Theoretical Issues in Ergonomics Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ttie20

The level of measurement of trust in automation

Jiajun Wei , Matthew L. Bolton & Laura Humphrey

To cite this article: Jiajun Wei , Matthew L. Bolton & Laura Humphrey (2020): The level of measurement of trust in automation, Theoretical Issues in Ergonomics Science, DOI: 10.1080/1463922X.2020.1766596

To link to this article: https://doi.org/10.1080/1463922X.2020.1766596

Published online: 14 Aug 2020.

THEORETICAL ISSUES IN ERGONOMICS SCIENCE
https://doi.org/10.1080/1463922X.2020.1766596

The level of measurement of trust in automation

Jiajun Wei[a], Matthew L. Bolton[a] and Laura Humphrey[b]

[a]Department of Industrial and Systems Engineering, University at Buffalo, The State University of New York, Buffalo, New York, USA; [b]Autonomous Controls Branch, Aerospace Systems Directorate, Air Force Research Laboratory, Wright-Patterson AFB, Ohio, USA

ABSTRACT
Psychometrics are increasingly used to evaluate trust in the automation of systems, many of them safety-critical. There is no consensus on what the highest level of measurement is for trust. This is important as the level of measurement determines what mathematics and statistics can be meaningfully applied to ratings. In this work, we introduce a new method for determining what the maximum level of measurement is for psychometrically assessed phenomenon. We use this to determine the level of measurement of trust in automation using human ratings about the behaviour of unmanned aerial systems performing search tasks. Results show that trust is best represented at an ordinal level and that it can be treated as interval in most situations. It is unlikely that trust in automation can be considered ratio. We discuss these results, their implications, and future research.

ARTICLE HISTORY
Received 31 August 2019
Accepted 5 May 2020

KEYWORDS
Psychometrics; measurement theory; trust; automation; level of measurement; scale of measure

Relevance to human factors/ergonomics theory

Treating a measure at the appropriate level of measurement is critical to performing meaningful mathematical operations and comparisons on psychometric data. By presenting a novel method for determining what level of measurement is most appropriate for a psychological concept that is being measured by a psychometric, this work is making a significant theoretical contribution to human factors and ergonomics theory. Further, by using this method to assess the maximum level of measurement that could be used for trust in automation, this work provides the human factors and ergonomics society with valuable information about how to model and analyze trust in automation.

1. Introduction

With the rise of autonomous systems and more sophisticated automation, researchers have become increasingly interested in designing automation that humans will trust. There are

CONTACT Matthew L. Bolton ✉ mbolton@buffalo.edu
This paper is an expanded version of a paper by the same authors that was presented at the 2019 International Annual Meeting of the Human Factors and Ergonomics Society (Wei, Bolton, and Humphrey 2019). This journal article reports the full set of experimental results (the conference paper only contained a small subset) and also presents a more in depth discussion.

different ways of measuring trust in automation (Hoff and Bashir 2015). Because trust is purely psychological (as opposed to a psychological manifestation of a physical phenomenon), it is typically measured using psychometric rating scales. For these, humans use introspection to convert their psychological state into a quantifiable rating on a predetermined scale. Unfortunately, it is not clear what the level of measurement is for trust.

A scale's level of measurement defines the relative meaning of numbers measured on that scale. As such, the level of measurement determines what mathematical and statistical comparisons and operations can be meaningfully employed on measures made on the scale (B. H. Cohen 2013; Stevens 1946). Thus it is important that measures be handled with respect to the most appropriate level.

Despite this importance, there does not appear to be any work that has investigated what level of measurement is most appropriate for trust in automation. Furthermore, there does not appear to be an established method for determining what the maximum level of measurement is for any given psychometric scale. In this work, we set out to fill this gap. To accomplish this, we introduce a new method for determining what the maximum level of measurement is for a psychological phenomenon measured by a psychometric. We then use this method to assess trust in automation. In what follows, we provide background on trust in automation and the levels of measurement of psychometrics. We then outline our method and describe how we used it to evaluate the maximum level of measurement of trust using a human subjects experiment with an unmanned aerial system (UAS) task. We report results of this analysis and discuss their implications for the measurement and modelling of trust in automation.

2. Background

2.1. Levels of measurement

In psychology, there are generally four levels of measurement (Stevens 1946): nominal, ordinal, interval, or ratio. At the nominal level, the numbers represent mutually exclusive categories or identities (i.e. player number on a sports team). At the ordinal level, the numbers only indicate order (i.e. class rank). For the interval level, the distances between numbers have meaning. However, because there is no meaningful zero (zero does not mean the complete absence of the phenomenon being measured), ratios are meaningless (i.e. temperature in Fahrenheit or Celsius). Finally, at the ratio level, ratios between numbers have meaning by virtue of there being a meaningful zero (i.e. distance).

The level of a given scale determines what mathematical and statistical operations can be meaningfully applied to numbers measured on that scale (Stevens 1946). Nominal scales are compatible with equalities (and inequalities), counts, modes, set membership, and contingency correlation; ordinal scales support greater-than and less-than comparisons, medians, percentiles, and rank-order statistics; interval scales allow for the computation of means, standard deviations, product moment correlations, and most parametric statistics; and ratio scales are compatible with percent changes, geometric means, coefficients of variation, and the full range of parametric statistics.

Critical to the determination of meaningful operations is the concept of permissible transformations. Scales that fall within each level of measurement can be converted to other scales at the same level that measure the same phenomenon through these permissible transformations. Nominal scales can be transformed into other nominal scales with a one-to-one transformation: a function that preserves the identity of each element. Ordinal scales can be converted to other ordinal scales with a strictly increasing function: a function that preserves the order of the elements. An interval scale can be transformed into another interval scale with a linear transformation of the form $y = a{\cdot}x + b$: a function that scales the measure on the original scale x to the new one y by scaling the original measure by a factor a and moving the position of the zero or intercept with b. Finally, ratio scales can be converted to other ratio scales with a ratio transformation $y = a{\cdot}x$, without the need to move the zero. These transformations determine what mathematical operations can be meaningfully performed on numbers. Specifically, for a comparison or mathematical operation between numbers on a given scale to be meaningful, it must hold when the numbers are permissibly transformed to different scales at the same level. Examples of this are shown in Table 1.

2.2. Levels of measurement and psychometrics

Trust is predominantly measured on psychometric scales. In psychometric rating scales, humans use introspection to convert some attribute of their psychological state or subjective experience into a number on a predetermined scale. These scales are used everywhere from review scores on Amazon, to disease diagnoses, to the engineering and design of safety-critical human-automation interaction (HAI). They are also widely used in scientific research in psychology, medicine, and engineering. In the latter two cases, this can include subjective assessment of trust in automation.

Although the full process is not always followed, psychometric scales are properly created and evaluated in a standardised, scientific process (Kline 1986). Most importantly, developed scales must be 'valid' and 'reliable' (Eignor 2013). Validity relates to the ability of the scale to actually measure the phenomenon it is intended to measure (Messick 1995). Practically, this is evaluated through subjective assessment by expert judges (face validity) or by showing that the measures collected on the scale correlate with things associated with the attribute being measured. Reliability relates to the ability of a scale to produce consistent results over repeated measurements and across participants (Ghiselli, Campbell, and Zedeck 1981).

Despite the rigours of scale development and evaluation, there is no clear consensus about the level of measurement of psychometric scales. Because of the mathematical and statistical power offered by the interval level's support for means, standards deviations, and most parametric statistics, practitioners prefer to treat most psychometric ratings as interval

Table 1. Meaningful and Meaningless Expressions Based on Transformations.

