INSTITUTE FOR DEFENSE ANALYSES

Initial Validation of the Trust of Automated Systems Test (TOAST)

Heather Wojton, Project Leader

Daniel Porter
Stephanie Lane
Chad Bieber
Poornima Madhavan

January 2019

Approved for public release.

IDA Non-Standard Document
NS D-10416

Log: H 2018-000555

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under Central Research Project C9089, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Mr. Robert R. Soule and consisted of John Haman from the Operational Evaluation Division, and Mike Gilmore from the System Evaluation Division.

For more information:
Heather Wojton, Project Leader
hwojton@ida.org • (703) 845-6811

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Non-Standard Document NS D-10416

Initial Validation of the Trust of Automated Systems Test (TOAST)

Heather Wojton, Project Leader

Daniel Porter
Stephanie Lane
Chad Bieber
Poornima Madhavan

Initial Validation of the Trust of Automated Systems Test (TOAST)

Heather Wojton[1]

Daniel Porter[1]

Stephanie Lane[1]

Chad Bieber[1]

Poornima Madhavan[1]

[1] Institute for Defense Analyses (IDA)

Initial Validation of the Trust of Automated Systems Test (TOAST)

Trust is a key determinant of whether people rely on automated systems in the military and the public. However, there is currently no standard for measuring trust in automated systems. In the present studies we propose a scale to measure trust in automated systems that is grounded in current research and theory on trust formation, which we refer to as the Trust in Automated Systems Test (TOAST). We evaluated both the reliability of the scale structure and criterion validity using independent, military-affiliated and civilian samples. In both studies we found that the TOAST exhibited a two-factor structure, measuring system understanding and performance (respectively), and that factor scores significantly predicted scores on theoretically related constructs demonstrating clear criterion validity. We discuss the implications of our findings for advancing the empirical literature and in improving interface design.

Initial Validation of the Trust of Automated Systems Test (TOAST)

Systems that act with little to no input from humans are increasingly common. The U.S. military, for example, uses automated systems to perform search and rescue missions and to assume control of aircraft to avoid ground collision. The public uses these systems widely to perform tasks as trivial as vacuuming to those as critical as regulating nuclear power plants. Automated systems use sensors and computer programming to perform tasks with little or no human intervention (Boulanin & Verbruggen, 2017). Ideally, these systems should improve people's safety and performance by completing complex, repetitive tasks efficiently and aiding or replacing humans in hazardous environments. In reality, however, automated systems are imperfect and can behave in ways that produce unintended consequences. This imperfection can result in misuse or disuse of the systems under certain conditions.

Trust is a key determinant of whether people will rely on automated systems (Hoff & Bashir, 2015; Lyons, et al., 2016). Research demonstrates that people rely on automation when they believe that it will help them achieve their goals in situations characterized by uncertainty and vulnerability (Lee & See, 2004). Accidents occur, however, when people trust automated systems inappropriately. The Costa Concordia cruise ship sunk in 2012, for example, because the captain placed too little trust in the ship's automated navigation system, manually diverging from the route it selected, crashing into a shallow reef and killing 32 passengers (Levs, 2012). It is critical, therefore, that automated systems are designed to facilitate appropriate levels of trust.

Currently, there is no standard for measuring trust in automation. In general, researchers measure it using custom scales or validated measures of theoretically related constructs, making it difficult to draw comparisons across studies or craft criteria to determine whether automation is facilitating appropriate levels of trust (Merritt, 2011). Furthermore, the published scales that do exist perform inconsistently, exhibiting unreliable scale structure (Jian, Bisantz, Drury, & Llinas, 2009). In the present studies, we propose a scale, grounded in existing research and theory on the factors that drive trust formation, and evaluate its structure and validity in independent samples.

Trust in Automation Framework

Research suggests that three factors form the foundation of trust in automated systems: knowledge of the system's purpose, performance, and underlying processes (Lee & See, 2004; Lee & Moray, 1994). Purpose refers to a person's knowledge of why the system was built and how the designer intended for users to employ it. Evidence suggests that purpose is a stronger determinant of trust with users who have little to no experience employing the system (Hoff & Bashir, 2015). Imagine that your new smart phone comes preloaded with a map application you have not seen before. You need to navigate to a meeting at a restaurant, so the question becomes: do you use it? Since you have never used this app before, your decision likely depends on what you know or assume about it. If you know nothing at all, research suggests that you are likely to assume the system works and will navigate you to your desired location. Further, multiple studies clearly demonstrate that in the absence of information about performance, people expect novel systems to perform perfectly (Beck, Dzindolet, & Pierce, 2007; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003), driving higher levels of trust.

Trust can rapidly deteriorate, however, if users' expectations for perfect performance do not match their experience (Madhavan & Weigmann, 2007). In the example above, you expect the app to take you to the restaurant to meet your friend. If instead, it tries to drive you into a river, you might think twice about relying on it in the future.

As users gain experience with automated systems, performance—the system's ability to execute specific tasks—replaces purpose as the primary driver of trust (Hoff & Bashir, 2015). Research demonstrates that trust in automation changes based upon training (Koustanaï, Cavallo, Delhomme, & Mas, 2012) and real-time performance. In particular, trust increases when the system performs consistently, without error, and in a way that is consistent with users' expectations (Biros, Daly, & Gunsch, 2004; Guznov, Nelson, Lyons, & Dycus, 2011; Merritt & Ilgen, 2008; de Visser & Parasuraman, 2011).

