American Psychological Association

Technology, Mind & Society
2021 Conference Proceedings

ISBN:

http://dx.doi.org/10.1037/xxxxxxx

Applying Freeze Off-line Probe Technique for Coding Human Machine Trust Behaviors

Katharine Woodruff[1] , Gianna Avdic McIntire[1], Peggy Wu[2], Bret Israelsen[2], Patrick "Dice" Highland[3], Glenn Taylor[4], Colton Thompson[3], Mathew Cover[3], Thomas "Mach" Schnell[3]

[1]Collins Aerospace
[2]Raytheon Technologies Research Center
[3]The University of Iowa, Operator Performance Laboratory
[4]SoarTech

As autonomy becomes more complex, it is increasingly important to measure the amount of trust exhibited by a human towards a machine, regardless of the level of trustworthiness of the machine. There are several challenges associated with current methods that rely primarily on self-report subjective ratings: 1) self-reports are dependent on the self-awareness of subjects and may not be reliable, 2) low sampling rates may not result in an adequate level of granularity for analysis for high tempo, highly dynamic scenarios, and 3) in high risk scenarios like air-to-air combat, subjects may be cognitively overloaded and unable to use the think aloud protocol, or their surveys may simply not be feasible or practical. To overcome these challenges, we were inspired by the "freeze on-line" approach of the Situation Awareness Global Assessment Technique (SAGAT) and modified it for use in an off-line video-based approach to avoid the high logistics costs of performing this assessment during flight or while in a flight simulation. Subjects were first recorded during flight. In an extended post-hoc video replay, time was frozen during predetermined events, and subjects were asked to report their thought process at the time. Their comments and other behaviors observed in the video were then recorded and coded by study staff after the debrief. We report on the development of the codebook and the resulting Inter Rater Reliability calculations.

Keywords: Objective Measures, Human Machine Trust, Behavior Coding

Autonomous systems continue to advance in their roles as task supporters to task executers. This has led to changes in the human-autonomy relationship. The roles have shifted to a dynamic team-based relationship which has caused a redistribution in task responsibilities (Madhavan & Wiegmann, 2007). This dynamic relationship has made it necessary to measure factors that influence the performance of the team. Analogous to human-human teams, trust has emerged as an important element of human-machine heterogeneous teams. Trust has been 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). At this point, the human is the only agent who can intervene and take over controls from the other party. Therefore, it is up to the human to dictate whether both agents are properly tasked. This optimized relationship occurs when both parties are performing tasks within their capabilities. The human will need to monitor the autonomy and take over controls when the situation exceeds the autonomy's abilities. If a human does not intervene during poor machine performance, they are *over trusting*, and performance will suffer. On the other hand, if a human intervenes while the autonomy is in a context where it can outperform the human, the operator is *under trusting* the autonomy and the overall human+machine team performance will suffer. There exists an appropriate level of trust based on the capabilities of the autonomy. However, the human operator forms a mental model of trust based on factors including, but not limited to the machine's reliability and competence. For this reason, it is important to be able to objectively measure the human's exhibited trust. A real time measure of the human's exhibited trust would allow for corrective actions, either off-line in the design of the autonomy, or on-line through interventions.

This article was published [to be completed by publisher].
The authors acknowledge the following people: Dr. Anna Skinner (DARPA) and Dr. Joseph Lyons (AFRL).
This research was supported in part by the Defense Advanced Research Projects Agency (DARPA Air Combat Evolution program, Technical Area 2 "build and calibrate trust in air combat local behaviors", Soar Technology, Inc contract #FA8650-20-C-7044. This study was awarded by the Air Force Research Laboratory (AFRL). The authors have the following conflict of interest to disclose: none. Distribution A: Approved for Public Release, Distribution Unlimited.

Correspondence concerning this article should be addressed to Katharine Woodruff, Collins Aerospace, Cedar Rapids, IA, 52498, USA. Email: katharine.woodruff@collins.com

Currently, trust is primarily measured through subjective responses pre and post hoc data collection. Subjective responses have been validated as a reliable source for measuring trust (Schaefer, 2016). One downfall to subjective responses is that they must either be administered before or after a task is completed. Therefore, they lack the granularity of fluctuations in trust that may occur throughout a task. This is insufficient for tasks that occur in complex and dynamic environments where trust levels may significantly vary within seconds. One example of such a complex task is Within Visual Range (WVR) Air-to-air Combat Maneuvering (ACM), also called "dogfighting".