	If X and Y are interval with	If X and Y are ratio with
Expression	$f(x) = ax + b$	$f(x) = ax$
$x_1 - x_2 = k(x_3 - x_4)$	$f(x_1) - f(x_2) = k(f(x_3) - f(x_4))$ $\therefore (ax_1+b) - (ax_2+b) = k((ax_3+b) - (ax_4+b))$ $\therefore x_1 - x_2 = k(x_3 - x_4)$ $\therefore$ The expression is **meaningful**	$f(x) = ax + b\ f(x_1) - f(x_2) = k(f(x_3) - f(x_4))$ $\therefore (ax_1) - (ax_2) = k((ax_3) - (ax_4))$ $\therefore x_1 - x_2 = k(x_3 - x_4)$ $\therefore$ The expression is **meaningful**
$x_1 = kx_2$	$f(x_1) = k\,f(x_2)$ $\therefore ax_1+b = k(ax_2+b)$ $\therefore x_1 = kx_2+(k-1)b/a$ $\therefore$ The expression is **meaningless**	$f(x_1) = k\,f(x_2)$ $\therefore ax_1+b = kax_2$ $\therefore x_1 = kx_2$ $\therefore$ The expression is **meaningful**

X and Y are numerical sets at a given level of measurement; $x_1 \ldots x_4 \in X$; $f(x)$ is a function $f: X \rightarrow Y$; a and b are constants; and $\therefore$ is 'therefore.' An expression is meaningful in a scale if it holds after transforming each $x \in X$ with f.

(Furr and Bacharach 2013; Guilford 1954). However, there is controversy around this subject. Most psychometrics experts do not think psychometric scales are capable of providing ratio measures (Furr and Bacharach 2013; Guilford 1954), yet common measures such as the NASA-TLX (for measuring mental workload) (Hart and Staveland 1988) and ratings used in predictive trust models (Lee and Moray 1992; Muir 1987) appear to treat psychometrics as if they are ratio. In contrast, many researchers in ergonomics and measurement theory (Annett 2002; Barrett 2003; Cliff and Keats 2003; Michell 1997, 2008; Trendler 2009) doubt that subjective ratings can be treated as anything more than ordinal. In fact, Stevens (1951) himself held this view, stating that 'as a matter of fact, most of the scales used widely and effectively by psychologists are ordinal scales.'

Researchers that use subjective psychometric rating scales appear to avoid these concerns by relying on Stevens' (Stevens 1975) definition of psychological measurement: 'the assignment of numerals to objects and events according to a rule.' That is, as long as the rule used for obtaining measures on a psychometric scale is consistent with a given level of measurement, then the data can be treated as if it is on that level. For example, in the NASA-TLX (Hart and Staveland 1988) (or other similar rating scales), a participant gives a rating by marking a position on a line on a sheet of paper. An analyst then measures the position of the mark to obtain a rating. Because this measurement is done on a ratio scale, the data collected by NASA-TLX is treated as ratio in supported calculations. However, this practice is suspect because the rule used for assigning a number to a scale does not inherently speak to the level of the psychological quantity being measured. Thus, while still widely used, Stevens' definition of measurement is not well accepted among measurement theorists (Luce 1997).

2.3. Levels of measurement and psychophysics

While we have not been able to identify any specific studies that investigate the maximum level of measurement of psychological phenomena that are measured with psychometrics, there have been such efforts within psychophysics. Psychophysics represent the human's psychological representation of measurable physical quantities.

The psychophysics that have been subjected to level-of-measurement analyses relate to Stevens' power law (Stevens 1956), which links physical stimulus intensity to its perceived intensity. To produce a power law, humans make ratio judgments about the relative magnitudes of different stimuli represented on a ratio scale. Prominent researchers have expressed scepticism that humans are capable of making true ratio judgments (Laming 1997). Attempts have been made to check this (Ellermeier and Faulhammer 2000; Zimmer 2005) by assessing whether judged ratio differences between measured physical stimuli follow multiplicative and commutative properties. These found that judgments satisfied the commutative property, but not the multiplicative one. While this is sufficient to indicate that humans can make ratio judgments in power law experiments, it would be more convincing if both properties held (Narens 1996). Further, Bolton (2008) found that psychophysical power laws could be fit to ordinal numbers generated in computationally simulated power law experiments.

This work is relevant because it shows that there are serious doubts about the level of measurement used for psychological phenomena, even when they are representations of physical ratios. The work also shows that the level of measurement can be evaluated by checking for properties (i.e. multiplicativity and commutativity) between measurements. However, comparable evaluations of psychometrics are more challenging because, unlike with psychophysics, there are no physical measures that can be used as the basis for comparisons.

2.4. Trust in automation

Trust in automation and its measurement with psychometric rating scales has become increasingly important as systems become more and more automated. Accurate measures of human trust will be needed, for instance to identify cases in which humans have 'mis-calibrated' trust in autonomy (Hoffman et al. 2013). That is, accurate measures can help autonomy designers identify and address issues such as under-reliance on autonomy (low trust in reliable autonomy) and over-reliance on autonomy (high trust in unreliable autonomy). Below we discuss the definition of trust in automation used in this research, the components of trust in automation, and the way trust measures have been treated in the literature.

2.4.1. Definition of trust in automation

There are many definitions of trust in the literature (Lee and See 2004). For the purpose of this work, we adopt the definition of Lee and See (2004) that trust is 'the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability' because it has been been widely adopted in the human-automation interaction and human factors communities.

This definition of trust as an attitude suggests that trust is a synthesis or evaluation of beliefs into a trust continuum on which one thing may be more trusted than another. Thus, trust can be at least ordinal. There is nothing in the definition to indicate whether or not there can be meaningful distances between levels of trust. Thus trust could be interval. If at least an interval level of measurement can be assumed, it is possible that trust as an attitude could be ratio. This is because it is conceivable that zero trust could constitute the absolute minimum trust attitude a person could have about something: that no trust exists for the given target at all. However, it is also conceivable that no such zero exists. Given that the widely accepted definition of trust does not inherently suggest that trust is at a given level of measurement, there is a need for an empirical evaluation of this.

2.4.2. The components of trust

There are different components of trust based on who the person giving trust is and the trustee. For human trust in automation, performance, process and purpose have been identified as the three bases of trust in automation (Lee and Moray 1992; Lee and See 2004). Performance is the 'current and historical operation of the automation and includes characteristics such as reliability, predictability and ability' (Lee and See 2004); process is the 'degree to which the automation's algorithms are appropriate for the situation and able to achieve the operator's goals' (Lee and See 2004); and purpose is the 'degree to which the automation is being used within the realm of the designer's intent' (Lee and See 2004).

We will ultimately use these different components of trust to inform how we elicit different trust levels from human subjects in our planned experiment.

2.4.3. Usage of trust measures

Given that the dominant definitions of trust in automation do not precisely define its level of measurement, we surveyed the literature to determine what level of measurement the community has used for trust. Because nobody explicitly describes the level of measurement that they use, this had to be inferred by the statistics and model relationships that were employed. This has revealed that researchers treat trust as being at ordinal, interval, and ratio scales.

Ordinality is generally revealed through the use of non-parametric statistics (such as the Kruskal–Wallis test and rank order correlations) to assess or discover significant differences in trust ratings between experimental conditions (Clare, Cummings, and Repenning 2015; Cramer et al. 2009; T. A. Kazi et al. 2005; T. Kazi et al. 2007; Ma and Kaber 2007; Perkins et al. 2010; Wei and Bell 2012).

Intervality was assumed to be revealed through the use of parametric statistics such as t-test, analyses of variance, and standard deviations from means. This is by far the dominant practice in the research community (see for example, Abe and Richardson 2006; Bagheri and Jamieson 2004; Bailey and Scerbo 2007; Bass, Baumgart, and Shepley 2013; Biros, Daly, and Gunsch 2004; Bisantz and Seong 2001; Davenport and Bustamante 2010; De Vries and Midden 2008; Dzindolet et al. 2003; Madhavan, Wiegmann, and Lacson 2006; Manzey, Reichenbach, and Onnasch 2012; Pak et al. 2017; Rovira, McGarry, and Parasuraman 2007; Rovira, Pak, and McLaughlin 2017; Rovira and Parasuraman 2010; Seong and Bisantz 2008; Visser et al. 2012; Visser and Parasuraman 2011; Wang, Jamieson, and Hollands 2009).

While less common than the other two levels, there is also evidence of researchers treating trust at a ratio level. For example, Lee and Moray (1992) studied the dynamics of trust in a supervisory control simulation. They calculated percentage change as an index for showing the dynamics of trust. Since percentage change is a statistic that can only be meaningfully applied to ratio data (Stevens 1946), we infer that they assumed trust is at the level of ratio.