In some cases, users may come to understand the processes (algorithms and operations) that automated systems use to perform tasks through either training or direct experience. Few studies address the impact that knowledge of underlying processes plays in trust formation. In theory, however, such knowledge should shift users' focus away from a simple accounting of success and failure toward the qualities and characteristics of the automation (Lee & See, 2004). This shift encourages appropriate levels of trust by clarifying the conditions under which the automation is likely to perform as expected. If you knew that the map app was created by physically scanning real maps, and that it sometimes mistakenly read a river to be a road, you may be less likely to stop trusting the system altogether and instead take a more nuanced approach where you trust it under some conditions but not others.

Current Measurement Approaches

Current approaches to measuring trust in automated systems vary across researchers, and few (if any) capture the factors that researchers believe form the basis for trust judgments. The most common methods include measuring trust-related behaviors, such as system reliance or compliance (Hoff & Bashir, 2015), and quantitative self-reports of trust or theoretically related constructs (Merritt, 2011; Lerch, Prietula, & Kulik, 1997). Using trust-related behaviors as proxies for trust is problematic. Trust is an expectation and although it shapes reliance and compliance, it does not completely determine it (Lee & See, 2004; Muir, 1987). Biros and colleagues (2004) demonstrate, for example, that in high-workload situations, users rely on automated systems to keep up with task demands even after reporting that they distrust the system. Additionally, the use of trust-related behaviors as proxies of trust makes it difficult (if not impossible) to discern how knowledge of system purpose, performance, and underlying processes affect trust judgments.

Quantitative self-report measures trust in a more direct way and, as a consequence, is not subject to the same limitations as indirect measures like reliance and compliance. In fact, researchers use quantitative self-report to measure beliefs and expectations across a range of domains, including health and interpersonal and intergroup relations (among others). Generally, researchers prefer to use multi-item scales that research demonstrates are both reliable and valid in order to increase confidence in their findings, facilitate between-study comparisons and meta-analyses, and develop standards for what constitutes good performance. Critically, however, researchers interested in trust in automation largely design and implement custom scales to test their hypotheses, decreasing confidence in their findings and hindering direct comparisons across studies. In addition, these custom-made scales often fail to reflect the diversity of factors that form the theoretical basis for trust judgments – system purpose, performance, and underlying processes.

One notable exception is the Trust in Automation Scale (TAS) developed by Jian and colleagues (2000). The TAS is a multi-dimensional, 12-item scale that is designed to measure both distrust (items 1-5) and trust (items 6-12). The authors developed the TAS in a multi-phase study with college students in which they described trusting people, trusting machines, or trust in general (phase 1); rated the extent to which words derived from these descriptions reflect trust or distrust (phase 2); and compared the similarity of 30 words associated with trust and distrust with each of the other words in the subset (phase 3). Subsequently, the authors submitted these comparisons to factor analysis and cluster analysis.

The TAS is, to our knowledge, the most widely used measure of quantitative self-report among studies in which researchers chose not to create a custom scale. There are, however, reasons to question its validity. In phase 1, for instance, the authors did not filter out words describing trust in general or trust between humans, and retain only those describing trust in machines for use in later phases. It is possible, therefore, that the scale reflects trust in an agent – human, machine, or otherwise – rather than trust in automation. Additionally, evidence suggests that the scale structure is unreliable. In phase 3, Jian and colleagues (2000) report results from a factor analysis that are difficult to interpret because of cross-loading. More recently, Lyons and colleagues (2011) demonstrated a two-factor structure consistent with the trust and distrust subscales but, again, some items loaded onto both factors, failing to exhibit simple structure. Furthermore, the distrust subscale attributes human characteristics, such as being deceptive or having intent, to automated systems. It is unclear, therefore, whether scale scores are driven primarily by distrust or by a tendency to personify automation. Related to this point, our unpublished research on military personnel suggest that some populations find it aversive to personify machines, leading them to respond to scale items in unanticipated ways.

The present studies seek to provide a reliable and valid measure of trust in automation that is grounded in the trust formation framework discussed earlier. In particular, we propose a multi-item scale called the Trust of Automated Systems Test (TOAST), designed to measure three factors: knowledge of system purpose, performance, and underlying processes, and evaluate its performance in both military-affiliated (study 1) and civilian (study 2) samples. The proposed scale is presented in Table 1. In the section below, we briefly outline the methods used to evaluate scale performance, including both scale structure and validity.

Evaluating Scale Structure and Validity

Scale performance is determined by the extent to which the scale is valid – that is, the extent to which the scale reliably measures the construct that it is designed to measure. There are several methods for establishing validity. One common method, referred to as face validity, involves asking subject-matter experts or participants to rate how well a set of items appears to measure a specific construct, such as trust in automation, and retaining items with the highest scores. Face validity is, arguably, among the weakest forms of evidence that a scale is valid, measuring only what the scale *appears* to measure. More rigorous methods require that the scale demonstrate a reliable, theoretically predictable structure and that scale scores predict theoretically related outcomes.

