USAARL-TECH-TR--2024-10

UNITED STATES ARMY AEROMEDICAL RESEARCH LABORATORY

Measuring Trust in Automation in Operational Aeromedical Settings: A Systematic Review of the Literature

Bethany Ranes, Jordayne Wilkins, Emily Kenser, & Marissa Caid-Loos

Notice

Qualified Requesters

Qualified requesters may obtain copies from the Defense Technical Information Center (DTIC), Fort Belvoir, Virginia 22060. Orders will be expedited if placed through the librarian or other person designated to request documents from DTIC.

Change of Address

Organizations receiving reports from the U.S. Army Aeromedical Research Laboratory on automatic mailing lists should confirm correct address when corresponding about laboratory reports.

Disposition

Destroy this document when it is no longer needed. Do not return it to the originator.

Disclaimer

The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation. Citation of trade names in this report does not constitute an official Department of the Army endorsement or approval of the use of such commercial items.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
08-12-2023	Technical Report	NOV 2022 - JUL 2023

4. TITLE AND SUBTITLE

Measuring Trust in Automation in Operational Aeromedical Settings: A Systematic Review of the Literature

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6.3

6. AUTHOR(S)

Ranes, B.[1,2], Wilkins, J.[1,3], Kenser, E.[4], & Caid-Loos, M.[5]

5d. PROJECT NUMBER

MO230061

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

U.S. Army Aeromedical Research Laboratory
P.O. Box 620577
Fort Novosel, AL 36362

8. PERFORMING ORGANIZATION REPORT NUMBER

USAARL-TECH-TR--2024-10

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command
Military Operational Medicine Research Program
504 Scott Street
Fort Detrick, MD 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

USAMRDC MOMRP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release: distribution is unlimited.

13. SUPPLEMENTARY NOTES

[1]U.S. Army Aeromedical Research Laboratory, [2]Goldbelt Frontier, LLC, [3]Oak Ridge Institute for Science and Education, [4]U.S. Army Aviation Center of Excellence, [5]Lyster Army Health Clinic

14. ABSTRACT

As military environments integrate more complex technological systems, operators increasingly require more assistance in the form of automation. When used properly, automation has the potential to significantly enhance performance; however, proper use is predicated on the operator's trust in the automation (TIA). TIA, like trust among people, is a complex construct that is influenced by biological, psychosocial, and behavioral aspects that affect (and are affected by) how a user feels about a system. While options for measuring TIA have rapidly expanded in the past decade, there has been little consideration for how well these measures can perform in operational environments. The purpose of this review was to explore the literature produced over the previous ten years to identify all means of measuring TIA, evaluating the quality of the studies that used each measure, and rating how well each measure would perform in an operational aeromedical environment. A recommendation of 28 behavioral, physiological, and user-reported TIA measures is provided, as well as a list of 23 measures with a cautious recommendation (including caveats for use) and six TIA measures that are not recommended. While these recommendations offer a starting point for testing TIA in aeromedical settings, further research is required to test how well these recommended measures actually perform in an operational aeromedical environment.

15. SUBJECT TERMS

trust, automation, aviation, performance, trust measurements, operational environment, TIA

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Loraine St. Onge, PhD
UNCLAS	UNCLAS	UNCLAS	SAR	29	19b. TELEPHONE NUMBER *(Include area code)* 334-255-6906

This page is intentionally blank.

Summary

As military aeromedical environments integrate more complex technological systems, operators increasingly require more assistance in the form of automation. When used properly, automation has the potential to significantly enhance performance; however, proper use is predicated on the operator's trust in the automation (TIA). TIA, like trust among people, is a complex construct that is influenced by biological, psychosocial, and behavioral aspects that affect (and are affected by) how a user feels about a system. While options for measuring TIA have rapidly expanded in the past decade, there has been little consideration for how well these measures can perform in operational environments. The purpose of this review was to explore the literature produced over the previous ten years to identify all means of measuring TIA, evaluating the quality of the studies that used each measure, and rating how well each measure would perform in an operational aeromedical environment (aeromedical appropriateness). A recommendation of 28 behavioral, physiological, and user-reported TIA measures is provided, as well as a list of 23 measures with a cautious recommendation (including caveats for use) and 6 TIA measures that are not recommended. While these recommendations offer a starting point for testing TIA in aeromedical settings, further research is required to test how well these recommended measures actually perform in an operational aeromedical environment.

This page is intentionally blank.

Acknowledgements

This research was supported in part by an appointment to the Oak Ridge Institute for Science and Education Research Participation Program at the U.S. Army Aeromedical Research Laboratory administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Army Medical Research and Development Command. The authors would also like to thank Mr. Anthony Waterman for his gracious and valuable assistance in the completion of this research.

This page is intentionally blank.

Table of Contents

List of Figures and Tables

This page is intentionally blank.

Introduction

Complex technology advancements are rapidly increasing and being integrated into a wide variety of operational environments; this is especially true in military environments. As the complexity of these technologies increase and as their role becomes more centralized to critical tasks, humans who are interacting with advanced technology on a regular basis may find it more difficult to oversee successful operation without assistance. Automated systems have the potential to significantly improve the interactions between advanced technology and human operators, optimizing the strengths of both human and non-human components in order to maximize the frequency of successful outcomes. However, for an automation to be effective, it must be used appropriately by the human operator, and appropriate use is largely dictated by the human operator's level of trust in the automation. For this reason, trust in automation (TIA) has emerged as one of the most significant considerations for engineering the next generation of complex technological innovations.

Defining Trust in Automation

Trust is generally 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, p. 51), and this definition has been found through decades of research to be suitable for describing TIA as well (Kohn et al., 2021). Trust is a critical component in ensuring proper human engagement and performance in any type of collaboration, and this is no different with collaborations with automation, which includes any technology that "actively selects data, transforms information, makes decisions, or controls processes" (Lee & See, 2004). Just like trust among humans, TIA is highly complex and is dynamically influenced by many intertwining factors that can have immediate and significant impacts on a user's subsequent behavior during an automation-enhanced task (Lee & See, 2004; Hoff & Bashir, 2015; Mayer et al., 1995). In general, these factors have been identified by researchers as human-based, automation-based, and environment-based (Lee & See, 2004; Hoff & Bashir, 2015). Human-based trust factors are those that are related to the operator themselves, and include personality traits, pre-existing knowledge, ethnicity, age, and gender (Lu & Sarter, 2019). Automation-based trust factors are those that are inherent to the system being used, and include system reliability, ability, robustness, and predictability. Environment-based factors are perhaps the most complicated and difficult to measure element of trust in an interaction, and include things like prior experience, societal impact, culture, team collaboration, and task type. TIA is also a highly dynamic construct based upon one's experience across time, and experiences can influence trust-related beliefs and behaviors in both the short-term (e.g., during an automated task) and long-term (e.g., a long career with experiences across multiple types of automations).

TIA is also complicated by its bi-directional nature, as optimal TIA lies in the middle of a spectrum between under-trusting and over-trusting a system. The overall level of TIA impacts the level of vigilance and sustainable attention an operator will give toward automation (Krausman et al., 2022). When operators place too much trust in a system (also called *complacency*), it can lead to an increased risk of mistakes, incidents, and accidents related to the user being "out of the loop" (Krausman et al., 2022; Lu & Sarter, 2019). Low levels of trust cause disuse of an automated system and can lead to unnecessarily high levels of user workload driven by the need to constantly (and unnecessarily) monitor automated systems to ensure safety

and accuracy (Lu & Sarter, 2019). Increased workload can significantly increase the risk of mistakes and oversights that would otherwise be identified and accounted for by the automation (Lee & See, 2004). Several researchers have hypothesized that calibrating appropriate levels of user trust through measurement and modification of human-based, automation-based, and environment-based TIA factors will result in ideal user reliance for optimal performance outcomes (Lee & See, 2004; Mayer et al., 1995; Sanders et al., 2021).

Types of TIA Measures

Trust is an emotional construct. As with any other emotional variable, it is perceptual, which means the experience of trust is context-dependent, based on a complex interaction of environmental and psychophysiological components, and entirely unique to the individual experiencing it (Barrett, 2017). All these considerations make valid and reliable measurement of trust particularly difficult. In the case of TIA, measures are often used to inform the automation itself, and therefore require data that are not only valid and reliable but are also continuously collected and able to be analyzed and modeled at an interval or ratio level of measurement (Wei et al., 2020). Researchers have attempted to overcome the inherent obstacles of properly measuring TIA by developing a wide array of instruments and metrics, ranging from simple single-item self-report measures to algorithmically constructed values representing the integration of several different elements related to trust. While most researchers advocate the use of a multi-modal approach to capturing TIA (a combination of different types of TIA metrics collected simultaneously and interpreted holistically), single-type TIA metrics generally fall into three broad categories: user-reported, physiological, and behavioral.

User-reported measures.