3. Objective

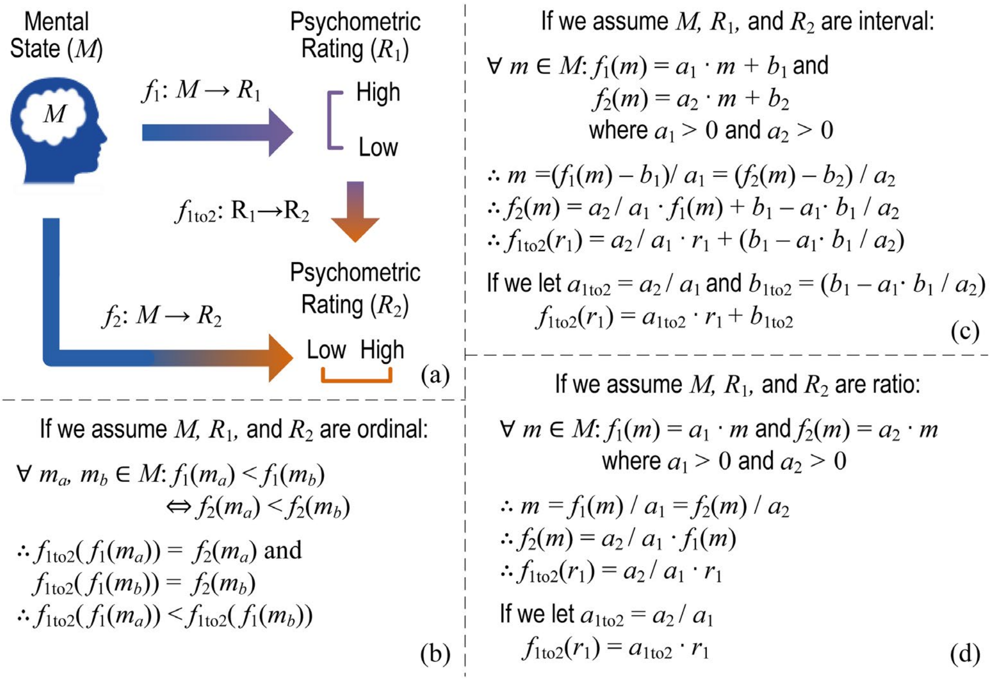
As the discussion above shows, the widely accepted definition of trust does not provide a clear view of what its level of measurement is. Furthermore, there is no definitive consensus about the level of measurement of trust. Because trust in automation is currently being used as a critical dimension for assessing and modelling human-automation interaction in emergent technologies, it is critical that we understand its level of measurement. This is true for two reasons. First, if the level of measurement is lower than how most researchers are treating it, then there is a risk of analysts using meaningless statistics and models for psychometric evaluations of trust. Second, if analysts are using levels of measurement below what trust actually is, they are missing out on powerful analysis and modelling possibilities.

However, there is no established method for determining what the maximum possible level of measure should be for a psychometric. Thus, in this research, we set out to identify such a method and use it to determine the level of measurement of trust. Below we present our method and describe how we used it to assess the level of measurement of trust using a human subjects experiment.

4. Method for assessing the level of measurement of a psychometric

We have developed a method for assessing the maximum level of measurement of psychometrics. To accomplish this, our method exploits meaningful transformations between scales at the same level of measurement. The relationship we use for this is shown in Figure 1.

In this, (Figure 1(a)) we assume two psychometric scales R_1 and R_2 that both measure the same psychological quality M without losing power by transforming M to a lower level. When asked to provide a rating for the same psychological quality and condition on these scales, we hypothesise that the human will implicitly apply transformations f_1 to convert M to R_1 and f_2 to convert M to R_2 respectively. While an analyst is not able to observe M, f_1, or f_2, he or she can find a model $f_{1\text{to }2}$ that converts between observed values made on R_1 to those on R_2. In this context, the form of $f_{1\text{to }2}$ gives us an indirect means of determining the level of measurement most appropriate for measuring M.


Mental State (M)
M
f_1: $M \rightarrow R_1$
Psychometric Rating (R_1)
High
Low
$f_{1\text{to}2}$: $R_1 \rightarrow R_2$
f_2: $M \rightarrow R_2$
Psychometric Rating (R_2)
Low High
(a)

If we assume M, R_1, and R_2 are ordinal:

$\forall\ m_a, m_b \in M$: $f_1(m_a) < f_1(m_b)$
$\Leftrightarrow f_2(m_a) < f_2(m_b)$

$\therefore f_{1\text{to}2}(f_1(m_a)) = f_2(m_a)$ and
$f_{1\text{to}2}(f_1(m_b)) = f_2(m_b)$
$\therefore f_{1\text{to}2}(f_1(m_a)) < f_{1\text{to}2}(f_1(m_b))$

(b)

If we assume M, R_1, and R_2 are interval:

$\forall\ m \in M$: $f_1(m) = a_1 \cdot m + b_1$ and
$f_2(m) = a_2 \cdot m + b_2$
where $a_1 > 0$ and $a_2 > 0$

$\therefore m = (f_1(m) - b_1) / a_1 = (f_2(m) - b_2) / a_2$
$\therefore f_2(m) = a_2 / a_1 \cdot f_1(m) + b_1 - a_1 \cdot b_1 / a_2$
$\therefore f_{1\text{to}2}(r_1) = a_2 / a_1 \cdot r_1 + (b_1 - a_1 \cdot b_1 / a_2)$

If we let $a_{1\text{to}2} = a_2 / a_1$ and $b_{1\text{to}2} = (b_1 - a_1 \cdot b_1 / a_2)$
$f_{1\text{to}2}(r_1) = a_{1\text{to}2} \cdot r_1 + b_{1\text{to}2}$

(c)

If we assume M, R_1, and R_2 are ratio:

$\forall\ m \in M$: $f_1(m) = a_1 \cdot m$ and $f_2(m) = a_2 \cdot m$
where $a_1 > 0$ and $a_2 > 0$

$\therefore m = f_1(m) / a_1 = f_2(m) / a_2$
$\therefore f_2(m) = a_2 / a_1 \cdot f_1(m)$
$\therefore f_{1\text{to}2}(r_1) = a_2 / a_1 \cdot r_1$

If we let $a_{1\text{to}2} = a_2 / a_1$
$f_{1\text{to}2}(r_1) = a_{1\text{to}2} \cdot r_1$

(d)

Figure 1. (a) Shows transformations between mental state M and scales R_1 and R_2. (b), (c), and (d) show that if M, R_1, and R_2 are ordinal (b), interval (c), or ratio (d) then $f_{1\text{to }2}$ is ordinal, interval, or ratio respectively. In all of the above, a_1, a_2, b_1, and b_2 are constants.

Specifically, by collecting psychometric ratings of M on two different scales (R_1 and R_2) for identical conditions, the level of measurement should be revealed by the transformation for converting measures collected on one scale to the other ($f_{1\text{to}\,2}$ in Figure 1).

If M is best represented at an ordinal level (Figure 1(b)), f_1 and f_2 will be ordinal transformations and thus $f_{1\text{to}\,2}$ will be ordinal. If M is best represented at an interval level (Figure 1(c)), f_1 and f_2 will be linear transformations and $f_{1\text{to}\,2}$ will also be linear. If M is best represented at a ratio level (Figure 1(d)), f_1 and f_2 will be ratio transformations and $f_{1\text{to}\,2}$ will also be ratio.

Because both ratio and interval transformations are in a linear form, characterising a transformation between any two data series observed on two different psychometric scales can be accomplished through a regression analysis. Because there can be error in the observation of both the predictor and the predicted measures, our method uses Deming regression (Deming 1943): a linear regression model that is able to account for this condition. Given that Deming regression does not use least squares in its fitting process, R^2 is not used. Thus, for this work, we use a Pearson's correlation coefficient (r) as the standard, regression-model-independent measure of how linearly related two measures are.