Confirmatory factor analysis (CFA) is a statistical technique that researchers use to determine how well a hypothesized scale structure fits the data. In particular, researchers specify the number of factors (constructs) the scale should measure and which items were designed to measure each factor. These decisions typically rely on existing theory or findings from earlier exploratory research. For our scale, the literature suggests that three factors form the basis for trust in automation, namely system purpose, performance, and underlying processes. The TOAST, therefore, consists of three distinct subscales whose items were designed to measure each of these factors. The mapping of items to factors is presented in Table 1. The present studies use CFA to evaluate how well this three-factor structure fits the observed data and whether alternative factor structures are superior.

Table 1. Proposed multi-item scale and hypothesized factor structure.

System Purpose	System Performance	Underlying Processes
• I understand what the system should do. • I understand the limitations of the system. • I understand the capabilities of the system.	• The system helps me achieve my goals. • The system performs consistently. • The system performs as it should. • I feel comfortable relying on the information provided by the system. • I think I could do a better job than the system. • I am concerned the system is vulnerable to hacking.	• I understand how the system executes tasks. • I wish I had more control over how the system executes tasks. • I am rarely surprised by how the system behaves. • I wish the system gave me more information. • I know when I should trust the system.

A critical next step in scale development is demonstrating that scale scores predict theoretically related outcomes. This is referred to as concurrent validity and is strong evidence that a scale measures what it is designed to measure. If we used the TOAST to measure trust in the automated map app discussed above, for example, we would expect that TOAST scores would positively correlate with people's propensity to rely on the app and the likelihood that they would recommend the app to others. Establishing these relationships is critical for demonstrating the validity of our scale and is thus a key part of our evaluation.

Overview of Present Studies

The goal of the present studies is to develop a valid measure of trust in automated systems that is grounded in existing theory and research on trust formation. In Study 1, Reserve Officer Training Corps (ROTC) Cadets read about an autonomous search and rescue system that was either reliable or unreliable and reported the degree to which they trusted the system using the TOAST. In Study 2, civilians reported on their experience using digital assistants, such as Alexa or Siri, and completed the TOAST to indicate their level of trust in these systems. Participants in both studies also completed measures of theoretically related constructs, such as their intentions to rely on these systems in the future and a single-item measure of system trust. We used CFA to evaluate the scale structure of the TOAST and simple correlations to demonstrate concurrent validity for each of its subscales. We hypothesized that:

H_1: The TOAST would demonstrate a three-factor structure consistent with the trust formation framework, clearly representing knowledge of system purpose, performance, and underlying processes.

H_2: Each subscale would demonstrate a positive correlation with theoretically related constructs, including custom measures of system trust and reliance intentions.

H_3: The pattern of findings in Study 1 would replicate in Study 2 despite using a different system and population.

Study 1

Method

Participants

Participants were ROTC Cadets attending two public Southern universities with large ROTC programs. We contacted local IRBs and commanding officers to obtain permission to recruit Cadets through an e-mail to their ROTC listserv or an advertisement in their weekly newsletter. We took special care to ensure that Cadets understood that their participation was completely voluntary and that their commanding officers did not require them to participate. Participants received a $10 Amazon gift card in exchange for their participation.

Ultimately, we collected 350 responses, 19 of which showed signs of being duplicate submissions and were discarded, leaving a final sample of 331 (64 Female, 260 Male, 7 Missing). Service affiliations were primarily through the Air Force (215), followed by the Army (51), Navy (37), and Marines (5), but 22 participants reported no military affiliation.

Materials

We prepared the survey materials using Snap 11 Professional and presented them online via Snap WebHost services.

Vignettes. We crafted two vignettes to manipulate the perceived reliability of a hypothetical system. Both vignettes began:

> Imagine that you are conducting a search and rescue mission aided by an unmanned ground vehicle called UGV S350. The vehicle is designed to move independently through a conflict zone, scanning buildings to determine if victims are trapped inside. When a victim is detected, the vehicle alerts the rescue unit by sounding an alarm. During the mission, UGV S350

The high-reliability version of the vignette continued:

> successfully alerted the rescue unit to the presence of both civilian and military personnel that were trapped during the conflict and aided in their recovery. All rescued parties were transported to a local hospital for treatment.

Whereas the low-reliability version read:

> falsely alerted the rescue unit that a victim was trapped inside a building, and the rescue unit was unable to locate a victim after performing a comprehensive search. This extended the duration of the mission and placed the rescue unit at risk.

Manipulation Check. Participants completed a manipulation check where they judged the reliability of the system on a 7-point scale, ranging from 1 = *Very unreliable* to 7 = *Very reliable*. As expected, participants who read the high-reliability vignette rated UGV S350 as significantly more reliable than those who read the low reliability vignette, $t(329) = -9.55$, $p < .001$, 95% CI = [-1.42, -0.94].

Reliance Intentions. We measured participants' intentions to rely on the system using two items. The first item read, "I would recommend others who are placed in this situation use UGV S350 to locate victims during search and rescue missions." and the second, "If placed in a similar situation in real life, I would rely on the UGV S350 to locate victims during search and rescue missions." Participants indicated their agreement with these items using a 7-point scale, ranging from 1 = *Strongly Disagree* to 7 = *Strongly Agree*.