Recording an individual's verbal or written rating of how they feel is the simplest and most common approach to measuring subjective variables, and TIA is no exception. User-reported measures are the most frequently used TIA data collection method, although there is significant variation in content and complexity of these measures (Wei et al., 2020). While many scientists quantify trust by simply asking study participants to rate their level of trust along a Likert-type scale (i.e., rating one's level of trust on a scale of 0 to 100), many more sophisticated TIA self-report measures have been developed to more precisely capture the different psychosocial aspects that have been highlighted by various models of interpersonal trust (e.g., Jian et al., 2000; Merritt et al., 2015). As with any standardized user-report measure, TIA questionnaires must demonstrate sound psychometric properties in order to be deemed a source of high-quality data. These properties include elements of validity (how well the instrument actually measures trust, and more specifically, how well it measures TIA) and reliability (how consistent its measurements are between subjects and over time). Availability of research on the psychometric properties of user-reported TIA measures is mixed, but common TIA instruments generally demonstrate high levels of internal consistency and acceptable correlation with other trust measures (Dolgov & Kaltenbach, 2017; Jessup et al., 2019). While some scientists have raised concerns that psychometric properties of self-report TIA measures have not been tested with enough scrutiny to rely upon them as valid sources of data (Madhavan, 2015), others have demonstrated that data from user-reported sources of TIA exhibit enough consistency and accuracy to be analyzed at an interval/ratio level of measurement (Wei et al., 2020). Practical criticisms of user-reported measures center on the fact that they require users to be consciously

aware of the construct being measured, and to also be forthcoming in their assessment. In the case of TIA, it is possible that users may not recognize subtle changes in their own levels of trust, may feel pressure to report a certain way, or may over- or under-report their levels of trust when compared to a more objective source of data (e.g., observing how reliant the user is on the automation's assistance) (Drnec et al., 2016). Conscious awareness also requires a fair amount of attention, so repeated user-reported data capture can pose a burden on cognitive workload and/or can become disruptive to the completion of the study task (Lee et al., 2018). For this reason, a large number of user-reported TIA metrics are focused more on trait-based components that contribute to TIA levels during an automated task (e.g., propensity to trust, personal beliefs about the general trustworthiness of computerized assistance, etc.).

Physiological measures.

Physiological measures address several of the weaknesses associated with user-reported measures; namely, that they can capture changes in user state that are too subtle for conscious awareness and that they collect data passively as to not interrupt the user while they complete their task(s). The continuous feed of interval/ratio level data produced by physiological sensors are particularly valuable to TIA applications, since these data can be fed directly to a system running an automation in order to model and seamlessly respond to dynamic changes in a user's state. For these reasons, physiological indicators are the preferred method of TIA measurement while an automated task is being completed. However, physiological measures are not without their own challenges and weaknesses when it comes to quantifying trust. A major practical consideration is that physiological sensors can introduce distracting equipment that is unfamiliar to users in their task environment, or the task environment may make reliable data capture impossible if sensors require a highly specific experimental setup (Balters & Steinert, 2017). A more fundamental challenge is that current physiological sensors are measures of autonomic arousal, and while they often demonstrate good sensitivity to changes in affective variables (like trust), they are by no means specific to them, which means that using physiological sensors to detect and quantify specific emotional states has not been well-supported in the literature (Balters & Steinert, 2017; Barrett, 2017). Physiological data do hold significant value as a measure of emotional intensity, and particularly as a measure of *dynamic* emotional intensity, which allows researchers to detect changes in affective states (Balters & Steinert, 2017). When correlated with more specific TIA measures, continuous detection of changes in affective intensity helps identify affective spikes and troughs related to TIA fluctuations that are likely to impact the user's interactions with the system (Lee & See, 2004). Physiological shifts in intensity are also strong measures of attentional awareness, which has been identified as a strong correlate and possible predictor of TIA during a user's engagement with an automated system. The unique inter-relationship between trust and attention makes physiological measures that are sensitive to attentional awareness particularly valuable for quantifying changes in TIA during an automated task (Parasuraman et al., 2008).

Behavioral measures.

Behavioral measures (sometimes referred to as observational measures) offer insight into demonstrations of TIA, which are significantly less prone to bias than a user's judgment and/or report of their own TIA level (Miller et al., 2016). Since the end goal of TIA research is not intended to be a comprehensive understanding of the "invisible" cognitive and affective elements of trust, per se, but is instead meant to use predictions of users' TIA-influenced behavior to engineer better automated systems, behavioral data inform outcomes more directly aligned with the mission of TIA research (Ajzen & Fishbein, 1980; Hoff & Bashir, 2015). Like physiological measures, behavioral measures are best utilized while an automated task is being completed by the user. Behavioral measures often target a user's explicit reliance on an automation (e.g., how frequently they adhere to an automated recommendation), and in order to be quantified, they rely on a standardized means of capturing each automated suggestion, the user's response to that suggestion, and the reliability/accuracy of the automated suggestion. While this is a minor hurdle for laboratory-based tasks that have been specifically engineered for the purposes of measuring TIA, it can pose a practical challenge for measuring TIA during real-world tasks and environments. As a result, the ongoing capture of behavioral measures is particularly critical when considering the design and engineering of automated systems, as the integration of these factors can provide continuous, unbiased TIA data that the system can respond to during an automated task in order to optimize TIA levels and subsequently optimize user performance
(Parasuraman et al., 2008; Chancey et al., 2015).

Constructing Scenarios for Testing TIA

Since trust is dependent upon the presence of a "situation characterized by uncertainty and vulnerability" (Lee & See, 2004), the testing scenario used to elicit feelings of trust (or the lack thereof) is also a crucial element for ensuring the ecological validity of TIA measures. Automations are meant as a tool for reducing uncertainty in a complex situation; if a task does not initially present the operator with any sense of uncertainty or challenge as to the correct response, then trust is unable to form at all because the operator does not need automation to achieve their goal (Moorman et al., 1993). Vulnerability is dependent upon the level of risk inherent to the scenario (Mayer et al., 1995). If there is no perceived risk as a result of failing to achieve a goal, then trust cannot form because the operator does not need to achieve their goal at all (Johns, 1996). For obvious ethical reasons, the necessity of risk poses a challenge for developing testing scenarios that elicit vulnerability. Most research scenarios handle risks abstractly; the operator must determine automation trustworthiness while simply imagining possible risks, or researchers assume that failing to properly complete a task is risk enough on its own (Kohn et al., 2021). This calls into question whether the measures of TIA acquired during these scenarios are consistent with an operator's actual interactions with an automation-assisted task in the real world. However, many researchers have found success in eliciting appropriate uncertainty and vulnerability by using scenarios from operational settings that have specific mission aims with clear ties to perceived risk in the event of failure (such as military operations or monitoring a self-driving car in high-traffic areas). For example, (Lyons & Stokes, 2012) successfully simulated uncertainty and vulnerability for a military context with a scenario of convoy route planning aids with the risk of a simulated attack. In these cases, scenarios are designed for specific operational tasks using cues and equipment that closely mirror real-life environments.

Purpose of This Review

While TIA has emerged as a popular research topic across a wide variety of disciplines, most of the work that has been completed to-date is either theoretical or tailored to a controlled laboratory setting. Despite a global recognition that continuous measures and real-time responses to changes in user TIA are critical for monitoring its dynamic nature, many TIA instruments are static in nature and many experiments have not taken into consideration the logistical challenges of an operational setting. The purpose of this review is to create an index of user-reported, physiological, and behavioral TIA measures from the scientific literature published over the past ten years, and to identify which TIA instruments and testing scenarios may be feasible for use in an operational aeromedical environment. While several studies of TIA have been tailored to a military aviation application (for fixed- and rotary-wing aircraft), most studies are taken from laboratory settings or other operational settings (such as automation systems in self-driving cars). It is the intent of this report to capture TIA metrics and scenarios from a broad spectrum of use cases that may be valuable for innovating technology for aeromedical applications.

Methods

A review of the current literature was conducted to collect a wide range of TIA instruments and scenarios that have been used in research studies over the past decade. While healthcare and aviation domains were included as specific areas of interest, most articles came from a broad search of the TIA literature. This was done intentionally to gain a broad perspective of TIA measures that have been used across multiple fields in order to identify opportunities to apply techniques and scenarios that have been used in other operational settings that may also be valuable for an aeromedical environment. Searches were conducted in three databases: Google Scholar, EBSCO, and the Defense Technical Information Center's (DTIC) collection of military technical reports. Each search included a Boolean list of terms that were designed to isolate articles referencing specific TIA measures and measurement techniques from articles discussing theoretical models of trust and TIA. A summary of the search terms is included in the table below.

This space is intentionally blank.