Thus, statistics produced by analyzing the relationship between R_1 and R_2 will give us the means to identify the measurement level of M. If there is not a strong non-parametric correlation between R_1 and R_2 (if they have a low Spearman's ρ), the data will not suggest a monotonically increasing relationship between measures and M will be at least **nominal**. If there is a strong non-parametric correlation between R_1 and R_2, the data will suggest a monotonically increasing relationship between measures and M will be at least **ordinal**. If there is a strong linear relationship between R_1 and R_2 (indicated by a Pearson's r) and a Deming regression model has a significant intercept, then M will be **interval**. If there is a strong linear relationship between R_1 and R_2 and the regression model does not have a significant intercept, then M will be **ratio**.

In this method, human judgments on only two scales are necessary for determining the level of measurement of a psychological attribute. However, by using more we can reduce the chance that any set will have the same arbitrary zeros. Thus, we use three scales to reduce the risk of concluding that a psychological phenomenon is ratio when it is actually interval.

5. Methods

We used a human subjects experiment to evaluate the level of measurement of trust. This study received approval from the University at Buffalo IRB under STUDY00002118.

5.1. Procedure

This experiment had participants arrive at the laboratory and sign an informed consent document. Participants observed a PowerPoint presentation that introduced them to the experimental task. Note that because trust of a given system has been shown to vary as a function of trust at previous times (a time series; Lee and Moray 1992), participants were instructed to treat the UAS in each trial as a separate, independent system. They then performed the experiment in which they watched simulations of UASs performing search tasks. The same simulations were observed in three blocks, where humans rated how much they would trust the automated controller they observed using three different judgment methods.

5.2. Participants

We recruited 36 University at Buffalo student participants. 13 were female and 23 were male. Their average age was 26.

5.3. Materials and apparatus

The experiment was run in a controlled, quiet, evenly-lighted laboratory. It was administered on desktop computers resting on a computer desk in front of which a participant would sit. Computers were equipped with 21 inch LCD monitors, optical mice, keyboards, and physical knobs (see Figure 3). The experiment was administered on the computers using software that was created for this project.

During the experiment, the software would depict a video of a UAS flying around a given area and performing search tasks (Figure 2). The simulations were created using UxAS and AMASE (Rasmussen, Kingston, and Humphrey 2018). This enabled simulations to represent realistic UAS dynamics and route planning. The UAS was depicted as a blue chevron shape

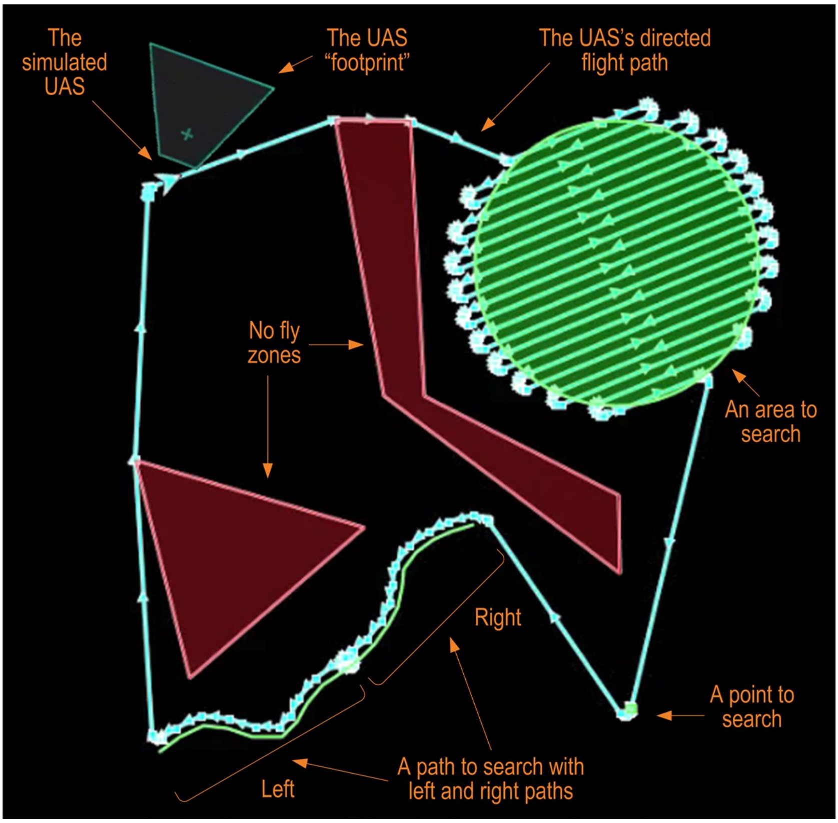


Figure 2. A screen of the UAS simulation. Annotations that did not actually appear in presented simulations are shown in orange.

moving through the area. A 'footprint' of the UAS's camera also showed the ground area the camera was capturing. A cross in the footprint indicated the centre of the camera's view. The smaller the footprint, the more focussed the camera.

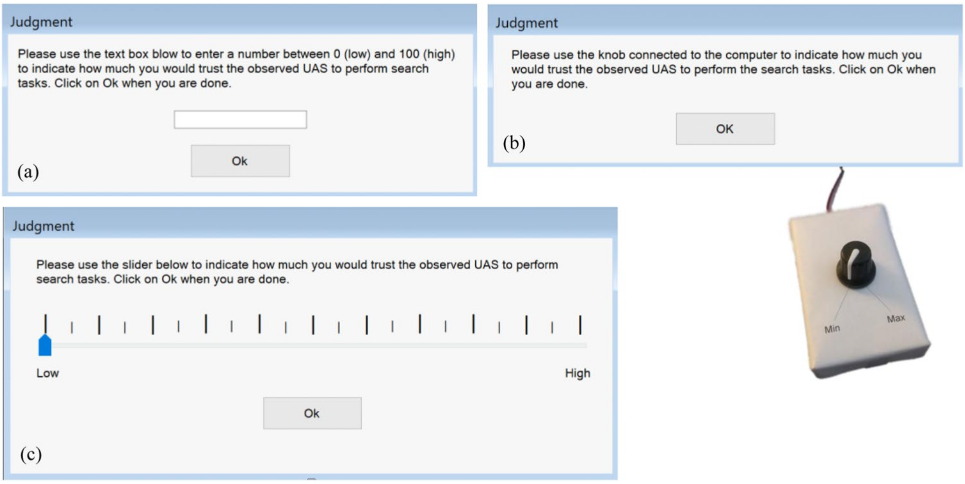
In simulations, the UAS always started in the upper left side of the area. The UAS was expected to complete three search tasks. In an area search, the UAS would search (cover) the space encompassed by the green circle with the camera footprint. In a point search, the UAS would have the footprint's cross pass over a specific spot in the lower right of the area. In a path search, the UAS would have the footprint's cross pass over the entirety of the green line. When all tasks were complete, the UAS would return to the starting point and loiter there. The UAS was expected to avoid flying into the two 'no fly zones' (red shapes). When the UAS's planned flight path was shown (as in Figure 2), it was depicted as a blue line.

After each simulation, participants were asked to provide ratings about their trust in the UAS with either (Figure 3): (a) a number between 0 and 100 (note that training explicitly showed that decimal values could be entered), (b) the position of a physical knob, or (c) the position of an on-screen slider.

5.4. Independent variables

The independent variables all related to the experimental trials. Specifically, trials varied along dimensions that would exhibit different levels of trust. This trial geometry included the possibility of all the factors shown in Table 2.

These factors were selected because their variation should produce a range of trust responses from participants. Specifically, each related to the 'three Ps' (Lee and See 2004) of automation that influence trust: its *purpose*, the *process* it uses, and its *performance*. The variety of tasks the UAS undertakes relate to *purpose*. The *Order*, *Density*, and *Path* relate to *process*. *Error*, *Skip*, and *NoFly* all relate to *performance*.



Judgment

Please use the text box blow to enter a number between 0 (low) and 100 (high) to indicate how much you would trust the observed UAS to perform search tasks. Click on Ok when you are done.

Judgment

Please use the knob connected to the computer to indicate how much you would trust the observed UAS to perform the search tasks. Click on Ok when you are done.