Trust. We measured system trust in two ways via a single item, "I trust the system.", and the TOAST (Table 1), a 14-item scale designed to measure trust of automated systems. Participants indicated their agreement with both the single item and the TOAST using a 7-point scale, ranging from 1 = *Strongly Disagree* to 7 = *Strongly Agree*.

Procedure

After providing electronic consent and filling out basic demographic information, participants were randomly presented with either the high- or low-reliability vignette. Following the vignette, they completed the manipulation check and reported their reliance intentions. Next, participants saw either the single system trust item or the TOAST. The measure they saw first was randomly determined, and the order of items within the TOAST was randomized. Finally, participants were debriefed and asked to enter a six-digit PIN. Afterward, they were automatically redirected to another website to enter their e-mail address and the PIN they had created. The PIN was used to verify that they had taken the survey. Their e-mail address was not connected to their responses at any point, and the PIN was scrubbed from the dataset prior to analysis.

Results

Estimation Details

All confirmatory factor models were estimated in the freely available open-source R package *lavaan*, version 0.5-23. Full information maximum likelihood (FIML) was used to estimate the confirmatory factor models, which uses all available case information to inform the likelihood (Enders & Bandalos, 2001). Therefore, under the assumption that the data are missing at random (MAR), the missing data present in Study 1 and Study 2 do not pose a problem. Finally, we use robust maximum likelihood (MLR) to appropriately scale the test statistics (e.g., the chi square statistic used in fit indices) and adjust the standard errors for the slight nonnormality present in the data.

In order to evaluate the goodness of fit of each model, we evaluate the χ^2 test statistic and its associated *p* value, the comparative fit index (CFI), and the root mean squared error of approximation (RMSEA). The χ^2 test statistic is a global measure of model fit, where a nonsignificant *p* value indicates good model fit. However, because of the known limitations of the chi square test statistic as a measure of model fit (Schermelleh-Engel, Moosbrugger & Müller, 2003), we present additional indices of fit. The CFI and RMSEA are two popular measures of model fit, where a CFI > .95 (Hu & Bentler, 1995) and an RMSEA < .05 (MacCallum, Browne, & Sugawara, 1996) indicate good model fit.

Confirmatory Factor Models

In Study 1, we fit a series of confirmatory factor models and evaluate their fit to the data: a three-factor model representing system purpose, performance, and process; a two-factor model representing system understanding (purpose and process combined) and system performance; and a one-factor model representing trust in automated systems. From these models, we evaluate measures of fit and model modification indices in order to select the model that best characterizes the data.

We begin by fitting a confirmatory three-factor model (purpose, performance, and process) to Study 1 data. This model did not fit the data well, $\chi^2(74) = 387.53, p < .01, CFI = .79, RMSEA = .11$. These indices of fit reflect poor, unacceptable model fit. Further, the reverse-scored items demonstrate negative and/or nonsignificant loadings on our first-order factors, indicating that these items are not behaving as intended. The poor performance of these reverse-scored items is further evinced by the low R-square values (< .05) for these items, indicating that less than 5% of the variance in these reverse-scored items is explained by the underlying latent variable. Therefore, all subsequent analyses exclude the reverse-scored items from the scale.

The removal of these items has implications for the factor structure of our scale. In our two-factor model, we remove the reverse-scored items from the set of possible items, we retain our system performance latent variable, and we merge the system purpose and process items into one first-order factor representing system understanding. The remaining items are presented in Table 2. The two-factor model represents a substantial improvement over the three-factor model and is characterized by an adequate fit to the data, $\chi^2(26) = 86.44, p < .01, CFI = .95, RMSEA = .08$. All factor loadings are significant, large, and positive. Further, all item R-squares are above .30. As expected, there is a significant, positive correlation between system understanding and system performance. The factor structure and associated loadings are presented in Figure 1.

Table 2. Revised multi-item scale.

System Understanding	System Performance
• I understand what the system should do.	• The system helps me achieve my goals.
• I understand the limitations of the system.	• The system performs consistently.
• I understand the capabilities of the system.	• The system performs the way it should.
• I understand how the system executes tasks.	• I am rarely surprised by how the system responds.
	• I feel comfortable relying on the information provided by the system.

Finally, we fit a one-factor model to the data in order to demonstrate that these items should not be considered to be one dimension of trust. A one-factor CFA fit to the final nine items represents a poor fit to the data, $\chi^2(27) = 254.79, p < .01, CFI = .79, RMSEA = .16$. The likelihood ratio test comparing the one-factor and two-factor models is significant, $\chi^2(1) = 168.35, p < .0001$, indicating a significant decrement in fit when moving to the more parsimonious one-factor model. Therefore, we retain the two-factor model moving forward.

Criterion Validity

In order to establish the criterion validity of our scale, we evaluated the relationship between the latent variables in our two-factor model with other theoretically related constructs – reliance intentions, trust, and reliability. In Study 1, we correlate our measures of system understanding and system performance with participants' rating of the extent to which they would 1) trust the system, 2) rely on the system themselves, and 3) tell others to rely on the system (see Table 3). We find that the subscales correlate significantly and positively with each of these items, suggesting that each subscale measures a unique aspect of trust in automation.