Table 1. Summary of Search Terms

Database	Search terms
Google Scholar	"trust in automation" OR "trust measures"
	"trust in automation" OR "trust measures" AND "medical"
	"trust in automation" OR "trust measures" AND "aviation"
	"measure trust in automation"
	"measure trust in automation" AND "medical"
	"measure trust in automation" AND "aviation"
EBSCO	"measure trust in automation medical aviation"
	"trust in automation" OR "trust measures"
	"trust in automation" OR "trust measures" AND "medical"
	"trust in automation" OR "trust measures" AND "aviation"
DTIC	"measure trust in automation"
	"trust in automation" OR "trust measures"
	"trust in automation" OR "trust measures" AND "aviation"
	"trust in automation" OR "trust measures" AND "medical"

Each search was filtered to restrict results to articles published within the past ten years (2013-2023), written in English, approved for public release, and peer-reviewed (except for technical reports pulled from DTIC, which were published after scientific review and routing procedures consistent with government regulations). Articles needed to include at least one measure of TIA, collected before, during, and/or after an individual's interaction with an automated system. Review articles discussing various TIA measures and articles that analyzed the psychometric properties of instruments designed to measure TIA were also included. Initial searches were conducted in March 2023. Results from each of the initial searches were reviewed for inclusion and exclusion independently by two reviewers. In addition to the filters applied to the initial search, the following types of papers were excluded: articles that discussed a theoretical model of TIA with no specific measures or recommendations for quantifying it, articles limited to philosophical considerations of TIA and its importance to developing automated systems, dissertations or theses, and conference presentations. Specific measures and scenarios were manually extracted by the reviewers and categorized based on whether testing scenarios were laboratory-based or field-based and whether TIA measures included were user-reported, physiological, or behavioral.

Final articles included in the review were rated independently by two reviewers on the research team. Quality ratings consisted of Likert scale responses (1-5) for 13 research quality criteria from the standardized Justification, Operationalisation, Replicability, Interval validity, Presentation, Interpretation, External validity, and Final Judgement (JORIPIEF) critical research review checklist developed for psychological research articles (Barber, 2003). A copy of the quality review checklist used by the reviewers is included in Appendix A. Any article that was rated with an overall score of less than 1.5 was excluded from the review, as this suggested poor enough research quality as to call into question the validity of its results. Reviewers were deemed qualified to evaluate research quality given a minimum five years of aeromedical research experience and a minimum of 12 academic credit hours in research methodology from an accredited university. In instances where overall research quality rating differed by two or more,

the reviewers discussed their ratings and modified initial scores as needed to reach a consensus.

In addition to quality ratings conducted by researchers for each study, each of the 159 TIA measures was independently evaluated for aeromedical appropriateness by three members of the research team. Reviewers were deemed qualified to evaluate aeromedical appropriateness if they were a currently rated active-duty fixed-wing aviator and/or if they had five or more years of experience as an aeromedical researcher. Raters for this review included two active-duty fixed-wing aviators (one rated Black Hawk pilot and one rated Apache pilot) and a U.S. Army Aeromedical Research Laboratory research technician with five years of aeromedical study experience. Aeromedical appropriateness was defined as the extent to which a measure could be properly implemented according to its best practices within the physical and operational confines of an active aeromedical environment. Low aeromedical appropriateness might include measures that require large, specialized equipment that could not fit inside of an aircraft and/or measurement tasks that would significantly disrupt a typical flight task; conversely, high aeromedical appropriateness would suggest measures that require no specialized equipment beyond what is already available in an aircraft and/or collect information on TIA with no discernable disruption to flight tasks. Since there are no standardized methods for rating aeromedical appropriateness, ratings were determined based on criteria that were developed by the research team for the purposes of this review. Ratings were made on a scale of 1 through 5 across six criteria, with 1 representing the lowest possible aeromedical appropriateness and 5 representing the highest possible aeromedical appropriateness. A copy of the aeromedical appropriateness checklist used by the reviewers is included in Appendix A. Similar to the methodology for achieving inter-rater agreement for research quality ratings, in instances where overall aeromedical appropriateness rating differed by two or more, the reviewers discussed their ratings and modified initial scores as needed to reach a consensus.

Results

Initial searches were filtered in all three databases to return only English-language results that had been published in the past ten years. The Boolean search terms included in Table 1 yielded a total of 3561 titles. Specific frequencies of each search are reported in Appendix B. Two researchers screened the list of titles and selected those which were relevant to the purpose of the review. Relevance at this stage of the review was determined based on whether the title alluded to operator TIA specifically, or if the study was exploring factors related to the proper use of an automation that could reasonably include TIA. Following the initial title review, both researchers conducted an independent review of abstracts to remove duplicate selections and further determine whether studies met criteria for inclusion; conceptual models of TIA with no specific measurement instruments were removed, as were conference presentations, dissertations, and theses. After these exclusions, a total of 159 abstracts were selected. Full-texts were acquired for all selected abstracts and were once again independently screened by the researchers to determine final inclusion based on whether the study included at least one specific measure of TIA and/or described a specific scenario used to elicit TIA. The research quality of each full-text article was also rated independently by the researchers on a scale of 1 (very poor) to 5 (very strong), based on the JORIPIEF checklist (Barber, 2003), which is a standardized checklist developed for the purposes of critically critiquing the overall quality of a research study in a consistent and replicable way (a copy of this checklist is included in Appendix A). While studies demonstrated a range of quality levels, no studies were deemed to be of poor enough quality to

merit exclusion from the review (i.e., scoring less than 1.5 on the JORIPIEF checklist). The mean overall quality rating was $\bar{x}$ = 3.66, with a median of η = 3.50. After all screening and exclusion efforts were completed, 31 articles remained and were included in the review. These articles included 6 detailed scenarios and 153 TIA measures: 94 were user-reported, 34 were physiological, and 25 were behavioral (see Figure 1).

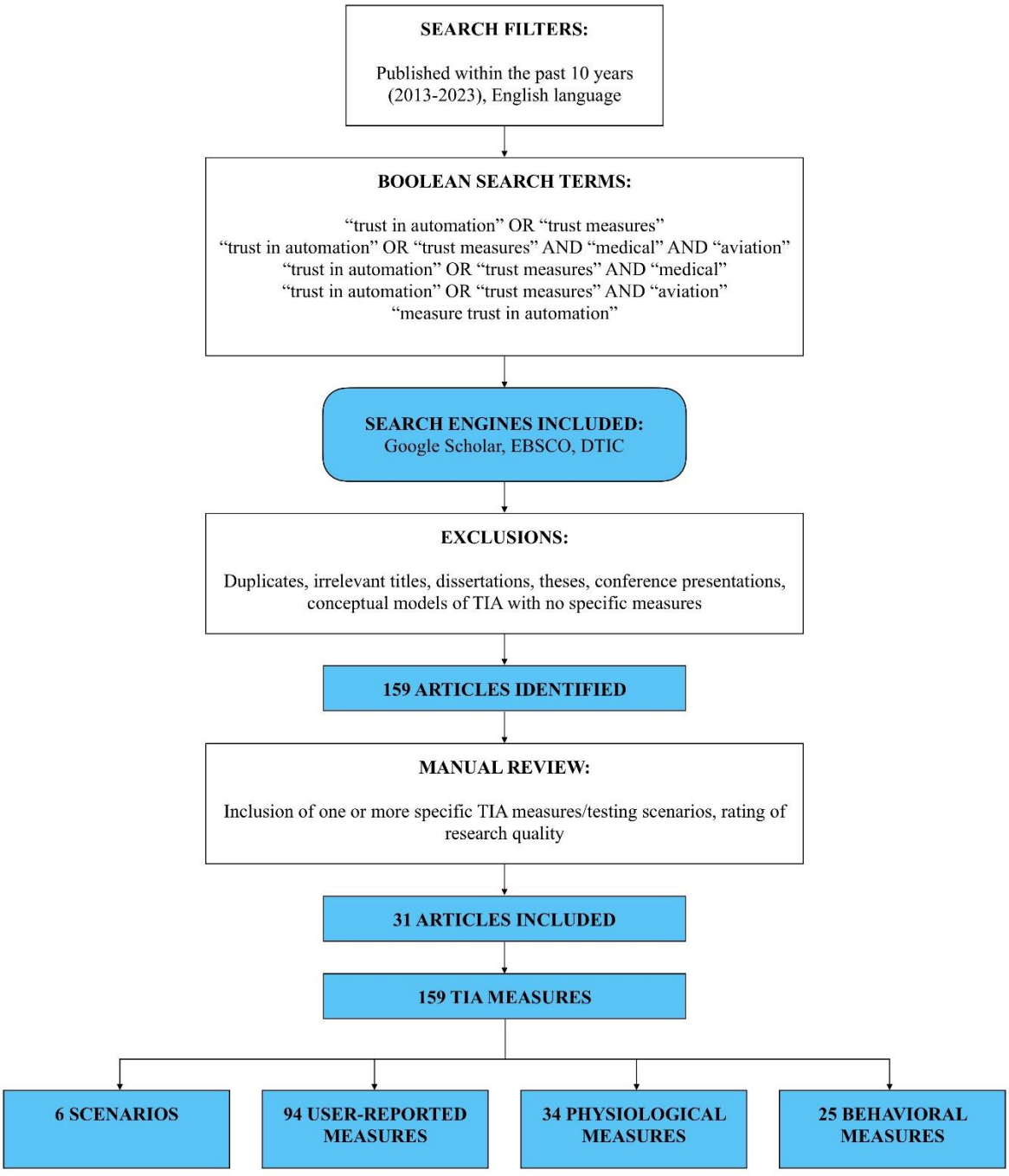

Figure 1. Search process and screening results.