The method in this paper was developed as part a research program where the overall goal was to objectively measure and calibrate the trust of a pilot in an autonomous co-agent during ACM. ACM is the art of maneuvering an aircraft in combat to gain a positional advantage over another aircraft. ACM is challenging as the "solution" depends on many factors such as energy states, vehicle performance capabilities, relative orientation of vehicles, and locations of both aircraft in the scenario. High granularity of trust levels is needed to calibrate the trust of the human to optimize performance of the human-autonomy interaction. One option to gain higher granularity is through the Experience Sampling Method (ESM) (Larson & Csikszentmihalyi, 2014). The method consists of participants periodically self-reporting to gain understanding on what people feel, do, and think throughout activities. However, this is not plausible for ACM because it would require too much task interruption and any data captured would not be representative of a naturalistic environment. To simulate ESM post-hoc, a video replay debrief was conducted after experimental data collection where the pilots were asked questions about their thought processes and significant events throughout a scenario. The comments of the pilots were recorded in-line with the audio and video data collected from the experiment. This paper describes the development and validation of the codebook that translates the qualitative comments into quantitative data. The codebook developed in this study serves as a steppingstone between subjective surveys and real-time objective classification of trust using pilot physiology and behaviors

Background

In this project we leverage the definitions and models of trust from Lee and See (2004) and Hoff and Bashir (2015). Lee and See (2004) defines trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" Trust is a situational multidimensional latent construct that is influenced by a plethora of factors. Lee and See (2004) defined three dimensions of trust: analytical, affective, and analogical processes.

Analytical Trust. Analytical trust is a result of rational analysis of own ability versus autonomy's ability to achieve the individual's goal. This is similar to Hoff and Bashir's definition of "situational trust". Analytical trust is influenced by the system performance, perceived risks/benefits, mood, the framing of the task and attentional capacity.**Analogical Trust.** Analogical trust is based on trust of analogical systems or other's trust of the system (the reputation). We believe this to be related to the learned trust component of Hoff and Bashir (2015) and extend Lee and See's (2004) definition of analogical trust to include the operator's own prior experience with the same system. This dimension of trust could be formed a-priori but is also adjusted during the course of an interaction. The initial analogical trust is based on prior knowledge of a system from others, or one's own prior experience with similar systems. The dynamic analogical trust is developed through interaction with the system. An operator will adjust their mental model of a system as they interact with it. With each failure or success, the operator will tweak the perceived capabilities of the system. Analogical trust is influenced by the predictability of a system, prior knowledge or a system, perceived risk of the situation, type and timing of errors, and validity of decisions or actions.

Affective Trust. Affective trust is the willingness to rely on automation, moderated by workload and affect. Hoff and Bashir (2015) define this as dispositional trust and it is influenced by the personality of the operator, the experience and expertise of the operator, the self-confidence, and background of an operator.

Although we are able to identify factors that affect trust, there is still little known about how these factors affect trust as parts and as a whole.

In the past, trust has been measured through subjective questionnaires and surveys. Some factors of trust are more easily measured than others. For example, dispositional trust can be captured with surveys before any experimentation has been done. This can be captured through formal scales such as the Mini IPIP (Goldberg, 1999) and/or through background and experience questions. The disposition of a person is somewhat unchanging therefore it can be measured before any exposure to a system. But the learned trust and situational trust are more dependent on the experimentation and exposure to an autonomous system. One way to capture this data without interrupting a task is by asking survey questions related to the influencing factors after a task is complete. Such survey questions should inquire about scenario focused factors such the Bedford Workload scale (Roscoe & Ellis, 1990) and the Situational Awareness Rating Technique (SART) (Taylor, 1990). However, this only gives the summary of a task's trust components. Throughout a task there are different levels of trust based on the behavior of the autonomy and the complexity of an action or decision. Referring back to the definition of trust, vulnerability can be affected by either the risk of a situation or the complexity of a task. As a dogfight develops the risk of the situation changes and the complexity of maneuvering fluctuates. Because of the speed of the aircraft, changes in situational risk occurs within seconds. Therefore, trust is expected to vary throughout a dogfight. These different levels are not captured with post hoc survey questions and pilots are likely to respond with summaries of a dogfight. Thus, it is necessary to capture these differences as a scenario develops is necessary to understand how the actions of an autonomous system influence the trust levels of a pilot. One way to capture these fluctuations is through the *freeze on-line technique* developed by Endsley (1995). This technique was originally developed to test the situational awareness of a subject throughout a task but has since been implemented in numerous experiments to