Judgment

Please use the slider below to indicate how much you would trust the observed UAS to perform search tasks. Click on Ok when you are done.

Figure 3. Software dialog boxes used for collecting human trust ratings. (a) Participants would enter a number between 0 and 100. (b) Participants would turn the physical knob connected to the computer. (c) Participants would use the computer's mouse to move a slider.

Table 2. The Scenario Geometry for UAS Simulations.

Variable	Description	Levels
Path	The UAS could show or not show its flight path	{Visible, Invisible}
Error	The UAS could fly its path and control its camera with levels of error (random turns and jitters)	{0, 0.2, 0.4, 0.6, 0.8, 1}
Order	The UAS could execute search tasks in any order	All of the possible orders
Skip	The UAS could skip up to one task or part of the line search	{None, Area, Point, Right_{Of} Line, Left_{Of} Line}
Density	The UAS could execute area searches with different densities (based on the camera's footprint size)	{Low, Medium, High, Highest}
NoFly	The UAS could fly into 'no fly zones'	{Occurs, DoesNotOccur}

Error levels are the proportion of global maximums used as local maximums for uniformly distributed error. For the UAS, the global maximum was 0.001° for latitude and longitude and 0.2 for rotation radians. For the footprint, the global maximum was 0.0003° for the latitude and longitude of each point boundary point.

5.5. Dependent measures

The dependent measures were human trust ratings made using each of the three judgment modalities (Figure 3). With the ask modality (Figure 3(a)), human trust was measured as a floating-point number from 0 to 100. With the knob (Figure 3(b)), human trust was measured as a floating-point number from 0 to 100 based on the position of knob between its minimum (0°) and maximum (300°) positions. With the slider (Figure 3(c)), human trust was measured as a floating-point number from 0 to 100 based on the left-to-right position of the slider.

5.6. Experimental design

We created a set of 96 trials: for each of the six possible *Error* levels, we generated 16 different trials. These had every possible combination where *Skip* was or was not None, each possible value of *Path*, and each possible value of *Density*. For each of the trials where *Skip* was not None, one of the options for *Skip* (see Table 2) was randomly assigned as well as a random *Order*. In 9 trials, the UAS flew into a no fly zone. We randomly selected 30 trials for use in the actual experiment. In 2 of these, the UAS entered a no fly zone. It is worth noting that because this work was only interested in eliciting a range of trust responses from participants (not analyzing the impact of different factors on trust), we did not employ a factorial design. This is discussed in more depth in Section 7.1.

Four additional training trials were selected that exhibited variation along all the scenario geometry dimensions. Two additional training trials, representing best and worst performance conditions, were also created. The best performance trial had the UAS complete all search tasks with no error, at the highest search density, and in the most efficient order. The UAS in the worst performance trial had the highest level of error and randomly flew through the search area, including no fly zones.

A participant was assigned three random orders of the 30 experimental trials, one for each of the three judgment modalities. Trials for a given modality were presented in blocks. Block order was counterbalanced between participants.

Training trials were presented in a consistent order. At the beginning of the experiment, participants saw training to introduce them to the experimental task and first judgment modality. In this, participants saw the 'best' trial, then the 'worst' trial, then four other trials. On-screen instructions introduced judgment modalities and scenario geometry features in each trial. Subsequent training blocks of three trials (which excluded the best and worst conditions) were presented between judgment modalities to introduce participants to the new modality. Training trial and presentation orders were consistent between participants regardless of the given judgment modality order.

5.7. Data analysis

For each participant, we used our new method to assess the level of measurement of trust by calculating non-parametric (Spearman's ρ) and parametric correlations (Pearson's r) and fitting Deming regression models between the judgments made for the different modalities. To determine if a regression model had a significant intercept, we used the jackknife method (NCSS 2016) to calculate a 95% confidence interval around the intercept and checked if it contained 0.

Using these statistics, we developed a heuristic to interpret results. This enabled us to determine if a given model provided weak or strong evidence that trust was at least at a given level of measurement and to synthesise evidence across a participant's models to draw conclusions about the level of measurement of trust. For each model: (a) Evidence for nominality was assumed by default. (b) Evidence for ordinality was expressed by a weak Spearman's correlation ($\rho \geq 0.1$; J. Cohen 1988). (c) Weak evidence for intervality was indicated by a moderate Pearson's correlation ($r \geq 0.3$). (d) Strong evidence for intervality was indicated by a strong Pearson's correlation ($r \geq 0.5$). (e) Weak evidence for a ratio scale was indicated by evidence for intervality and a non-significant intercept. (f) Strong evidence for a ratio scale was indicated by strong evidence for intervality, a non-significant intercept, and a small (20 unit) 95% confidence interval around the intercept. This heuristic is summarised in Table 3.

Then, across all three models for each participant: (a) Strong evidence of normality was assumed. (b) Weak evidence of ordinality was assumed if one or more models provided evidence of ordinality. (c) Strong evidence of ordinality was assumed if two or more models provided evidence of ordinality. (d) Weak evidence of intervality was assumed if two or more models provided evidence of intervality. (e) Strong evidence of intervality was assumed if two or more models provided strong evidence of intervality. (f) Weak evidence of a ratio level was assumed if all models had weak evidence of a ratio level. Note that this required every model to not have a significant intercept. This is because evidence of any intercept would indicate non-ratio trust. (g) Strong evidence of a ratio level was assumed if all the

Table 3. Heuristic Used to Assess Whether a Given Model Provided Weak or Strong Evidence of Trust in Automation Being at a Given Level of Measurement.

Level of Measurement	Evidence Strength Weak ○	 Strong ●
Nominal	………………… Assumed ……………………………	
Ordinal	…………………… $\rho \geq$ 0.1……………………………………	
Interval	$r \geq 0.3$	$r \geq 0.5$
Ratio	$r \geq 0.3$ and $0 \in$CI	$r \geq 0.5$ and $0 \in$CI and $\lvert$CI$\rvert \geq 20$

Entries spanning columns (using…) indicate the criteria for general evidence with not strong or weak designation. CI stands for confidence interval.

Table 4. Heuristic Used Across All of a Participant's Models to Determine If Weak or Strong Evidence of a Level of Measurement Was Provided.

Level of Measurement	Evidence Strength	
	Weak ○	Strong ●
Nominal	................................Assumed................................	
Ordinal	1+ with Evidence of Ordinal	2+ with Evidence of Ordinal
Interval	2+ with Evidence of Interval	2+ with Strong Evidence of Interval
Ratio	3 with Evidence of Ratio	3 with Evidence of Ratio, 2+ with Strong Evidence

Evidence (Weak, Strong, or otherwise) is based on the heurstic described in Table 3. Entries spanning columns (using…) indicate the criteria for general evidence with no strong or weak designation. CI stands for confidence interval.

models exhibited evidence of a ratio level and two or more exhibited strong evidence of this. See Table 4 for a summary of this.

We also analyzed our results across participants by computing Spearman ρ and Pearson r correlations and Deming regression models for the three judgment modality pairs. We analyzed these results using the same heuristics employed for the individual participant's models (Table 3).

6. Results

Analysis results and the synthesis of all three models for each participant are reported in Table 5 and Figure 4. Analyses revealed that only one participant (participant 25) exhibited strong evidence of a ratio level of measure for trust. Of the remaining participants, only five had weak evidence of a ratio level. Conversely, only two participants (3 and 22) showed no evidence of an interval level. Thirty four of the 36 participants showed evidence of an interval level, with 19 of these having strong evidence. All but one of the participants (22) had evidence of an ordinal level of measurement, where all but three of them were strong.

The aggregate results across all participants are shown in Table 6 and Figure 4. These results show that, when all of the data are considered together, the experiment exhibits strong evidence for the intervality of trust in automation.

7. Discussion and conclusions

This work constitutes the first research to identify the level of measurement of trust in automation. There is consistency in our results. Because all but one participant showed evidence of ordinality, and higher levels can always be accommodated by a lower level, it is safest to treat trust as ordinal. However, only two participants did not exhibit evidence of interval-level trust and the majority of participants had strong evidence for this level. Thus, given the significant increase in mathematical power offered by the interval level, our results indicate that it is safe to treat trust as interval. Conversely, only six participants exhibited evidence of a ratio level, and only one had strong evidence for this. This suggests that while some people may think about trust at a ratio level, it is not common.