Of the six scenarios that were described in the research, two were created specifically for the study at hand and not intended for generalized use, so they were not included in the analysis. The other four were standardized platforms created specifically to evaluate TIA and subsequent performance on the scenario task(s). All four of these standardized platforms were designed to emulate flight tasks and are suitable for testing automations that might be used in aerospace operations (if not an aeromedical environment, specifically). Two studies used the system monitoring task (SYSMON) from the Multi-Attribute Task Battery (MATB-II), which was developed by the U.S. Air Force to replicate common operational tasks that can be completed on a computer (National Aeronautics and Space Administration, 2016). One study used the Mixed Initiative Team Performance Assessment System (MITPAS), which is also a military-based performance scenario that was designed to evaluate performance of combined teams of humans and unmanned systems (Freedy et al., 2007). Another study used the Research Environment for Supervisory Control of Heterogenous Unmanned Vehicles (RESCHU-V), which was developed to test TIA and performance in the context of unmanned aerial vehicle (UAV) missions (Nehme, 2009). Evaluations of the studies that included these scenarios determined them to be of equal research quality (3.50/5.00). A summary of quality ratings for the four standardized platforms is provided in Table 2.

Table 2. Summary of Research Quality Ratings for TIA Testing Scenarios

Scenario name	***n***	**Mean research quality rating (range)**
Multi-Attribute Task Battery (MATB-II) SYSMON Task	2	3.50 (3,4)
Mixed Initiative Team Performance Assessment System (MITPAS)	1	3.50 (3,4)
Research Environment for Supervisory Control of Heterogenous Unmanned Vehicles (RESCHU-V)	1	3.50 (3,4)
All standardized scenario platforms	**4**	**3.50**

Behavioral measures included 18 measures of observed behaviors (a total of 10 unique types of behaviors were used overall) and seven measures of distinct communication patterns. A full list of these behavioral measures is included in Appendix C. Physiological measures included ten distinct methods of collection, and the most frequently used was eye tracking (n = 15). A full list of physiological measures is included in Table 3. Seventy-one of the 94 user-reported measures were collected using validated instruments (a total of 29 unique instruments were used overall). Seventeen studies collected TIA using a Likert scale (where operators were simply asked to rate their current levels of trust on an ordinal numerical scale), and six non-validated instruments were created specifically for use in a particular study. Since they were not intended to be scaled for broad research use, the six non-validated instruments were not included in our final analysis. A full list of the validated instruments (and the frequency that they were used among the studies in this review) is reported in Appendix C.

Table 3. Summary of Research Quality and Aeromedical Appropriateness Ratings for All TIA Measure Types

TIA measure type		*n*	Mean research quality rating (SD)	Mean aeromedical appropriateness rating (SD)
Behavioral	Observed behavior	18	3.61 (0.53)	4.11 (1.00)
	Communication pattern	7	5.00 (0.00)	4.14 (0.50)
	Overall behavioral	**25**	**4.00 (0.78)**	**4.12 (0.88)**
Physiological	Eye tracking	15	3.57 (0.73)	3.82 (0.35)
	Skin conductance	5	3.10 (0.55)	2.60 (1.03)
	Electroencephalogram (EEG)	3	3.17 (0.58)	2.61 (0.76)
	Electrocardiogram (ECG)	2	3.50 (0.50)	2.83 (0.24)
	Heart rate/HRV	2	3.50 (0.50)	3.00 (0.00)
	Functional magnetic resonance imaging (fMRI)	2	3.00 (0.71)	1.00 (0.00)
	Facial expression	2	3.75 (1.77)	4.33 (0.47)
	Blood volume pulse	1	5.00 (0.00)	4.67 (0.47)
	Functional fear-infrared spectroscopy (fNIRS)	1	3.50 (0.50)	3.67 (0.47)
	Physiological synchrony	1	5.00 (0.00)	4.50 (0.50)
	Overall physiological	**34**	**3.51 (0.76)**	**3.35 (0.96)**
User-reported	Validated instrument	71	3.66 (0.66)	3.29 (0.75)
	Likert scale	17	3.44 (0.35)	3.85 (0.32)
	Overall user-reported	**88**	**3.62 (0.62)**	**3.40 (0.71)**
	All included measures	**147**	**3.66 (0.70)**	**3.51 (0.83)**
	All measures median scores	**147**	**3.50**	**3.67**

The overall mean aeromedical appropriateness rating for all measures remaining in the analysis (n = 147) was $\overline{x}$ = 3.51, with a median of η = 3.67. A summary of all research quality and aeromedical appropriateness ratings for the TIA measures is provided in Table 3. Each measure's individual quality and aeromedical appropriateness rating was compared against the median; scores below the median were considered "low" while scores above the median were considered "high." These comparisons were used to categorize each measure based on strength of recommendation for using it in an operational aeromedical environment. Measures in the Strong category demonstrated high research quality and aeromedical appropriateness ratings. Measures in the Cautious category demonstrated a high rating in one area (research quality or aeromedical appropriateness) and low rating in the other. Measures in the Not Recommended category demonstrated low ratings of research quality and aeromedical appropriateness. Overall, there were 29 TIA measures in the High Recommendation category and 22 TIA measures in the Cautious Recommendation category. Six TIA measures were Not Recommended for use in an aeromedical research environment. A complete list of each recommendation category is included in Figure 2.

Measures with High Recommendation		
Measure Name	**Research Quality**	**Aeromedical Appropriateness**
B: Behavioral Synchrony and Entrainment	High	High
B: Combined Team Performance	High	High
B: Compliance with automation recommendations	High	High
B: Delegation	High	High
B: Intervention	High	High
B: Reliance	High	High
B: Response Time	High	High
Communication: Bottom-Up Measures of Communication Content	High	High
Communication: Flow	High	High
Communication: Network Analysis	High	High
Communication: Rate	High	High
Communication: Real-Time Event, Flow, and Coordination Tool (REFLECT)	High	High
Communication: Top-Down Measures of Communication Content	High	High
P: Blood Volume Pulse (BVP)	High	High
P: Eye Tracking	High	High
P: Facial Expression	High*	High
P: Physiological Synchrony	High	High
U/SR: Adapted Propensity to Trust Questionnaire (Jessup 2019)	High	High
U/SR: Checklist for Trust between People and Automation (Jian 2000)	High	High
U/SR: Complacency Potential Rating Scale (CPRS) (Singh et al, 1993)	High	High
U/SR: Cross-Cultural Automation Trust Scale (Chien et al 2014)	High	High
U/SR: Draper Trust Scales (Jackson et al, 2016)	High	High
U/SR: Dynamic Reporting of Trust (Desai, 2012)	High	High
U/SR: Integrated Model of Trust Scale (Muir & Moray, 1996)	High	High
U/SR: Perfect Automation Schema	High	High
U/SR: Propensity to Trust Scale (Mayer & Davis 1999)	High	High
U/SR: Trust in Automated Systems Test (TOAST; Wojton et al, 2020)	High	High
U/SR: Trust Scale (Merritt, 2011)	High	High
**Facial expression has mixed performance in affective literature as a measure of emotion detection. See discussion.*		
Measures with Cautious Recommendation		
Measure Name	**Research Quality**	**Aeromedical Appropriateness**
B: Manual verifications	Low	High
Likert Scale	Low	High
U/SR: Hypothetical Complacent Behavior	Low	High
U/SR: Implicit Association Test (Merritt 2013)	Low	High
U/SR: Negative Attitude towards Robots Scale (Nomura et al, 2006)	Low	High
U/SR: Propensity to Trust Machines Scale (Merritt & Ilgen 2008)	Low	High
B: Economic Trust Game: Stakes Invested	High	Low
B: Proximity	High	Low
Communication: Bag of Words	High	Low
P: Electrocardiogram (ECG)	High	Low
P: Functional Near-Infrared Spectroscopy (fNIRS)	High	Low
P: Heart Rate/Variability	High	Low
U/SR: Human Computer Trust Scale (HCTS; Madsen & Gregor, 2000)	High	Low
U/SR: Human-Robot Interaction Trust Scale (Yagoda & Gillan, 2012)	High	Low
U/SR: Interpersonal Trust Scale (Rotter, 1967)	High	Low
U/SR: Multi-Dimensional Measure of Trust (MDMT; Malle & Ullman, 2021)	High	Low
U/SR: Propensity to Trust Survey (Evans & Revelle, 2008)	High	Low
U/SR: Propensity to Trust Technology Scale (Schneider et al, 2017)	High	Low
U/SR: System Trustworthiness Scale (Schaefer et al, 2012)	High	Low
U/SR: Trust & Self-Confidence Scale (Lee & Moray, 1994)	High	Low
U/SR: Trust in Automation Scale (Korber et al, 2015)	High	Low
U/SR: Trust Perception Scale for Human-Robot Interaction (HRI; Schaefer, 2016)	High	Low
U/SR: Trust Scale (Lee & Moray, 1992)	High	Low
Measures that are Not Recommended		
Measure Name	**Research Quality**	**Aeromedical Appropriateness**
P: Electroencephalagram (EEG)	Low	Low
P: Functional Magnetic Resonance Imaging (fMRI)	Low	Low
P: Skin Conductance	Low	Low
U/SR: Godspeed Measure	Low	Low
U/SR: Multitasking Preference Inventory	Low	Low
U/SR: Trust Measure (Yan et al)	Low	Low

Figure 2. Measures listed by recommendation category.