test other performance measures of humans such as decision making, attentional capacities, and trust (Gevins, et. al., 1997, Klein, 2017, Pashler, 1999). The technique consists of a subject performing a task with periodic interruptions where the experimenter pauses the task, then asks the subject questions about the relevant factor(s) being tested. This method has been validated with an air-to-air combat battle management task (Endsley, 1995). However, the extremely dynamic nature of ACM does not allow for interruptions of a task without influencing the naturalistic environment. Moreover, conducting the freeze on-line technique during flight would use jet fuel at an inefficient rate. ACM does not have objectively "good" or "bad" maneuvering without context. For example, a maneuver may be perfectly good in one scenario but extremely bad in another based on the behavior and location of the opponent. Moreover, the validity of a maneuver may be influenced by the situational awareness, workload, and background of a pilot. For example, a pilot may perceive a situation to be fine if their attention is on another task and are missing key details. Likewise, pilots develop their own style of dogfighting, therefore perception of a situation can vary from pilot to pilot. For this experiment, we adapted the *freeze on-line technique* to become a *freeze off-line technique*. We asked the pilot to fly the test cards of the experiment and then replayed the video of the testcard. We developed a set of predefined selection criteria for pause points representative of significant events (e.g. pilot reaching for the override button, pilot performing a visual search outside the cockpit, pilot first noticing the enemy, within shooting distance of the enemy). At each pause point we asked the pilot for their perception of the autonomy and situation in various forms such as risk, benefit, trust levels, errors, etc. The conversations were guided by factors that influenced analytical and analogical trust. A researcher would then tag these significant events and comments in line with the audio and video. In interest of debrief time, we recorded qualitative comments recorded as opposed to having the pilot numerate each of the factors of interest. For example, after the autonomy performs a lead turn the pilot may describe it as "…it maneuvered in an expected way with a sufficient turn rate." The video replay provided multiple benefits. We gained insight into the thought processes of the pilot and their perception on the realism and face validity of our experimental design. However, the primary goal of this process was to help guide and develop our objective models of trust. Therefore, it was essential that we can consistently quantify the debrief qualitative data. We coded observations for use in analysis with objective data that was collected during the test card (i.e. physiological data, eye gaze, and interactions with avionics). In summary, the codebook, which we describe below, was used to quantify these comments in a systematic manner that then could be analyzed for their influence on trust with a higher level of temporal granularity.

Method

Here we describe the development of the codebook and a procedure to validate the coding through interrater reliability. We identified significant events and traits by extracting components from validated subjective measures that were relevant to the design of experiment scenarios. We give a brief description of the experimental tasks for background and context.

Materials. The simulator was a fixed based simulator with a stick and throttle that followed the aircraft performance of an L29 Delfin aircraft. All visuals were displayed and recorded in a virtual reality headset using Preprar3D™ and the Next Generation Threat System (NGTS). Video of the pilot in the simulator and flight state data were recorded. The pilot was instrumented with the following physiological sensors: 1. three wired electrodes on the chest measuring EKG (Martin et. al., 2019), 2. Empatica E4 wrist bracelet measuring blood volume pulse, 3-axis accelerometer, skin temperature, and galvanic skin response (GSR) (Garbarino, 2014), and 3. Head tracking and eye tracking using the Vive Pro Eye headset (Ogden, 2019).

Participants. All participants were male Air Force or Navy Fighter pilots between the ages of 20-60. Participants were recruited through email. There were 17 pilots with prior air combat experience.

Design. Subjects went through one day of experimental data collection. Upon arrival, subjects completed initial subjective surveys and received a briefing on the experiment. After the briefing the pilots were instrumented with physiological sensors before entering the simulator. Then, the pilots familiarized themselves with a Battle Management Task (BMT) before going through the experimental test cards. The BMT acted as a distractor task to monitoring the autonomy, which was flying the dogfight. Lastly, after the test cards were completed, the pilots participated in the video replay debrief.