Table 3. Concurrent validity for Study 1.

	Understanding	Performance	Trust	Reliance – Self	Reliance – Other	Reliability
Understanding	1					
Performance	0.465	1				
Trust	0.263	0.695	1			
Reliance – Self	0.192	0.519	0.568	1		
Reliance – Other	0.204	0.622	0.587	0.752	1	
Reliability	0.103	0.585	0.620	0.634	0.670	1

Measurement Invariance

Finally, we aim to demonstrate the invariance of the scale's measurement properties as it pertains to reliability of the system. Briefly, establishing measurement invariance allows researchers to evaluate a scale's invariance across multiple groups (e.g., age, gender, occupation). Several levels of invariance exist, which progress from the least restrictive to the most restrictive. Invariance is tested in a sequence of steps, where researchers begin with a model where all parameters are fixed to equality across groups. These constraints may then be freed in a series of steps, where factor loadings are freed first across groups (weak invariance), then item intercepts (strong invariance), then factor means (strict invariance). Here, we aim only to establish weak invariance, which is equality in the factor loadings across groups.

In Study 1, the system was entirely hypothetical, and individuals were randomly assigned to a low or high reliability scenario. Therefore, we aim to examine the invariance of the scale across these low and high reliability groups. Study 1 demonstrated partial measurement invariance. Specifically, all individual items could be fixed to equality without a significant decrement in model fit, with the exception of item 7. This item related to the underlying latent variable of performance more strongly for the low reliability group than the high reliability group. All other items could be held constant across the low and high reliability groups; therefore, with one exception, the items are measuring trust in the same way for both groups.

Study 2

Study 1 successfully demonstrated the validity of the TOAST in a military-affiliated population. However, the emergent two-factor structure did not match perfectly with the three-factor solution predicted by the trust formation framework discussed in the introduction. Our scale showed evidence of purpose and performance factors, but the items written to capture process did not break out as an independent dimension. Process is advanced knowledge of how a system executes tasks or makes decisions. Given that the Cadets had no real experience with this hypothetical system, were given only a single, brief case description, and received no information about its process, it is possible that process is real but indistinguishable from purpose in low-knowledge populations. However, while process is theorized to be a dimension of trust formation, it has received little direct empirical examination.

Study 2 was developed in part to extend the results of Study 1's novel, hypothetical system for military users to a real system with which the general population had varying levels of experience. We chose to investigate user experiences and trust of digital assistants such as Apple's Siri or Amazon's Alexa to accomplish these goals. Study 2 was also designed to address hypotheses about experience with automation unrelated to trust, and so while these are described in the method, they are not addressed in the analysis of Study 2.

Method

Participants

Participants were recruited via Amazon's Mechanical Turk (MTurk) and were paid $0.50 for their participation. 501 individuals submitted a code for payment on MTurk, though only 488 responses were submitted to Snap WebHost. Of the remaining 488 participants, 163 responded incorrectly to a multiple-choice item designed to check their attention. Specifically, the item read, "From the following list of countries, select England." Those who failed to select England were excluded. Additionally, 27 participants who passed the attention check completed the survey in under three minutes (2.5 seconds per item to read instructions, load pages, read each item, consider and then select a response), and 8 who passed both of these criteria gave absurd responses (the consequence in most cases of incorrectly playing music are "life-threatening"). This left us with a final sample of 288 (109 [38%] Female, 172 [60%] Male, 2 Other, and 5 No Response).

Materials

We prepared survey materials using Snap Professional 11 and presented them online using Snap WebHost.

Digital Assistant Exposure. We measured participants' exposure to digital assistants through two items. Specifically, we asked whether they "have either a smart phone or smart home (e.g., Amazon Echo or Google Home)" and whether they had used "Siri (Apple/iPhone)," "Cortana (Microsoft)," "Alexa (Amazon)," or "Google Assistant (Android/Google)." If they had used any, they were asked to indicate which they had used the most and, if not, about which one they knew the most. Their answer to this question was piped as text into later items, indicated below as {Piped DA name}.

Experience with Digital Assistants. We measured participants' subjective evaluation of their experience using digital assistants, their use habits, and their general technological intelligence using three items. Namely, "When it comes to digital assistants, how experienced would you say you are?" (1 – *Not at all*, 4 – *Somewhat*, 7 – *Extremely*), "I use digital assistants all the time" (1 – *Strongly Disagree*, 7 – *Strongly Agree*), and "My friends would say I'm pretty tech savvy," (1 – *Strongly Disagree*, 7 – *Strongly Agree*).

Frequency of Digital Assistant Use. We measured experience as frequency of digital assistant use in six categories: 1) Communication (Texting, Messaging, Calling), 2) Searching for information (e.g., Nearby locations, Facts), 3) Navigating, 4) Time management (e.g., Calendars, Alarms), 5) Playing music, and 6) Shopping. Participants estimated how frequently they engage in these tasks using a digital assistant on a 6-point scale: 0 (*Never*), 1 (*Less than once a month*), 2 (*Monthly*), 3 (*Weekly*), 4 (*Daily*), 5 (*Many times a day*). They also reported how often "you yourself do the following tasks in situations where you realistically could have used the digital assistant {Piped DA name} instead," and "you use the digital assistant {Piped DA name} to do the following tasks." As a related question, they rated proportionally "how often you use the digital assistant {Piped DA name} for these tasks instead of doing it yourself," on a 7-point scale, 0 (*Never*), 3 (*About half the time*), and 6 (*All the time*).