Discussion

High-level categorizations of TIA measurements based on ratings of research quality and aeromedical appropriateness resulted in a list of 29 measures that have a High Recommendation for use in an operational aeromedical research environment. This list includes instruments and techniques from all three types of TIA measurement. In addition, a list of 22 measures has been categorized as having a Cautious Recommendation, which means that a measure may pose some practical limitations in an aeromedical environment or is supported by research of less robust quality; caveats for measures with Cautious Recommendations (with additional support from the literature) are provided in the following discussion. Six TIA measurements are Not Recommended for use, and justification for this determination is included in this section.

TIA Measures with Strong Recommendation

Behavioral measures.

Of all the measures identified by this review, behavioral TIA measures were the highest rated for both the quality of research in which they were used as well as their appropriateness for an aeromedical environment. Behavioral measures included both observed behaviors (things that the operators did that could be collected passively by the automation software or easily observed by an investigator) and communication patterns (how the operators communicate with other team members and the automation during a mission). Most observed behavior measures (70 percent) and all communication measures identified in the literature were categorized in the High Recommendation group.

Seven distinct observable behavior measures were revealed in the analysis to be appropriate for the High Recommendation group. These measures included behavioral synchrony and entrainment, combined team performance, compliance with automation recommendations, intervention, reliance, delegation, and response time. These measures are uniquely advantageous, as they can generally be applied without disrupting cockpit operations. Behavioral synchrony and entrainment were originally developed as a measure of interpersonal trust among human teams but were modified to serve as a measure of TIA as well. In the context of TIA, behavioral synchrony and entrainment both refer to the adjustment of an operator's behavior to more closely align with or complement the recommendations and behaviors of the automation (Krausman et al., 2022). For example, an operator may adjust their technique for a repeated action to more closely mirror the technique modeled by the automation. Compliance is observed when the pilot's actions are in accordance with the system's signals, requests, or demands (Krausman et al., 2022). Generally, compliance can be operationally defined as a situation when the operator chooses to comply with an automated task or recommendation that is "correct" or enhances performance. Conversely, intervention represents the frequency of incidents when the operator overrides an automation to take manual control, even in situations where the automation makes accurate decisions (Kohn et al., 2021). The frequency of intervention incidents and the time elapsed before intervention occurs can both be used to quantify the magnitude of TIA (or perhaps more precisely, the level of distrust) demonstrated by the operator (Kohn et al., 2021). Reliance is the opposite of intervention, defined as a situation when the operator chooses not to intercede when the automation makes a decision (Kohn et al., 2021). Unlike intervention and compliance, which are active measures (i.e., the operator takes an

action based on the performance of the automation), reliance is a passive measure (the operator does not act to correct or takeover an automation). Unlike compliance, reliance can be operationally defined as the operator choosing to comply with an automated task or recommendation that is "incorrect" or degrades performance. Reliance occurs when an operators' level of TIA exceeds their level of self-confidence in performing the task themselves (Kohn et al., 2021). Compliance, intervention, and reliance can be easily measured in simulated environments where accuracy of the automation can be manipulated by investigators; however, it can also be calculated in standard flight missions if a record of automation recommendations, automation accuracy, and operator response is captured in the system. Delegation refers to an operator ceding control to automation, and signals high levels of TIA (Kohn et al., 2021; Xie et al., 2019). Unlike reliance or compliance, which also refer to an operator's tendency to accept an automated recommendation, delegation is a pre-emptive action taken by the operator to rely upon automation over the course of multiple actions/interactions. For this reason, delegation can only be measured if operators have the option within a system to give up control to the automation for a period of time (Xie et al., 2019). Lastly, response time refers to the time it takes for the operator to act following a prompt from the automation. High response times generally indicate lower TIA overall, but shorter response times only correlate with higher TIA if the operator ultimately complies with the automation (Kohn et al., 2021; Krausman, et al., 2022). Just as trust in a human co-pilot develops over time, operators' level of TIA will likewise change over time as pilots become more familiar with the technology. Each of these observable behavior TIA measures can be collected repeatedly over the course of a mission (or multiple missions) to capture patterns of growing or waning levels of TIA as operators interact with the system over time.

Communication patterns are another type of behavioral TIA measure that allows for passive data capture that does not interfere with operators' natural interactions with the automated system. While communication measures often reflect human-human interpersonal trust, aspects such as communication rates, flows, and content can be used in human-automation interactions to quantify TIA during a mission, particularly in situations where a team of multiple human operators is working together with an automated system (Krausman et al., 2022). Unlike observable behaviors, which can frequently be used in any automation-enhanced system with little or no special modification, communication measures generally require more sophisticated data capture and analysis to offer TIA insights; as a result, a number of special tools have been developed to facilitate the use of communication patterns as a continuous measure of TIA. Three such tools that are strongly recommended for inclusion in operational aeromedical environments are the Bag of Words tool; Real-time Event, Flow, and Coordination Tool (REFLECT); and Network Analysis Tool developed by the Army Research Laboratory as part of their Human-Autonomy Teaming Trust Toolkit (HAT^3; Neubauer et al., 2022). These specialized tools provide a streamlined method of measurement for all other communication-based TIA measures with a strong recommendation based on this review: top-down and bottom-up content analysis, flow, and rate. The Bag of Words tool is a means of capturing and analyzing all communication content from a mission to assist with post-hoc top-down and bottom-up thematic analysis based on content categories that can be created and modified by the research team. REFLECT provides an organized visualization of communication flow among human operators and automated systems over the course of a mission (a standardized means of collecting and analyzing flow and rate) (Neubauer et al., 2022). The Network Analysis Tool is specifically intended to derive insights about TIA using communication data. While similar to REFLECT, Network Analysis

goes beyond a clear presentation of critical communication and includes robust statistical analyses that help researchers define TIA based on complex communication networks that form among human operators and automations during a team-based mission (Neubauer et al., 2022).

Physiological measures.

Four physiological TIA measures identified in the review were included in the High Recommendation group; this included eye tracking, facial expression, blood volume pulse (BVP), and physiological synchrony. Physiological synchrony described a data analysis technique used for capturing an array of physiological measures between different subjects as a means of determining similar levels of TIA across multiple operators (while baseline physiological data varies significantly between subjects, similarities in changes and patterns in physiological data have been correlated with similar patterns of TIA) (Krausman et al., 2022). Eye tracking was the most frequently used physiological measure of TIA across the literature (n = 15) and refers specifically to tracking operators' gaze on a screen to determine where they were directing their attention (as opposed to ocular measures of physiological arousal, such as pupillometry). Using externally-mounted camera systems for collecting eye tracking data significantly improves the aeromedical appropriateness of this measure, as head-mounted tracking hardware (e.g., goggles or glasses) presents limitations when used alongside required operational gear for aeromedical crew members. Similarly, the recommendation for BVP was based upon standards in the literature that used a small wrist-worn photoplethysmography (PPG) sensor to derive data (as opposed to the finger clip sensor frequently used in clinical settings) (Krausman et al., 2022).

While the facial expression studies included in this review were both well-designed and deemed to be of sufficient quality for the purposes of this review, it should be noted that outside of the TIA literature, facial expression has often had mixed performance in attempts to use it as a measure of emotional states. While facial expressions are undoubtedly an important tool for non-verbally evaluating other people's emotional states, their complexity and reliance on a vast array of contextual elements often complicate computerized efforts to use expressions to predict emotional state accurately (Gendron et al., 2018). Since they are often quite nuanced, issues such as lighting and head position can easily create problems in correctly interpreting facial expressions (Tarnowski et al., 2017). These issues are further complicated by the fact that many efforts to use computerized recognition and classification of facial expressions are trained on an outdated theory of emotion that suggests humans only experience six "basic" emotional states: joy, sadness, disgust, anger, fear, and surprise (Ekman, 1999; Li & Deng, 2022). More recent neuroscience has discounted the concept of basic emotions in favor of a dynamic model of emotions that follow much less predictable patterns of expression (Barrett, 2017). While research within the field of TIA has demonstrated some strong support for the use of facial expressions as a physiological measure, it is possible that the actual strength of this measure may vary based on the underlying theoretical shortcomings of using computerized facial expression as a means of "basic emotion" detection. One well-established solution to this issue in the literature has been combining facial recognition with other measures in order to develop a deep neural network approach that better accounts for the complexity of emotional expression (Li & Deng, 2022). Although facial expression has a High Recommendation (rather than a Cautious Recommendation), given its history of mixed performance, it is recommended that studies that incorporate facial expression as a physiological measure of TIA with the caveat of including it as

part of an array of measures (as was the case in both facial expression studies included in this review).