Our results do suggest that analysts should be extremely careful when handling subjective trust data. This is because some people are clearly only treating trust as if it is ordinal and some treat it as if it is ratio. Experimental results and models of trust that have processed

Table 5. Analysis of Individual's Experimental Results.

	y - Ask, x - Knob				y - Ask, x - Slider				y - Knob x - Slider				At Least			
ID	ρ	Model	Intercept CI	r	ρ	Model	Intercept CI	r	ρ	Model	Intercept CI	r	N	O	I	R
1	0.81	y = 0.59 x + 21.43*	[9.89, 32.97]	0.79	0.56	y = 0.91 x + 5.69	[−12.99, 24.37]	0.59	0.67	y = 1.54 x − 26.69	[−54.13, 0.75]	0.64	●	●	●	
2	0.37	y = 0.56 x + 38.59*	[22.27, 54.91]	0.47	0.51	y = 0.57 x + 45.33*	[32.84, 57.82]	0.53	0.33	y = 1.03 x + 12.05	[−3.06, 27.16]	0.39	●	●	○	
3	0.06	y = 0.93 x − 1.10	[−29.26, 27.06]	0.19	0.27	y = 1.12 x − 8.44	[−38.52, 21.63]	0.29	0.05	y = 1.20 x − 7.91	[−33.56, 17.73]	0.13	●	○		
4	0.93	y = 0.60 x + 22.86*	[18.00, 27.72]	0.96	0.90	y = 0.94 x + 4.29	[−2.76, 11.34]	0.94	0.90	y = 1.57 x − 31.00*	[−43.52, −18.49]	0.95	●	●	●	
5	0.73	y = 0.75 x + 17.35	[−4.49, 39.20]	0.70	0.91	y = 0.89 x + 14.34*	[0.36, 28.32]	0.91	0.76	y = 1.18 x − 4.00	[−20.49, 12.48]	0.78	●	●	●	
6	0.42	y = 0.76 x + 8.24	[−8.52, 25.00]	0.53	0.65	y = 0.95 x + 7.49	[−3.43, 18.42]	0.68	0.61	y = 1.25 x − 0.99	[−20.76, 18.78]	0.64	●	●	●	○
7	0.16	y = 0.56 x + 33.15	[−1.39, 67.70]	0.18	0.37	y = 0.54 x + 37.86*	[27.46, 48.26]	0.47	0.10	y = 0.97 x + 8.43	[−35.42, 52.28]	0.35	●	●	○	
8	0.69	y = 0.82 x + 14.42	[−1.19, 30.04]	0.72	0.56	y = 1.00 x − 2.98	[−30.32, 24.35]	0.54	0.54	y = 1.22 x − 21.21	[−45.87, 3.46]	0.53	●	●	○	○
9	0.13	y = 1.27 x − 43.92	[−1147.92, 1060.07]	0.02	0.33	y = 0.77 x + 19.35	[−0.45, 39.15]	0.33	0.57	y = 0.60 x + 49.92*	[32.34, 67.51]	0.56	●	●	○	
10	0.49	y = 0.62 x + 23.95*	[10.29, 37.61]	0.53	0.47	y = 0.95 x + 2.19	[−30.92, 35.31]	0.52	0.55	y = 1.53 x − 34.88	[−83.28, 13.53]	0.63	●	●	●	
11	0.52	y = 0.70 x + 26.44	[−0.34, 53.23]	0.46	0.53	y = 0.85 x + 33.30*	[11.90, 54.69]	0.57	0.29	y = 1.20 x + 9.75	[−11.11, 30.60]	0.39	●	●	○	
12	0.37	y = 0.58 x + 39.23*	[24.83, 53.62]	0.34	0.53	y = 0.92 x + 7.88	[−7.89, 23.66]	0.63	0.17	y = 1.59 x − 54.17*	[−90.89, −17.45]	0.20	●	●	○	
13	0.33	y = 0.74 x + 4.31	[−21.98, 30.61]	0.31	0.66	y = 0.67 x + 29.64*	[20.71, 38.57]	0.67	0.38	y = 0.90 x + 34.04*	[15.29, 52.78]	0.40	●	●	○	
14	0.75	y = 0.59 x + 29.20*	[13.27, 45.14]	0.78	0.64	y = 0.79 x + 35.76*	[23.66, 47.87]	0.63	0.47	y = 1.33 x + 11.06	[−16.56, 38.67]	0.52	●	●	●	
15	0.68	y = 0.77 x + 8.59*	[0.18, 17.00]	0.79	0.66	y = 1.08 x − 13.38*	[−24.43, −2.34]	0.81	0.73	y = 1.40 x − 28.67*	[−45.07, −12.27]	0.82	●	●	●	
16	0.54	y = 0.50 x + 40.76*	[29.59, 51.93]	0.54	0.44	y = 0.82 x + 32.66*	[18.81, 46.52]	0.47	0.48	y = 1.64 x − 16.12	[−39.23, 6.98]	0.53	●	●	●	
17	0.69	y = 0.44 x + 32.30*	[27.24, 37.36]	0.75	0.56	y = 0.93 x + 1.37	[−13.38, 16.11]	0.66	0.75	y = 2.11 x − 70.26*	[−102.17, −38.34]	0.79	●	●	●	
18	0.30	y = 0.50 x + 39.17*	[28.17, 50.18]	0.37	0.37	y = 0.89 x + 0.27	[−45.34, 45.88]	0.43	0.51	y = 1.78 x − 78.28	[−160.90, 4.33]	0.61	●	●	○	
19	0.63	y = 1.15 x − 25.64	[−126.00, 74.72]	0.67	0.63	y = 0.82 x + 16.38	[−7.53, 40.30]	0.87	0.62	y = 0.71 x + 36.57	[−19.58, 92.71]	0.69	●	●	●	○
20	0.59	y = 0.85 x + 4.10	[−23.34, 31.54]	0.65	0.55	y = 0.77 x + 22.17*	[8.46, 35.88]	0.57	0.39	y = 0.90 x + 21.24	[−9.31, 51.79]	0.54	●	●	●	
21	0.51	y = 0.55 x + 31.47*	[22.61, 40.34]	0.55	0.60	y = 1.01 x + 4.59	[−8.58, 17.75]	0.66	0.68	y = 1.85 x − 49.31*	[−72.70, −25.92]	0.71	●	●	●	
22	−0.15	y = −1.22 x + 118.54*	[94.18, 142.90]	−0.10	−0.05	y = −1.05 x + 110.42	[−435.57, 656.41]	−0.02	0.09	y = 0.86 x + 6.65	[−176.58, 189.89]	0.06	●			
23	0.37	y = 0.61 x + 7.95	[−7.79, 23.69]	0.31	0.06	y = 0.86 x + 7.49	[−147.17, 162.16]	0.09	0.04	y = −1.42 x + 130.54	[−413.61, 674.69]	−0.03	●	○	○	
24	0.49	y = 0.73 x + 3.20	[−14.64, 21.04]	0.52	0.40	y = 0.96 x + 0.95	[−15.74, 17.64]	0.36	0.55	y = 1.31 x − 3.08	[−21.52, 15.35]	0.58	●	●	●	○
25	0.59	y = 1.12 x + 0.68	[−8.21, 9.56]	0.61	0.60	y = 1.28 x − 0.94	[−8.97, 7.10]	0.60	0.41	y = 1.14 x − 1.44	[−8.67, 5.79]	0.35	●	●	●	●
26	0.78	y = 0.93 x − 10.35	[−30.55, 9.86]	0.76	0.82	y = 0.79 x + 9.40	[−0.58, 19.38]	0.86	0.63	y = 0.85 x + 21.18	[3.77, 38.60]	0.67	●	●	●	
27	0.41	y = 0.65 x + 23.63	[−1.18, 48.43]	0.27	0.45	y = 0.94 x + 12.65	[−9.32, 34.63]	0.46	0.58	y = 1.46 x − 17.01	[−49.36, 15.34]	0.56	●	●	○	
28	0.25	y = 0.70 x + 18.60	[−17.03, 54.23]	0.30	0.17	y = 1.27 x − 21.00	[−72.23, 30.23]	0.37	0.59	y = 1.80 x − 56.42*	[−97.55, −15.28]	0.50	●	●	○	
29	0.59	y = 1.25 x − 38.75	[−186.82, 109.32]	0.34	0.80	y = 0.79 x + 22.20*	[2.04, 42.35]	0.75	0.57	y = 0.63 x + 48.63*	[12.82, 84.43]	0.47	●	●	○	
30	0.45	y = 0.53 x + 27.98*	[12.10, 43.85]	0.49	0.75	y = 1.30 x − 18.31	[−46.59, 9.96]	0.75	0.36	y = 2.43 x − 86.54*	[−149.95, −23.13]	0.41	●	●	○	
31	0.71	y = 0.79 x + 18.09	[−6.04, 42.22]	0.86	0.57	y = 1.12 x − 8.67	[−56.73, 39.39]	0.63	0.63	y = 1.41 x − 33.66	[−83.17, 15.84]	0.76	●	●	●	○
32	0.81	y = 0.62 x + 13.53*	[3.66, 23.40]	0.79	0.75	y = 0.97 x + 18.61*	[10.21, 27.01]	0.74	0.81	y = 1.56 x + 8.19	[−4.18, 20.55]	0.73	●	●	●	
33	0.09	y = 0.75 x + 12.69	[−13.36, 38.74]	0.14	0.10	y = 0.93 x + 16.17	[−5.05, 37.38]	0.30	0.17	y = 1.24 x + 4.62	[−31.70, 40.95]	0.19	●	○	○	
34	0.57	y = 0.75 x + 2.82	[−17.90, 23.53]	0.52	0.59	y = 1.43 x − 27.86*	[−52.41, −3.32]	0.62	0.39	y = 1.90 x − 40.87	[−82.83, 1.09]	0.41	●	●	●	
35	0.40	y = 1.08 x + 17.42*	[4.32, 30.51]	0.40	0.47	y = 0.98 x − 1.38	[−18.15, 15.39]	0.46	0.61	y = 0.91 x − 17.42*	[−30.89, −3.96]	0.57	●	●	○	
36	0.24	y = 1.22 x − 26.72	[−1227.75, 1174.30]	0.01	0.49	y = 1.17 x − 23.79	[−98.26, 50.68]	0.37	0.26	y = 0.96 x + 2.41	[−75.03, 79.86]	0.10	●	●	○	
												Num ●:	36	32	19	1
												Num ○:	0	3	15	5
												Total:	36	35	34	6