There were 13 test cards in an experiment run. The first 4 test cards were measuring baseline physio responses. Test card 5-13 were the experimental test cards where the pilot would perform the BMT while monitoring the autonomous pilot ("autonomy") perform the ACM. The scenarios started with the enemy or the "bandit" appearing about 20 miles away from the pilot's own ship at 180-degree Angle Off Tail (AOT)[2]. The autonomy of the own ship maintained its heading and flew towards the bandit. This path was consistent until the two aircraft are 1-2 miles away from each other and about to pass across each other's canopies; this is referred to as the "merge". The autonomy may start performing the "lead turn" at the merge. If done successfully (with a proper timing and turn rate), the lead turn allows the own ship to gain a positional advantage behind the bandit. In this experiment, the lead turn was performed in one of three manners, 1: high reliability autonomy and successfully if no pilot intervention, 2: low reliability autonomy and would end in a stalemate (a stalemate was defined as when the own ship well outside of the bandit's weapon engagement zone after 5 minutes) if no pilot intervention, and 3: unsuccessfully and own ship would be destroyed if the pilot did not intervene with a "paddle" override. In this study, paddle represented a pilot intervention (e.g., decision and action) to deliberately disengage the autonomy and take over the dogfighting task. Once the own ship was positioned behind the bandit 1 mile away in the 30-degree cone

[2] Angle Off Tail refers to the angle that is created between the tails of two aircraft paths.

of the nose, it would be considered to be in the Weapon Employment Zone (WEZ) as this was a "guns only engagement". Once the own ship was in the WEZ the pilot would call a kill (e.g., a pilot would verbally call out "guns"), which signified a successful mission completion.

Codebook Development. The codebook was developed by leveraging existing validated trust measures and translating them into a set of behavioral measures that were operationalized to the program scenario. We selected components from the HRI Trust Perception Scale (Schaefer, 2016) and the factors identified in (Hoff & Bashir, 2015). The HRI trust scale was developed to measure human-robot trust perceptions over time as an overall percentage. The scale consists of 40 items and has a condensed version consisting of 14 items. We selected items that were relevant to our experiment design and the autonomy the pilot was interacting with. Next, we examined the trust factors that (Hoff & Bashir, 2015) identified and selected task relevant items. Each of those scales consist of over 20 components, therefore there was a need to condense these scales in the interest of time. Down-selection was based on the task parameters, hypotheses support, and subject matter expert (SME) conversations about the nature of dogfights. One of the hypotheses in the program is that the cross-check ratio of a pilot will be indicative of trust levels. The cross-check ratio is the ratio of time that a pilot spends monitoring the autonomy instead of the displays related to the distractor task. Autonomy-related information was shown in the tactical situational display (TSD), the heads-up display (HUD), and outside the canopy/window. The TSD displays the 2D locations of the own ship and the bandit on a screen in the cockpit. The BMT screen is where the pilots are performing the distractor task of battle managing. The HUD displays statistics about the own ship flight state data such flight path angle, angle of bank, airspeed and altitude through a combining glass in the cockpit. Pilots are trained to make "cross-checks" or scans of the different screens and displays within a cockpit, as well as outside the window, to help maintain situational awareness (Brown et. al., 2002). We hypothesized that one indicator of decreased trust in the autonomy is increased attention, and that attention manifests as the pilot looking at the TSD, HUD, and the outside world more than the BMT. Because of this hypothesis, the codebook also included prompts for when the pilot was looking at the TSD. The are trust scales that we use and have been validated but were not adapted well to the freeze on-line technique, so we made adaptations for it to be practical.

Each category in the codebook was rated based on the rater's interpretations of subjective data tags obtained through subject video replay debrief and in accordance with codebook definitions. All categorical variables were transformed into quantitative ordinal variables prior to the IRR calculations. The following categories and significant events were identified as relevant to the ACM scenarios.

Distance Threshold. This category indicated when the bandit was within a distance that increased the risk of the situation based on the parameters of the task. Pilots would indicate at which range they started paying more attention to the TSD based on the distance of the bandit and not because of the behavior of the autonomy. This was usually when the bandit was 3-5 miles away or at the merge where pilots are able to gain visuals of the bandit in the HUD (Highland et. al., 2020). Distance threshold is a categorical variable where 1 indicates the event. This category was developed from SME conversations and task parameters.

Error Type. This category was an adaptation of the category "Error Type" from (Hoff & Bashir, 2015) to indicate what types of error a pilot perceived the autonomy to make. Errors were split into 4 subsets: 1.) lead turn in the wrong direction, 2.) lead turn at an inefficient rate, 3.) lead turn too early or too late, and 4.) the lead turn was not present. Raters assigned a value to an error that corresponded with the definitions above.