User-reported measures.

The high frequency of user-reported TIA measures in studies is unsurprising given the ease and convenience of these instruments. This same ease and convenience also makes many user-reported measures well-suited to administration in an operational environment, granted that they can be administered before and/or after a mission in order to avoid distracting operators from their aeromedical tasks. Just over one-third (36.7 percent) of all the user-reported TIA measures identified in this review received a High Recommendation for use in operational aeromedical environments. For a complete list, see Figure 2. These instruments are strongly recommended based on a combination of the strength of research quality for the studies that included them, as well as brevity of the instruments and the ability to administer according to their recommended best practices prior to and/or after an automation-assisted aeromedical mission (rather than having to interrupt operators immediately after aeromedical tasks to administer the instruments during the mission).

TIA Measures with a Cautious Recommendation

Behavioral measures.

While observable behavior can be a great tool for gauging trust between teammates (whether human or automated systems), in the confines of a helicopter cockpit, not only are the behavioral indicators skewed by the limited available actions a pilot can take, but the means to record those behavioral indicators are hindered by the limited available options for camera placement. For this reason, behavioral TIA measures that require visual confirmation, namely manual verifications (when the operator checks an automated recommendation manually prior to accepting it) and proximity (how physically close to a system the operator tends to be), would not be as effective when pilots cannot move freely within the cockpit. Despite high quality ratings of research studies that included these measures, the limited physical movements afforded by an aircraft may hinder the quality of manual verifications and proximity data captured during an aeromedical mission. Manual verifications and proximity are recommended for inclusion only in the study of aeromedical tasks that are not completed within a physically confined operational space. The Economic Trust Game (sometimes called “the investment game”) is a well-established method for measuring trust in a partner (or automation) in an economic context (Berg et al., 1995). In the game, two partners (or an operator and an automated system) are paired together. The first partner (i.e., the operator) is given a certain amount of money; they are asked to invest or gift any amount of that money (even if it is zero) to their partner (i.e., the automation) and told that the researcher will triple the amount they decide to give. The second partner (automation) is then asked to give some amount of their tripled earnings back to the first partner (the operator). Trust is quantified by the amount of money the first partner chooses to give. While the Economic Trust Game may be a valid form of measuring interpersonal trust, it has limited use in studying TIA (Kohn et al., 2021) and it is not easily replicated in an operational environment. The specific economic context is not well-aligned with aeromedical operations, so attempting to initiate an Economic Trust Game within an aeromedical environment is likely to be highly distracting to the operator. Other behavioral measures are

context independent, and therefore much more suitable for a variety of operational environments.

Physiological measures.

Three physiological TIA measures were included with a Cautious Recommendation; not for any concerns related to quality of studies, but because of equipment and signal issues that impact the ability to collect the measures in an operational aeromedical environment. While simple heart rate can be captured accurately with a variety of sensors (many that pose very little interference to operators during aeromedical tasks), heart rate variability is best derived from a full ECG signal, which can pose issues related to wearing electrodes on the chest underneath the operator's uniform. ECG signals captured from sensors that do not require chest-worn electrodes often demonstrate lower levels of signal strength and reliability, which can affect the validity of the measure (Lin et al., 2014). fNIRS uses a head-worn set of sensors similar to an electrode array used in an EEG. However, unlike EEG electrodes, fNIRS sensors do not require conductive gel and have been designed with ultra-portability in mind to accommodate use in ergonomic and operational environments (Cay, 2017; Ayaz et al., 2019). Despite an attempt to make fNIRS sensors portable, most systems still require a heard-worn sensor array that interferes with operators' helmets in an aeromedical setting.

User-reported measures.

Likert scale measures of TIA were included with a Cautious Recommendation based on research quality ratings that were lower than the overall median. In fact, it is possible that the use of Likert scales had an influence on research quality ratings due to the high variability in the specific implementation of this measure and the fact that Likert scales are not often validated for their target construct, making them a less reliable subjective measure than validated instruments (Michalopoulou, 2017). Despite their high frequency of use across the studies included in this review (n = 17), the specific details of each Likert scale tended to vary significantly from study to study, with scales that offered anywhere from four to 100 options for operators to choose from. Some studies indicated that larger scales were used to improve precision (and subsequent quality) of ratings, but this assumption is not supported by several decades of research that have concluded that humans are unable to discern precision on a scale beyond seven options (a finding that has been frequently applied to other aspects of cognitive science as the "seven plus-or-minus-two" rule) (Miller, 1956). While several applications of the "seven plus-or-minus two" rule have been criticised in light of new evidence (Cowan, 2015), Monte Carlo computer simulations have repeatedly confirmed no significant improvements to reliability or validity of Likert scales beyond seven levels (Cicchetti et al., 1985; Lissitz & Green, 1975). While the use of an ordinal scale to evaluate levels of TIA can be a sufficient (and convenient) means of capturing robust data (Wei et al., 2020), in order to achieve the highest quality data, it is recommended that any Likert scale be validated alongside an existing TIA instrument before scaled use as an autonomous measure, and that scales contain no more than seven rating options.

Although they tend to offer a higher standard of validity and reliability than a Likert scale, 15 of the 29 validated user-reported TIA measures (52 percent) were rated as Cautious Recommendations, largely because of poor aeromedical appropriateness. Other than the Hypothetical Complacent Behavior measure (Merritt et al., 2015), the Implicit Association Test (Merritt et al., 2013), the Negative Attitudes Toward Robots Scale (Nomura et al., 2006), and the

Propensity to Trust Machines Scale (Merritt & Ilgen, 2008) (which are included as Cautious Recommendations due to the low quality ratings of the studies they were used in), each of the validated instruments included as Cautious Recommendations were determined to have low levels of aeromedical appropriateness. This is a result of their implementation requiring operators to respond to questions during or immediately following automation-assisted tasks; this "real-time" operator input interrupts the flight task at hand and introduces additional cognitive workload that can impact performance (and subsequently impact TIA itself). To avoid distracting interruptions from the tasks, survey data from these instruments can be collected prior to or immediately after an aeromedical mission scenario that incorporated automation-assisted tasks. For the instruments that are included as Cautious Recommendations due to low ratings of research quality, if they are to be included in aeromedical research, it is recommended that they be used alongside other validated TIA measures rather than as the sole measure of operator TIA to better evaluate their validity as part of a multimodal array.

TIA Measures that are Not Recommended

Physiological measures.

Physiological TIA measures that are Not Recommended for an operational aeromedical environment included EEG, fMRI, and skin conductance. All three of these measures were derived from studies where research quality was rated below the overall median for all studies included in the review (see Table 2), calling into question the reliability of their effectiveness across conditions. These measures also presented significant logistical challenges that would be difficult or impossible to overcome in an operational aeromedical environment. In the case of fMRI, no ambulatory options currently exist, rendering it impossible to use within an aircraft. EEG and skin conductance both have some limited options for ambulatory data collection, but still frequently require electrodes placed on the scalp and forehead (EEG) or hands (skin conductance) to achieve sufficient signal strength. Given the limitations for quality electrode placement on the head and hands during flight operations in an aircraft, these measures were rated as having low aeromedical appropriateness.

User-reported measures.

Three user-reported instruments were determined to be poor fits for aeromedical research and have been classified as Not Recommended. These include the Godspeed Measure (Bartneck et al., 2009), the Multitasking Preference Inventory (Poposki & Oswald, 2010), and a human-computer trust measure (Yan et al., 2011). In addition to their use in studies that were determined to have research quality below the sample median, these instruments were also deemed to be misaligned with aeromedical tasks, automations, and environments.

Limitations

While all efforts were made to conduct as thorough a review as possible, this analysis does not include an exhaustive account of all research published on TIA. As new studies emerge, the findings of this review may not reflect the current state of the science. Given the heterogeneity of study designs included in this review, some TIA measures were used as singular measures, while more frequently, measures were used as part of a multimodal array. For this

reason, scientific quality refers to the quality of the study using each of the measures rather than the quality of any singular instrument. The purpose of this review was to identify which of the many available measures of TIA showed promise for use in an operational aeromedical environment. Recommendation ratings reflect the potential for use in an aeromedical setting and are not intended to be interpreted as a rating of the instrument's general quality as a measure of TIA. Future studies should expand upon these recommendations by conducting tests of feasibility and validity of the measures in an actual operational setting to determine how each instrument performs in an aeromedical context.

Conclusions

Although TIA is a complex construct that incorporates behavioral, physiological, and psychosocial elements, great advances have been made over the past decade to create valid and reliable measures of TIA across multiple domains. Given the unique environmental barriers that exist for aeromedical environments, and the critical role of TIA in technology systems that require a growing reliance on automation, it is essential that special consideration is given to determine which emerging TIA measures can be feasibly and reliably administered in an operational aeromedical environment. This review provides strong recommendations for 28 TIA measures with high potential of success in aeromedical settings, as well as 23 measures that should be used cautiously with expected caveats and six measures that are not recommended for use in aeromedical environments. Future research can continue to build upon these recommendations by testing each of these measures (and those that emerge) to evaluate their feasibility of administration and psychometric fidelity in operational aeromedical settings.