Note: ρ is the Spearman's correlation coefficient. * indicates a statistically significant intercept. CI denotes a 95% confidence interval. r is the Pearson's correlation coefficient. N, O, I, and R stand for Nominal, Ordinal, Interval, and Ratio respectively. Circles indicate whether the three models for a given participant (ID) provided strong (●), weak (○), or no (a blank) evidence for the associated level of measurement. The $\rho = 0.10$ observed for participant 3 for the y - Ask and x - Slider condition rounded up to 0.10.

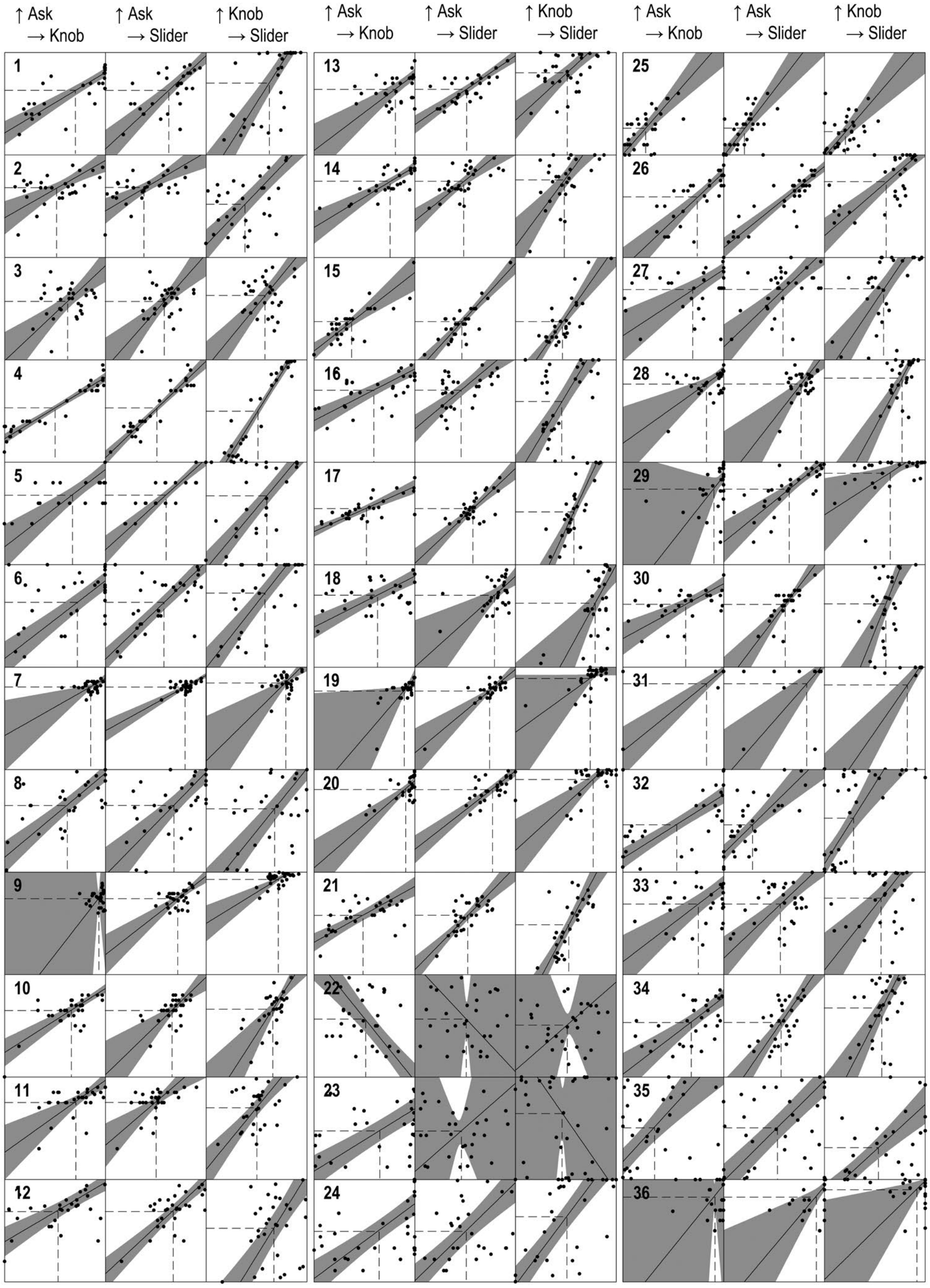


Figure 4. Plots showing the data collected from participants for each pair of the three pairs of judgment modalities. The three plots for a given participant (the number in top left of plots) are clustered horizontally. In each plot, pairs of judgment values for comparable trials are points, the fitted Deming regression lines are black, 95% confidence intervals are grey areas, and average values for each judgment modality are dotted lines. All plots go from 0 to 100 on both the x and y axes.

Table 6. Analysis of Aggregated Experimental Results.

						At Least			
Judgment Modalities		ρ	Model	Intercept CI	r	N	O	I	R
y - Ask,	x - Knob	0.61	y = 0.74 x + 15.08*	[12.50, 17.67]	0.62	•	•	•	
y - Ask,	x - Slider	0.61	y = 0.92 x + 9.14*	[6.35, 11.92]	0.63	•	•	•	
y - Knob,	x - Slider	0.58	y = 1.25 x − 8.08*	[−13.66, −2.50]	0.58	•	•	•	
					Overall:	•	•	•	

Note. See Table 5 for a description of this table's notations.

trust as if it is ratio (like those in Lee and Moray 1992) should be re-examined to see if they still hold with trust being interval.