Consequences of Digital Assistant Use. Participants rated the consequence of incorrectly completing each of the six tasks mentioned above on an 8-point scale: 0 (*Nothing at all*), 1 (*Trivial*), 2 (*Unimportant*), 3 (*Somewhat Unimportant*), 4 (*Somewhat Important*), 5 (*Important*), 6 (*Critically Important*), and 7 (*Life Threatening*). In particular, we asked them, "In most cases, how important would the consequences be if the following tasks were done incorrectly?" and "The most I would trust the digital assistant {Piped DA name} with is a task where the consequences are: _____".

Trust. As in Study 1, we measured system trust in two ways via a single item, "I trust the digital assistant {Piped DA Name}," and the TOAST. Participants indicated their agreement with both the single item and the TOAST using a 7-point scale, ranging from 1 = *Strongly Disagree* to 7 = *Strongly Agree*.

Reliance Intentions. Participants reported their reliance intentions using a single item, "I rely on the digital assistant {Piped DA name}," on a 7-point scale, ranging from 1 = *Strongly Disagree* to 7 = *Strongly Agree*.

Procedure

The study was advertised on MTurk as “A Survey on Digital Assistant Use.” After electronically consenting, participants read a brief definition of what we meant by digital assistants. Participants were told that experience using digital assistants was not required to participate so they would not feel that they had to lie about their experience to be eligible for the study, thereby decreasing data quality.

> We would like to know about your experience with voice activated digital assistants such as iPhone's Siri, Microsoft's Cortana, Amazon's Alexa, or Google Assistant ("OK Google"). These systems respond to voice commands and execute tasks on your phone or smart home (e.g., Amazon Echo or Google Home). When we talk about digital assistants, we are talking about any of these systems or ones like them.
>
> You do not need experience with these systems to participate today.

In the first block of questions, participants reported whether they owned a smart phone or home, and answered the subjective digital assistant experience items. On the next page, they indicated their familiarity with various digital assistants and which ones they had used. If participants indicated that they had never used a digital assistant, we asked them to report which digital assistant they knew the most about. Their answers to these items determined which system we asked them to consider for the rest of the survey.

Next, in four blocks presented in random order, participants completed the TOAST and the single-item measures of trust, reliance intentions, and the highest-consequence task they would use their digital assistant to complete. The TOAST items were presented on the same page in a random order. The other three items appeared on separate pages one at a time.

After this randomized block set, participants reported how frequently they used their digital assistant to complete six tasks. The order was set to the manual use then digital assistant use for the tasks on one page, followed by the relative digital assistant to manual use estimate on a separate page. In the final block, participants rated these tasks on how important the consequences are if the system completes the tasks incorrectly.

Afterward, participants filled out basic demographics questions. We embedded an attentional check item within the demographics form asking “From the following list of countries, select England.” After demographics were collected, we debriefed participants and sent them a code to submit to MTurk for their compensation.

Results

Confirmatory factor models

In Study 2, we aim to demonstrate the replication of our final model in an entirely independent sample. We progress through the analyses in the same order as Study 1, where we begin with a three-factor model, transition to a two-factor model, and finally test a one-factor model.

We begin by fitting a three-factor model to the data, which represents a poor fit to the data, $\chi^2(74) = 581.37, p < .01, CFI = .84, RMSEA = .12$. We see the same results as in Study 1 – namely, the model does not fit the data well, and a very small amount of variance in the reverse-scored items can be explained by the underlying latent variables (all R-square values < .06).

Next, we proceed to fitting our two-factor model representing system understanding and performance. We successfully replicate our two-factor model from Study 1 in Study 2. The model exhibits good fit to the data, $\chi^2(26) = 70.90, p < .01, CFI = .98, RMSEA = .05$. All factor loadings are significant, large, and positive. Further, all item R-square values are sizable and positive, indicating that the underlying latent variables are explaining a moderate to large portion of the variance in the items. These findings are presented in Figure 2.

Finally, we fit a one-factor model to Study 2 data, which exhibited a poor fit, $\chi^2(27) = 254.79, p < .01, CFI = .79, RMSEA = .16$. A likelihood ratio test comparing the one-factor and two-factor models is significant, $\chi^2(1) = 183.89, p < .0001$, indicating that the one-factor model is significantly worse than the one-factor model, and should therefore be rejected.

Criterion Validity

In Study 2, we correlate our measures of system understanding and system performance with trust and reliance intentions (see Table 4). For Study 1 and Study 2, we see significant, positive correlations between our latent variables and our measures of criterion validity. In both studies, we see that the system performance latent variable correlates more strongly with our criteria, indicating that positive perceptions of system performance may be more important for trust and reliance than understanding the system.

Table 4. Concurrent validity for Study 2.