References

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Prentice-Hall.

Ayaz, H., Izzetoglu, M., Izzetoglu, K., & Onaral, B. (2019). The use of functional near-infrared spectroscopy in neuroergonomics. In *Neuroergonomics* (pp. 17-25). Academic Press.

Balters, S., & Steinert, M. (2017). Capturing emotion reactivity through physiology measurement as a foundation for affective engineering in engineering design science and engineering practices. *Journal of Intelligent Manufacturing, 28,* 1585–1607.

Barber, P. (2003). Critical analysis of psychological research II: delivering a course for inclusion in the core curriculum for psychology. *Psychology Learning and Teaching*, *3*(1), 15–26.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain.* Houghton Mifflin Harcourt.

Bartneck, C., Croft, E., Kulic, D., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*(1), 71–81.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142.

Cay, G. (2017). *Design of a wearable fNIRS neuroimaging device with an internet-of-things architecture* [Master's thesis, University of Rhode Island]. Digital Commons @ URI.

Chancey, E. T., Bliss, J. P., Proaps, A. B., & Madhavan, P. (2015). The role of trust as a mediator between system characteristics and response behaviors. *Human Factors, 57*(6), 947–958.

Chien, S. Y., Lewis, M., Sycara, K., Liu, J. S., & Kumru, A. (2018). The effect of culture on trust in automation: Reliability and workload. *ACM Transactions on Interactive Intelligent Systems (TiiS), 8*(4), 1–31.

Cicchetti, D. V., Shoinralter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater relaibility: A Monte Carlo investigation. *Applied Psychological Measurement, 9*(1), 31–36.

Cowan, N. (2015). George Miller's magical number of immediate memory in retrospect: Observations on the faltering progress of science. *Psychological Review, 122*(3), 536–541.

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., & Yanco, H.. (2012). Effects of changing reliability on trust of robot systems. *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 73-80).

Dolgov, I., & Kaltenbach, E. K. (2017). Trust in automation inventories: An investigation and comparison of the human-vomputer trust and trust in automated systems scales. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*(1).

Drnec, K., Marathe, A. R., Lukos, J. R., & Metcalfe, J. S. (2016). From trust in automation to decision neuroscience: Applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction. *Frontiers in Human Neuroscience, 10*, 290.

Ekman, P. (1999). Basic emotions. In T. Dalgleish, & M. Power, *Handbook of Cognition and Emotion* (pp. 45-60). Wiley and Sons.

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality, 42*(6), 1585–1593.

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. *International Symposium on Collaborative Technologies and Systems* (pp. 106-114). Institute of Electrical and Electronics Engineers.

Gendron, M., Crivelli, C., & Barrett, L. (2018). Universality reconsidered: Diversity in making meaning of facial expressions. *Current Directions in Psychological Science, 27*(4), 211–219.

Gibson, A. M., Capiola, A., Alarcon, G. M., Lee, M. A., Jessup, S. A., & Hamdan, I. A. (2023). Construction and validation of an updated perfect automation schema (uPAS) scale. *Theoretical Issues in Ergonomics Science, 24*(2), 241–266.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434.

Jackson, K., Prasov, Z., V. E., & Jones, E. (2016). A heuristic based framework for improving design of unmanned systems. *Proceedings of the Human Factors and Ergonomics Society 2016 Annual Meeting. 60*, pp. 1696-1700. Sage Publications.

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiloa, A. (2019). The measurement of the propensity to trust automation. *International Conference on Human-Computer Interaction, 11575*, 476–489.

Jian, J.-Y., Bisanta, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53–71.

Johns, J. (1996). A concept analysis of trust. *Journal of Advanced Nursing, 24*, 76–83.

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y. C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology, 12*, e604977.

Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018), 6,* pp. 13-30. Springer International Publishing.

Krausman, A., Neubauer, C., Forster, D., Lakhmani, S., Baker, A. L., Fitzhugh, S. M., & Schaefer, K. E. (2022). Trust measurement in human-autonomy teams: Development of a conceptual toolkit. *ACM Transactions on Human-Robot Interaction (THRI), 11*(3), 1–58.

Lee, B. C., Chung, K., & Kim, S.-H. (2018). Interruption cost evaluation by cognitive workload and task performance in interruption coordination modes for human-computer interaction tasks. *Applied Sciences, 8*(10), 1780.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*(1), 153–184.

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

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243–1270.

Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13,* pp. 1195–1215.

Lin, W. H., Wu, D., Li, C., Zhang, H., & Zhang, Y. T. (2014). Comparison of heart rate variability from PPG with that from ECG. *The International Conference on Health Informatics: ICHI 2013* (pp. 213-215). Springer International Publishing.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: a Monte Carlo approach. *Journal of Applied Psychology, 60*(1), 10–13.

Lu, Y., & Sarter, N. (2019). Eye tracking: A process-oriented method for inferring trust in automation as a function of priming and system reliability. *IEEE Transactions on Human-Machine Systems, 49*(6), 560–568.

Lyons, J.B. & Stokes, C.K. (2012). Human-human reliance in the context of automation. *Human Factors, 54*(1), 112–121.

Madhavan, P. (2015). Perception of trust in automation. In R. R. Hoffman (Ed.), *The Cambridge Handbook of Applied Perception Research* (pp. 488-509). Cambridge University Press.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *11th Australian Conference on Information Systems, 53,* pp. 6-8. Australian Association for Information Systems.Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology, 84*(1), 123–136.

Mayer, R., Davis, J., & Schoorman, F. (1995). An integrative model of organizational trust. *The Academy of Management Review, 20*(3), 709–734.

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors, 53*(4), 356–370.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors, 50*(2), 194–210.

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., & Shirase, L. (2015). Automation-induced complacency potential: Development and validation of a new scale. *Human Factors, 57*(5), 740–753.

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., & Shirase, L. (2019). Automation-induced complacency potential: development and validation of a new scale. *Frontiers in Psychology, 10,* 225.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward autmoation on trust in an automated system. *Human Factors, 55*(3), 520–534.

Michalopoulou, C. (2017). Likert scales require validation before application: Another cautionary tale. *Bulletin of Sociological Methodology, 134*(1), 5–23.

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., & Ju, W. (2016). Behavioral measurement of trust in automation: the trust fall. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 60*, pp. 1849-1853. SAGE Publications.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Moorman, C., Deshpande, R., & Zaltman, G. (1993). Factors affecting trust in market-research relationships. *Journal of Marketing, 57*(1), 81–101.

Muir, B., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*(3), 429–460.

National Aeronautics and Space Administration. (2016). *MATB-II overview*. https://matb.larc.nasa.gov/

Nehme, C. E. (2009). *Modeling human supervisory control in heterogeneous unmanned vehicle systems.* Massachusetts Institute of Technology, Department of Aeronautics and Astronautics.

Neubauer, C., Baker, A. L., Fitzhugh, S. M., Kellihan, B., Jagielski, J., & Krausman, A. S. (2022). *Human-autonomy teaming trust toolkit (HAT3) executive summary* [Report No. ARL-TR-9622]. DEVCOM Army Research Laboratory.

Neubauer, C., Gremillion, G., Schaefer, K. E., Perelman, B. S., La Fleur, C., & Metcalfe, J. S. (2020). *Analysis of facial expressions: Explaining affective state and trust-based decisions during interaction with automation* [Report No. ARL-TR-8945]. DEVCOM Army Research Laboratory.

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies, 7*(3), 437–454.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making, 2*(2), 140–160.

Poposki, E. M., & Oswald, F. L. (2010). The multitasking preference inventory: Toward an improved measure of individual differences in polychronicity. *Human Performance, 23*(3), 247–264.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality, 35*(4), 651–665.

Sanders, T. L., Kaplan, A. D., MacArthur, K., Volante, W. G., & Hancock, P. A. (2021). Trust and human factors. In *The Neurobiology of Trust* (pp. 77-98). Cambridge University Press.

Schaefer, K. (2016). Measuring trust in human robot interactions: Development of the "Trust Perception Scale-HRI". In R. Mittu, & D. Sofge, *Robust Intelligence and Trust in Autonomous Systems* (pp. 191-218). Springer.

Schaefer, K. E., Sanders, T. L., Yordon, R. E., Billings, D. R., & Hancock, P. A. (2012). Classification of robot form: Factors predicting perceived trustworthiness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56*, pp. 1548-1552. SAGE Publications.

Schneider, T. R., Jessup, S. A., Stokes, C., Rivers, S., Lohani, M., & McCoy, M. (2017). The influence of trust propensity on behavioral trust. *Meeting for the Association for Psychological Society.*

Semsar, A., & Nazari Shirehjini, A. (2017). Multimedia-supported virtual experiment for online user–system trust studies. *Multimedia Systems, 23*, 583–597.

Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation-induced "complacency": Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology, 3*(2), 111–122.