Forced Reliance. Forced reliance was used to capture when the pilot could not tell what the autonomy is doing or where it was positioned in relation to the bandit based on information from the HUD or the TSD and did not paddle. A pilot may not take over controls because they have low situational awareness and although they are not comfortable with the situation or fully trusting the autonomy, they feel like they have to rely on it. Forced reliance is a categorical variable where 0 indicates no forced reliance and 1 indicates forced reliance. This category was created from SME conversations and task parameters.

Look at autonomy for reason other than lack of trust. A pilot may look outside in the HUD or at the TSD and still have full trust in the autonomy. The goal of tagging this event is to disentangle the signal of looking outside or at the TSD for lack of trust versus looking outside or at the TSD for another reason such as calling the WEZ. Look at autonomy is a categorical variable.

Meet Needs of Mission. This category was extracted from the HRI Trust Perception Scale (Schaefer, 2016) to assess the pilot perception of whether the autonomy was meeting the needs of the mission by successfully performing its assigned tasks. The pilots were told that the mission objective was to "not die" and only paddle off if they thought the autonomy was putting their life at risk. Raters assigned a value between 0 and 1, where 0 was indicative of autonomy not meeting the needs of a mission and 1 completely meeting the needs of a mission. Meets the needs of mission is a quantitative discrete variable.

Paddle. Paddle refers to a deliberate disengagement of the autonomy and to take over the dogfighting task. It is highly indicative of low trust or distrust. Raters assigned a value of 1 only when the subject explicitly called out paddle. Paddle is a categorical variable developed from the task parameters. This category was developed from SME conversations and task parameters.

Predictable. This category was extracted from the HRI Trust Perception Scale (Schaefer, 2016) to assess the pilot's perception of whether autonomy was behaving in a predictable manner. Raters assigned a value between 0 and 1, where 0 was indicative of the autonomy being completely unpredictable and 1 meaning the autonomy was completely predictable.

Situational Risk Perception. Situational risk perception category was assessed for its contribution to trust (Hoff & Bashir, 2015). The risk of the situation is dependent upon the distance of the bandit to the own ship and how/when the autonomy maneuvers.

Situational risk perception is a quantitative discrete variable from 0 to 1. Where 0 is no risk in the situation and 1 is high risk.

Surprise. Surprise category was used to indicate when the pilot is surprised by something in the experiment. We suspected that the pilot may be surprised by the battle management task, displays of the tasks, or the general cadence of the task at some point. Surprise is a categorical variable where 0 indicates no surprise and 1 indicates surprise. This category was developed from SME conversations.

Thinking about Paddling. The act of 'paddling' the autonomy involves the pilot pressing a control switch to de-activate the autonomy pilot. Pilots 'paddle' the autonomy when they no longer believe the autonomy is capable of performing as desired and they wish to take over the task manually. The exact time when the autonomy is deactivated is precisely recorded by the computer. However, knowing when a pilot is beginning to think about deactivating the autonomy is very helpful in understanding when trust begins to degrade. Raters assigned a value of 1 only when the subject explicitly stated thoughts of paddling. Thinking about paddling is a categorical variable.

Transparency. This category was an adaptation of the category "Provides Feedback" from the HRI Trust Perception Scale to assess the pilot's perception of mistakes of the autonomy. Raters assigned a value between 0 and 1, where 0 indicates the autonomy was not transparent, and 1 that the autonomy was completely transparent.

Trust Level. This category was extracted from the HRI Trust Perception Scale (Schaefer, 2016) to assess the pilot's perception of whether autonomy was trustworthy or not. Raters assigned a value between 0 and 1, where 0 was indicative of the autonomy being completely untrustworthy and 1 meaning the autonomy was completely trustworthy. Trust level is a quantitative discrete variable.

Uncomfortable/Feeling Sick. whenever the pilot was feeling uncomfortable or sick throughout the test cards. Uncomfortable/feeling sick is a categorical variable where 0 indicates no sickness and 1 indicates feeling of sickness. The pilots are sitting in the simulator cockpit with the headset this may cause discomfort at some point and wanted to capture that feeling as it could affect their mood. This way, if their mood is affected by their comfort level, it is not mistaken for lack of trust or frustration with the performance of the autonomy. This category was created based on SME conversations.