	Understanding	**Performance**	**Trust**	**Reliance**
Understanding	1			
Performance	0.725	1		
Trust	0.556	0.718	1	
Reliance	0.431	0.598	0.673	1

Measurement invariance

In Study 2, the system is a common digital assistant, such as Siri, Cortana, or Alexa. Individuals who elect to complete Study 2 on MTurk therefore enter the study with pre-existing variability in their level of experience with these automated systems in everyday life. In order to test weak measurement invariance, we performed a median split on self-reported experience, which separated individuals into low and high self-reported experience. Although there are shortcomings associated with dichotomizing continuous variables (see, e.g., MacCallum, Zhang, Preacher, & Rucker 2002), we do so here only as a convenience for testing measurement invariance.

Study 2 demonstrated measurement invariance without qualification. That is, the factor loadings for all items could be constrained to equality across low and high levels of experience without a significant decrement in model fit. This test reveals that the items are "tapping into" the underlying latent variable in the same ways for members of both groups.

Discussion

As reliance on automated systems increases among the military and the public, the need for validated instruments to measure their trust of these systems will only continue to grow. The goal of the present studies was to develop such an instrument. In particular, we designed a multi-dimensional scale, the TOAST, to measure trust in automated systems and tested its validity across two studies. Overall, the TOAST (presented in Table 2) demonstrated a reliable, two-factor structure, and scores obtained on these factors demonstrate strong positive correlations with theoretically related constructs – clear evidence of criterion validity. We discuss these results in greater detail below.

Overall, the TOAST consistently measures two dimensions of trust: system performance and understanding. These factors emerged in two separate studies employing different methods and populations. Study 1 demonstrated that the TOAST was sensitive enough to detect differences in trust among military-affiliated operators for notional systems of varying reliability. Study 2 replicated the factor structure obtained in Study 1 in a civilian sample with real systems. The study further extended the TOAST to differences in experience that were impossible to measure in Study 1, finding that regardless of experience, the factor structure remained invariant. In neither study did we find support for a third dimension of trust; participants did not differentiate between knowledge of a system's purpose and understanding of its underlying processes, even when they had significant experience with the system. However, the robustness of the TOAST across systems and user bases provides strong evidence of its value as a tool for measuring trust of a variety of different systems.

While the TOAST appears to measure trust in a consistent way and be sensitive to important variations in systems, a key limitation of this set of studies was that we were only able to measure reliance behavior through intentions or self-report. In future studies, we will directly evaluate whether the TOAST predicts real reliance behaviors during live military exercises. Additionally, our measure of experience was self-reported, and we were unable to disentangle the causal pathway of whether trust increased with use or if it led to self-selection of users.

This initial validation of the TOAST provides academic, industry, and government researchers and designers with a tool to measure operators' trust of their system and determine whether the interface will likely facilitate appropriate reliance behavior. While other trust scales exist, they have not undergone rigorous reliability and validity testing, and [operational] use of these scales suggests they may be aversive or inappropriate for some populations (e.g., military personnel). Our studies provide evidence that the TOAST overcomes these shortfalls while providing a more nuanced analysis of operator trust.

Figure 1. Final two-factor solution from Study 1 (unstandardized estimates shown).

