strategies. Survey evaluators can use the mitigation attachment for initial mitigation strategies. Additional

care should be taken to consider all data collected (e.g., usability, reliability, trust) as a systematic evaluation of the system under test.

Definition List:

Human Factors refer to the inherent individual characteristics of the user and their perceptions of the automation (Henshel et al., 2015).

- *User Confidence* refers to the perception of one's ability to effectively interact with the system in a consistent manner (Dolgov, et al., 2017).
- *Transparency* can be described as the capability of the automation to provide information to the user on its current state and behavior to assist in user understanding (Westin. Borst, & Hilburn, 2016).
- *Understandability* implies that the user knows how and why the automation is performing specific tasks (Sheridan, 2019).
- *Familiarity* references past experiences of the user with the automation, supposing some historical context for how the automated system works (Sheridan, 2019).

Automation Factors refer to the situational characteristics of the automation outside of the user individual characteristics (Henshel et al., 2015).

- *Faith* refers to confidence that the automation will perform the intended actions (Miller & Perkins, 2010).
- *Feedback* refers to the information provided from the system related to the outcome of actions and contextual *future* actions (Schaeffer et al., 2016).
- *Effectiveness* refers to the extent to which a system can complete its mission under established constraints (Dordick, 1965).
- *Suitability* refers to the appropriateness of the automation capabilities to carry out the tasks (Smith, Allaham, & Wiese, 2016).

Aviation Systems – Trust Survey

Trust Questions	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
I am confident in my ability to utilize the system.							
The system provides transparent system information.							
I understand what the system is doing.							
I am familiar with the system operation.							
I have faith that the system will perform the intended task.							
The system provides appropriate feedback on *current* system actions.							
The system provides appropriate feedback on *future* system actions.							
The system effectively accomplishes its tasks.							
The system is suitable for carrying out this type of task.							

Comments:

Mitigation Recommendations

Human Factors	Mitigation Recommendation
User Confidence	Improve system training and knowledge of system limitations.
Transparency	Improve delivery of useful information to the user.
Understandability	Inform users on how and why the automation performs specific tasks to allow for accurate mental models of system processes.
Familiarity	Provide more training or opportunity to work with the system and historical context of use.
Automation Factors	**Mitigation Recommendation**
Feedback	Ensure the system provides appropriate user information and, when required, contextual future actions of the automation.
Effectiveness	Ensure the system can accomplish required mission tasks to standard.
Faith	Ensure automation consistently performs the intended actions.
Suitability	Ensure the system is being used for the appropriate tasks.