Determining Inter-Rater Reliability with Three Raters Method. The subjective codebook definitions were used to rate and quantify the subjective qualitative data collected through subject video replay debrief. The subjective qualitative data tags for one subject were then exported into a spreadsheet and shared with three raters along with a subjective behavioral codebook. To quantify the subjective qualitative data and assess the validity and reliability of the subjective behavioral codebook, the raters quantified each data tag in terms of key constructs defined in the codebook (e.g., "meet mission objectives", or "trust", or "predictable"). Their responses were then used to compute inter-rater reliability (IRR). IRR is a measure of agreement or consistency between the raters in terms of their interpretations or assessment decisions. The level of consistency is important to inferring results with high confidence and has direct implications to both construct validity and instrument reliability. The software application Statistical Package for Social Sciences (SPSS) was used to compute the IRR. SPSS offers several methods to compute IRR, including Cohen's kappa (k), weighted kappa (k_w), and reliability analysis $ICC_{(IRR)}$, where each may be appropriate depending on the number of raters compared and type of data used. In this study, the subjective qualitative data was quantified using the subjective behavioral codebook definitions, and subsequently transformed into quantitative discrete data. IRR was calculated using two methods. First, Cohen's kappa was calculated using the weighted kappa method, which is appropriate when more than two raters are involved, and the variables can be considered quantitative ordered categories. Then, the IRR result was cross validated using the reliability analysis method by computing interclass correlation coefficient.

Results

Inter-Rater Reliability with Weighted Kappa. Cohen's kappa (k) accounts for the disagreement between the two raters but does not measure the degree of disagreement. Weighted kappa (k_w) is a variant of Cohen's kappa (k) that measures the degree of disagreement between the two raters by using the weighting schemes to take into the account the closeness of agreement between different categories. Cohen's kappa (k) is appropriate to use when comparing the responses between two raters, and when variables are nominal. This study uses multiple raters and quantitative variables that are ordered and discrete, thus it leverages the weighted kappa measure (k_w). Depending on a level of agreement between multiple raters, IRR may be interpreted as no agreement ($k_w < 0$), slight ($0 < k_w < 0.2$), fair ($0.2 < k_w < 0.4$), moderate ($0.4 < k_w < 0.6$), substantial ($0.6 < k_w < 0.8$), and perfect ($0.8 < k_w < 1$).

Weighted kappa was computed to determine the degree of agreement between three raters in interpretation of subjective qualitative data tags using the subjective behavioral codebook definitions. Results met "perfect" weighted kappa criteria ($0.80 < k_w < 1$, $p < 0.05$) between the three raters (Table 1).

Table 1. ***Cohen's Weighted Kappa Values.***

	Weighted	Asymptotic			95% Asymptotic Confidence Interval	
	Kappa[a]	Std. Error[b]	z[c]	Sig.	Lower Bound	Upper Bound
Rater_1 – Rater_2	.87	.03	13.66	0.00	.82	.92
Rater_1 – Rater_3	.85	.02	13.44	0.00	.81	.90
Rater_2 – Rater_3	.83	.02	12.84	0.00	.79	.88

a. The estimation of the weighted kappa uses linear weights.
b. Value does not depend on either null or alternate hypotheses.
c. Estimates the asymptotic standard error assuming the null hypothesis that weighted kappa is zero.

Inter-Rater Reliability through Interclass Correlation Coefficient ($ICC_{(IRR)}$)

Interclass correlation coefficient ($ICC_{(IRR)}$) is often used to assess the reliability index in IRR analysis. $ICC_{(IRR)}$ reflects the *degree* of rater agreement. It is calculated by mean squares and determined through analysis of variance in agreement between raters. This

method is appropriate to use when comparing more than two raters and when variables are quantitative and discrete. Depending on $ICC_{(IRR)}$ values, IRR may be interpreted as poor ($ICC_{(IRR)} < 0.50$), moderate ($0.50 < ICC_{(IRR)} < 0.75$), good ($0.75 < ICC_{(IRR)} < 0.90$), and excellent ($ICC_{(IRR)} > 0.90$, $p < 0.05$).

To assess IRR using reliability analysis method, an absolute agreement Interclass ($ICC_{(IRR)}$) was computed using a two-way mixed effects model. Results suggest that there was an excellent agreement between the three raters ($ICC_{(IRR)} = 0.96$, $p = 0.00$) (Table 2), validating the results obtained using the weighted kappa (k_w) method.