Tarnowski, P., Kolodziej, M., Majkowski, A., & Rak, R. (2017). Emotion recognition using facial expressions. *Procedia Computer Science, 108C*, pp. 1175-1184. International Conference on Computational Science.

Ullman, D., Aladia, S., & Malle, B. F. (2021). Challenges and opportunities for replication science in HRI: A case study in human-robot trust. *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 110-118).

Wei, J., Bolton, M. L., & Humphrey, L. (2020). The level of measurement of trust in automation. *Theoretical Issues in Ergonomics Science, 22*(3), 274–295.

Wojton, H. M., Porter, D., Lane, S., Bieber, C., & Madhavan, P. (2020). Initial validation of the trust of automated systems test (TOAST). *The Journal of Social Psychology, 160*(6), 735–750.

Xie, Y., Bodala, I. P., Ong, D. C., Hsu, D., & Soh, H. (2019). Robot capability and intention in trust-based decisions across tasks. *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (pp. 39-47).

Yagoda, R. E., & Gillan, D. J. (2012). You want me to trust a ROBOT? The development of a human-robot interaction trust scale. *International Journal of Social Robotics, 4*, 235–248.

Yan, Z., Kantola, R., & Zhang, P. (2011). A research model for human-computer trust interaction. *Trust, Security, and Privacy in Computing and Communications (TrustCom)* (pp. 274-281). IEEE.

Appendix A. Research Quality and Aeromedical Appropriateness Rating Checklists

Table A1. JORIPIEF-Based Research Quality Checklist Used by the Reviewers

Quality criteria	**Not at all**	**Poorly**	**Sufficiently**	**Well**	**Very well**
A case has been made for conducting the investigation reported.	1	2	3	4	5
The study hypotheses were clearly reported.	1	2	3	4	5
Sufficient detail has been provided for replication of the study to be done.	1	2	3	4	5
The internal logic of the research design is satisfactory.	1	2	3	4	5
The results have been analyzed appropriately, correctly, and thoroughly.	1	2	3	4	5
There are no serious flaws in the presentation of the paper (e.g., language, formatting).	1	2	3	4	5
The data have been presented effectively and in sufficient detail.	1	2	3	4	5
The data and analyses were interpreted satisfactorily and correctly.	1	2	3	4	5
The findings can be sufficiently generalized.	1	2	3	4	5
The scientific importance of the study has been established.	1	2	3	4	5
The findings achieve progress with the topic matter.	1	2	3	4	5
The study competently achieves its declared objectives.	1	2	3	4	5
There are no grounds for doubting the decision to accept the article for publication (given the circumstances that applied when it was written).	1	2	3	4	5
			Total of all rating scores		
			Overall quality rating (Total / 13)		

Table A2. Aeromedical Appropriateness Checklist Used by the Reviewers

Aeromedical appropriateness criteria	**Not at all**	**Poorly**	**Sufficiently**	**Well**	**Very well**
This measure/instrument physically fits within the available space of the aircraft.	1	2	3	4	5
This measure/instrument can be administered according to its instructions/best practices without physically disrupting or damaging the aircraft.	1	2	3	4	5
This measure/instrument can be administered according to its instructions/best practices without distracting crewmembers from their mission tasks.	1	2	3	4	5
This measure/instrument does not require special equipment that must be installed in the aircraft.	1	2	3	4	5
The quality of the data collected from the measure/instrument is not affected by the aircraft.	1	2	3	4	5
The quality of the data collected from the measure/instrument is not affected by the crewmember focusing on their mission tasks.	1	2	3	4	5
			Total of all rating scores		
			Overall appropriateness rating (Total / 6)		

Appendix B. Search Terms and Hit Frequency

Table B1. Number of Hits for Specific Searches During Initial Data Collection

Database	Search terms	Number of hits
Google Scholar	“trust in automation” “trust measures”	608
	“trust in automation” “trust measures” “medical”	238
	“trust in automation” “trust measures” “aviation”	126
	“measure trust in automation”	228
	“measure trust in automation” “medical”	85
	“measure trust in automation” “aviation”	58
EBSCO	measure trust in automation medical aviation (+2013-2023)	1973
	“trust in automation” “trust measures”	133
	“trust in automation” “trust measures” "medical"	15
	“trust in automation” “trust measures” "aviation"	8
DTIC	“measure trust in automation”	10
	“trust in automation” “trust measures”	41
	“trust in automation” “trust measures” “aviation”	20
	“trust in automation” “trust measures” “medical”	18
	Total hits	3561
	Total titles included	207
	Total abstracts included (non-duplicates)	159
	Total full-text articles included	**31**

Appendix C. Lists of User-report and Behavioral Measures with Frequencies.

Table C1. List of Validated Instruments and Frequency of Use Throughout the Reviewed Literature

Validated instruments	Frequency
Checklist for Trust between People and Automation (Jian et al., 2000)	11
Complacency-Potential Rating Scale (Singh et al., 1993)	5
Human Computer Trust Scale (Madsen & Gregor, 2000)	5
Propensity to Trust Scale (Mayer & Davis, 1999)	4
Propensity to Trust Machines Scale (Merritt & Ilgen, 2008)	3
Integrated Model of Trust Scale (Muir & Moray, 1996)	3
Trust Perception Scale for Human-Robot Interaction (Schaefer, 2016)	3
Cross-Cultural Automation Trust Scale (Chien et al., 2014)	3
Human-Robot Interaction Trust Scale (Yagoda & Gillan, 2012)	3
System Trustworthiness Scale (Schaefer et al., 2012)	2
Interpersonal Trust Scale (Rotter, 1967)	2
Propensity to Trust Technology Scale (Schneider et al., 2017)	2
Perfect Automation Schema (Gibson et al., 2023)	2
Draper Trust Scales (Jackson et al., 2016)	2
Implicit Association Test (Merritt et al., 2013)	2
Trust Scale (Merritt, 2011)	2
Measures of Trust & Trustworthiness (Mayer & Davis, 1999)	1
Trust in Automation Scale (Körber, 2019)	1
Trust & Self-Confidence Scale (Lee & Moray, 1994)	1
Multitasking Preference Inventory (Poposki & Oswald, 2010)	1
Adapted Propensity to Trust Questionnaire (Jessup et al., 2019)	1
Negative Attitude towards Robots Scale (Nomura et al., 2006)	1
Trust Scale (Lee & Moray, 1992)	1
Operator Tendency to Trust Automation (Merritt & Ilgen, 2008)	1
Trust in Automated Systems Test (TOAST) (Wojton et al., 2020)	1
Dynamic Reporting of Trust (Desai et al., 2012)	1
Yan Trust Measure (Semsar & Nazari Shirehjini, 2017)	1
Godspeed Measure (Bartneck et al., 2009)	1
Hypothetical Complacent Behavior (Merritt et al., 2019)	1
Muir and Moray Trust Scale (4-item) (Muir & Moray, 1996)	1
Propensity to Trust Survey (Evans & Revelle, 2008)	1
Muir and Moray Trust Scale (8-item) (Muir & Moray, 1996)	1
Multi-Dimensional Measure of Trust (MDMT) (Ullman et al., 2021)	1

Table C2. List and Description of Observable Behaviors and Frequency of Use Throughout the Reviewed Literature

Behavioral measures	Frequency	Description
Compliance with automation recommendations	11	The operator uses recommendations given by the automation
Reliance	5	The operator makes no attempt to override the automation (even when mistakes are made)
Response time	5	How long it takes the operator to act after a prompt or alert from the automation
Manual verifications	4	The operator confirms the accuracy of an automation before taking action
Economic Trust Game: Stakes invested	3	How much the operator is willing to risk/bet on an automation's accuracy (often reported as a monetary amount)
Delegation	3	Allowing automation to handle a task when the operator could do it manually
Intervention	3	Overriding automation/taking manual control (even when the automation is accurate)
Behavioral synchrony and entrainment	3	The operator begins to model behaviors or communication styles similar to the automation
Combined team performance	3	Objective scored performance on the automation-assisted task being completed by the operator
Proximity	2	The operator's physical closeness to an automation technology (used most often with robots)
Communication: Bag of Words	1	Computerized language model that analyzes sematic content of communications between the operator and automation
Communication: Bottom-up measures of communication content	1	Similarity of communication tone between the operator and automation
Communication: Flow	1	How often the operator and automation are communicating with each other
Communication: Network analysis	1	A computer modeling output of the communication flow and rate between the operator(s) and automation
Communication: Rate	1	The frequency and duration of interactions between the operator and automation
Communication: Real-Time Event, Flow, and Coordination Tool (REFLECT)	1	Computerized tool for visualizing the flow, content, and event coordination between operators and automations during a task
Communication: Top-down measures of communication content	1	Similarity of content being discussed between the operator and automation

U.S. ARMY AEROMEDICAL RESEARCH LABORATORY

FORT NOVOSEL, ALABAMA

Optimizing

HUMAN PROTECTION AND PERFORMANCE

since 1962

All of USAARL's science and technical informational documents are available for download from the Defense Technical Information Center.
https://discover.dtic.mil/results/?q=USAARL

U.S. ARMY

FUTURES COMMAND

MRDC