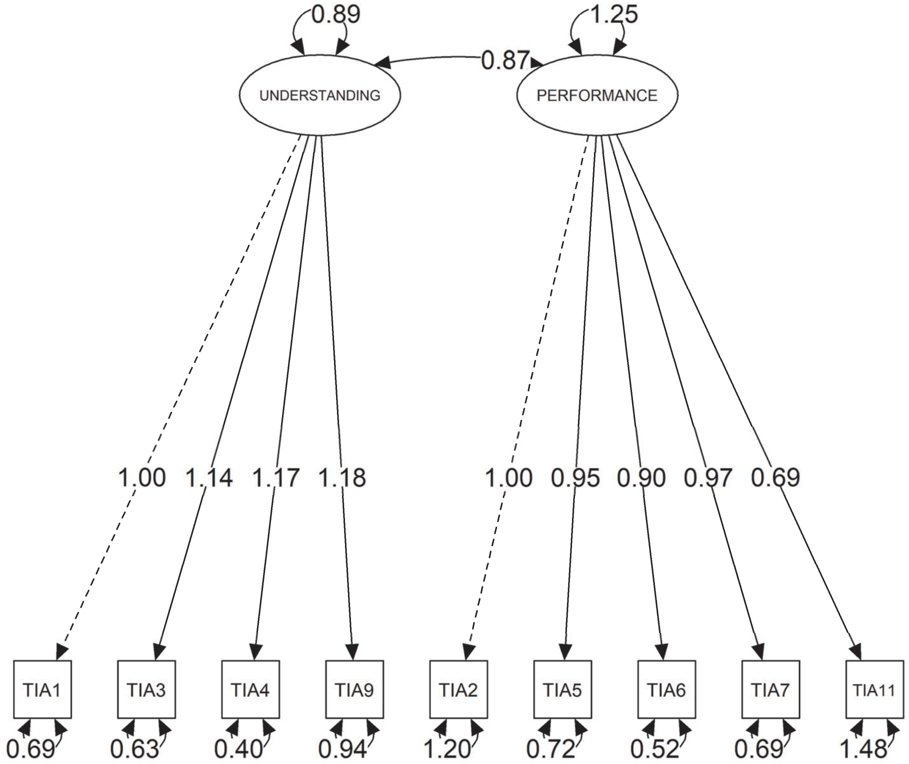
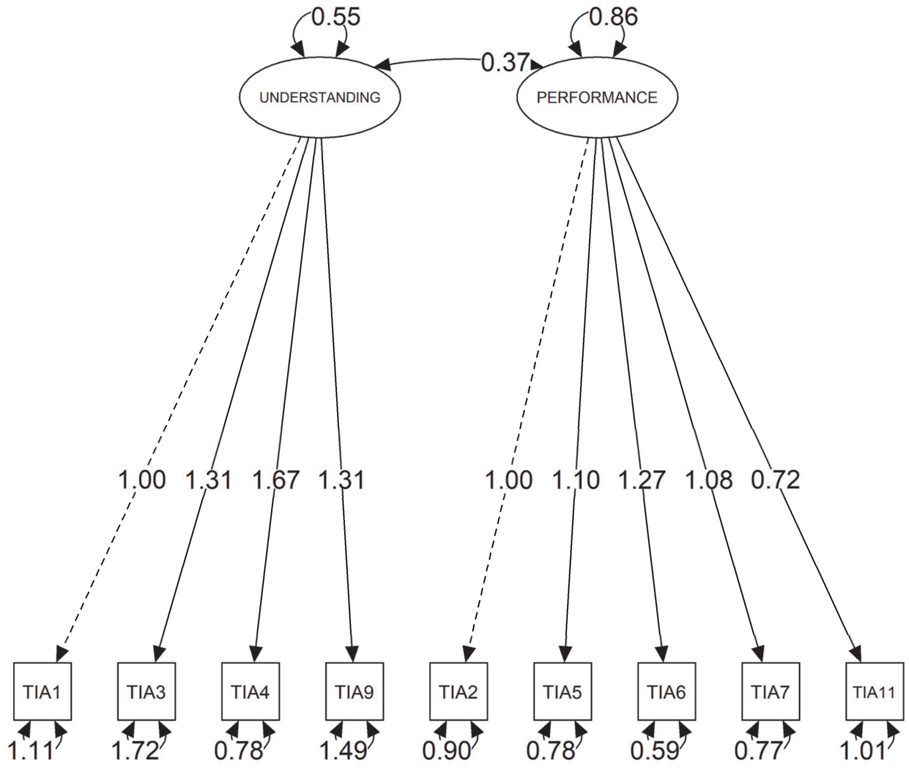


Figure 2. Final two-factor solution from Study 2 (unstandardized estimates shown).



References

Beck, H., Dzindolet, M., & Pierce, L. (2007). Automation usage decisions: Controlling intent and appraisal errors in a target detection task. *Human Factors*, 429-437.

Biros, D., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, 173-189.

Boulanin, V., & Verbruggen, M. (2017). *Mapping the development of autonomy in weapons systems.* Solna, Sweden: Stockholm International Peace Research Institute.

de Visser, E., & Parasuraman, R. (2011). Adaptive aiding of human-robot teaming effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making*, 209-231.

Dzindolet, M., Peterson, S., Pomranky, R., Pierce, L., & Beck, H. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 697-718.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*(3), 430-457.

Guznov, S., Nelson, A., Lyons, J., & Dycus, D. (2011). The effects of automation reliability and multi-tasking on trust and reliance in a simulated unmanned control task. *ADFA*, 1-6.

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

Hu, L. T., Bentler, P. M., & Hoyle, R. H. (1995). Structural equation modeling: Concepts, issues, and applications. *Evaluating Model Fit*, 76-99.

Jian, J., Bisantz, A., Drury, C., & Llinas, J. (2009). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 53-71.

Koustanaï, A., Cavallo, V., Delhomme, P., & Mas, A. (2012). Simulator training with a forward collision warning system effects on driver-system interactions and driver trust. *Human Factors*, 709-721.

Lee, J., & Moray, N. (1994). Trust, Self-Confidence, and Operators' Adaptation to Automation. *International Journal of Human-Computer Studies*, 153-184.

Lee, J., & See, K. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 50-80.

Lerch, F., Prietula, M., & Kulik, C. (1997). The Turing effect: The nature of trust in expert system advice. In P. Feltovich, K. Ford, & R. Hoffman, *Expertise in context: Human and machine* (pp. 417-448). Cambridge, MA: MIT Press.

Levs, J. (2012, January 15). *What caused the cruise ship disaster?* Retrieved from CNN: http://www.cnn.com/2012/01/15/world/europe/italy-cruise-questions/index.html

Lyons, J., Koltai, K., Ho, N., Johnson, W., Smith, D., & Shively, R. (2016). Engineering trust in complex automated systems. *Ergonomics in Design*, 13-17.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130.

Madhavan, P., & Weigmann, D. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, 277-301.

Merritt, S. (2011). Affective processes in human-automation interactions. *Human Factors*, 356-370.

Merritt, S., & Ilgen, D. (2008). Not all trust is created equal: Dispositional and history-based trust in automation interaction. *Human Factors*, 194-210.

Muir, B. (1987). Trust between human and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 527-539.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23-74.