Table 2. ***Intraclass Correlation Coefficient.***

	Intraclass Correlation [b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.88[a]	.84	.92	23.90	96	192	.00
Average Measures	.96[c]	.94	.97	23.90	96	192	.00

Two-way mixed effects model where people effects are random and measures effects are fixed.
a. The estimator is the same, whether the interaction effect is present or not.
b. Type A intraclass correlation coefficients using an absolute agreement definition.
c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Discussion and Future Work.

We successfully formulated a codebook operationalized to the WVR ACM scenario for coding trust-related behaviors that would otherwise be captured using self-report surveys. The codebook will also be used to tag outliers and help explain any anomalies that coincide with the objective data. The timestamped tags will be used in combination with physiological and behavioral data to derive objective methods to recognize trust behaviors in real time. This method provides greater temporal granularity without interrupting pilots during a flight, albeit at the cost of additional debrief time. We believe the same method can be adapted for other high-risk high tempo scenarios where ESM may be impossible (e.g. in flight) or impractical.

Following the collection of experimental data, we will code debrief data and use it in regression modeling for validation of a Structural Equation Model (SEM) (Israelsen et al., 2021).

Acknowledgement

The work was supported by the Defense Advanced Research Projects Agency (DARPA) Air Combat Evolution program, Technical Area 2 “build and calibrate trust in air combat local behaviors”, Soar Technology, Inc contract #FA8650-20-C-7044, awarded by the Air Force Research Laboratory (AFRL). The authors would like to thank Dr. Anna Skinner (DARPA) and Dr. Joseph Lyons (AFRL) for their advice and support. The views expressed herein are those of the authors and do not represent the official policy or position of the US Government, Department of Defense, the US Air Force, of DARPA.

References

Brown, D. L., Bautsch, H. S., Wetzel, P. A., & Anderson, G. M. (2002). *Instrument scan strategies of F-117A pilots*. LOGICON TECHNICAL SERVICES INC DAYTON OH.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human factors*, *37*(1), 32-64.

Endsley, M.R., & Garland, D.J. (2000). Direct measurement of situational awareness: Validity and use of SAGAT. *Situation Awareness Analysis and Measurement*, 147-174.

Garbarino, M., Lai, M., Bender, D., Picard, R. W., & Tognetti, S. (2014, November). Empatica E3—A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 4th International Conference on WirelessMobile Communication and Healthcare TransformingHealthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)* (pp. 39-42). IEEE.

Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cerebral cortex (New York, NY: 1991)*, *7*(4), 374-385.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, *7*(1), 7-28.

Highland, P., Williams, J., Yazvec, M., Dideriksen, A., Corcoran, N., Woodruff, K., ... & Schnell, T. (2020). Modelling of unmanned aircraft visibility for see-and-avoid operations. *Journal of Unmanned Vehicle Systems*, *8*(4), 265-284.

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

Israelsen, B., Wu, P., Woodruff, K., Avdic-McIntire, G., Radlbeck, A., McLean, A., ... & Javorsek, D. A. (2021, March). Introducing SMRTT: A Structural Equation Model of Multimodal Real Time Trust. In Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (pp. 126-130).

Klein, G. A. (2017). *Sources of power: How people make decisions*. MIT press.

Larson, R., & Csikszentmihalyi, M. (2014). The experience sampling method. In *Flow and the foundations of positive psychology* (pp. 21-34). Springer, Dordrecht.

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

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277-301.

Martin, P., Calhoun, P., Schnell, T., & Thompson, C. (2019, June). Objective measures of pilot workload. In *63RD Setp Symposium Proceedings*.

Ogdon, D. C. (2019). HoloLens and VIVE pro: virtual reality headsets. *Journal of the Medical Library Association: JMLA*, *107*(1), 118.

Pashler, H. E. (1999). *The psychology of attention*. MIT press.

Roscoe, A. H., & Ellis, G. A. (1990). *A subjective rating scale for assessing pilot workload in flight: A decade of practical use*. ROYAL AEROSPACE ESTABLISHMENT FARNBOROUGH (UNITED KINGDOM).

Schaefer, K. E. (2016). Measuring trust in human robot interactions: Development of the “trust perception scale-HRI”. In *Robust Intelligence and Trust in Autonomous Systems* (pp. 191-218). Springer, Boston, MA.

Taylor, R. M. (1990). Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design. Situational Awareness in Aerospace Operations (AGARD-CP -478). *Neuilly Sur Seine, France: NATO-AGARD*.