Despite this caveat, the results are generally positive for the research community. Specifically, our results suggest that it is reasonable to treat trust at the interval level of measurement, which is the overriding standard in the research community. Thus, our results do not suggest that there is a serious problem related to level of measurement with the vast majority of the trust research. This point is further supported by the work of Jaccard, Wan, and Jaccard (1996), who showed that ordinal data generally can be treated as interval without a significant impact on statistical outcomes.

Our conclusion is further supported by the results of our aggregate analyses. This is because, when all of the data are considered together, all three judgment modality pairs exhibit strong evidence of intervality.

It is worth noting that one of our participants (participant 5; Figure 4) appeared to have set levels of values that he or she used throughout the experiment. While some might be tempted to interpret this as indicating an ordinal level, such levels are an issue of resolution rather than level of measurement. In fact, this participant ultimately exhibited strong evidence of treating trust as an interval level (Table 5). Such results can provide some validity to the practice of researchers using low resolution scales with only 5 or 7 levels (see for example Bass, Baumgart, and Shepley 2013) and analyzing them with interval level statistics.

We fully acknowledge that our study may not be universally generalisable to all situations where trust in automation is relevant given that we only tested one application domain and one task within that domain. As such, future work should seek to see if our results extend to other areas where trust in automation is important. However, even with this limitation, this work does show that our method can identify different levels of measurement and that different people can treat trust as being ordinal, interval, or ratio. The study also provides compelling evidence that the interval level that most researchers assume is valid. Thus, we think this research is significant.

Beyond these contributions, this research has a number of implications for future research. These are explored below.

7.1. Insights into the application of our method

Our trust experiment is the first to use our method for identifying the level of measurement of trust. For the most part, this experiment was successful. However, there are a few things that could be done differently to improve the application of the method. First, several of the participants exhibited very wide confidence intervals around the intercept. For some participants, this seemed to occur because they did not produce many ratings on the lower

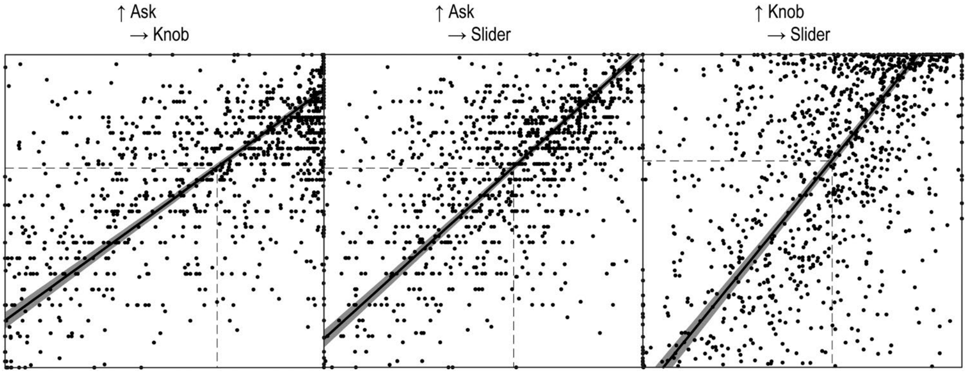


Figure 5. Plots showing the data collected from all participants or each pair of judgment modalities. Pairs of judgment values for comparable trials are points, the fitted Deming regression line is black, 95% confidence intervals are grey areas, and average values for each judgment modality are dotted lines. All plots go from 0 to 100 on both the x and y axes.

end of the scale (see, for example, participants 7, 9, 19, 20, 28, 29, and 36). This makes it difficult to determine if strong evidence of a ratio level was present. To address this, future work should work hard to ensure that the range of trials used in experiments will elicit a full range of participant responses, especially those near the lower end of the scale. Second, the 90 trials (30 for each of the three judgment modalities) required approximately 2 or more hours of participant time. This coupled with the tedious nature of the trials resulted in participants becoming noticeably bored. This may have impacted their ability to perform the experimental task with the rigour we desired. Thus, future work should work to either minimise the experiment's time or make trials more engaging. Third, the independent variable levels in experiments were selected to elicit a range of trust values from participants (something that was clearly accomplished in the results; Figures 4 and 5). These were not selected to give us insights into the impact the different factors have on trust ratings as would be possible with a full or partial factorial design. This work was not specifically interested in investigating how the considered factors impact trust. Furthermore, a full factorial design would either necessitate a prohibitively large number of trials or restrict the number of factors that could be used to influence trust. Thus, we chose not to use a factorial design. However, this choice does limit the insights we can obtain from analyzing the impact of the experimental factors. Future use of our method should employ more traditional experimental designs if insights into the effects of the independent variable levels is of importance.

7.2. Level of Measurement of Other Psychological Phenomenon

There are many psychometrics scales used in ergonomic research for measuring things like workload and situation awareness. None of these have had their level of measurement assessed. Thus, it is possible that there are problems with the ways these measures are treated in the literature. In particular, the Situation Awareness Rating Technique (SART, which is used for measuring situation awareness) and the NASA-TLX (which is used to measure mental workload) both combine multiple psychometric ratings that assess different phenomena into a single score by multiplying the other ratings by a scaling factor and adding them together. For SART, these are 'Demand on Attentional Resource,' 'Supply of Attentional Resource,' and 'Understanding of the Situation' (Selcon and Taylor 1990; Taylor 1989). For NASA-TLX, these are 'Mental Demand,' 'Physical Demand,' 'Temporal Demand,' 'Performance,' 'Effort,' and 'Frustration' (Hart and Staveland 1988). By adding these ratings together, the methods are assuming that they are at least interval. Further, since this process does not account for an intercept that would move arbitrary zeros between the different psychometric scales, the computation appears to assume that these values are ratio. Future work should adapt our method for use with these measures and assess the level of measurement for both their primary psychometric as well as the scales used in their computation.

7.3. Negative Trust and Distrust

Our results have potential implications for the measure of distrust. In particular, it is an open issue whether distrust is negative trust or an orthogonal measure (Jian, Bisantz, and Drury 2000). From a level-of-measurement perspective, distrust being negative trust would imply that trust is ratio because it implies a meaningful zero (a meaningful inflection point where trust (positive) transitions to distrust (negative)). In other words, distrust cannot be negative trust if trust is interval or ordinal because this would mean that negative trust would simply be lower than a positive value. Because our results suggest that trust is interval, this would indicate that distrust is not negative trust. This should be explored in more depth in the future.

7.4. Limitations of Stevens' Levels of Measurement

There are limitations with Stevens' (1946) levels of measurement (Velleman and Wilkinson 1993). The majority of these are based on potential deficiencies of Stevens' topology. In particular, there are numbers that are not properly accounted for by Stevens' level. For example, percentages constitute a scale that has a meaningful zero but does not support meaningful ratio transformations (Velleman and Wilkinson 1993). Because of such discrepancies, researchers (Mosteller and Tukey 1977; Velleman and Wilkinson 1993) have proposed alternative topologies of measurement. While these are not standard and rarely used, future work should investigate how our results and methods could address these other systems.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Air Force Research Lab/Universal Technology Corporation under Prime Contract FA8650-1.6-C-2642/Subcontract 18-S8401-13-C1.

Notes on contributors

Jiajun Wei received the M.S. degree in applied psychology (focusing on engineering psychology) from Zhejiang University, Hangzhou, China, in 2014. He received the Ph.D. in Industrial Engineering from the University at Buffalo, the State University of New York, in 2019. His research interests include human factor engineering, human–computer interaction, cognition, judgment, and decision making.

Matthew L. Bolton received the B.S. degree in computer science, the M.S. degree in systems engineering, and the Ph.D. degree in systems engineering from the University of Virginia, Charlottesville, VA, USA, in 2004, 2006, and 2010, respectively. He is an Associate Professor with the Department of Industrial and Systems Engineering at the University at Buffalo, the State University of New York. His research focuses on the use of human performance modeling and formal methods in the analysis, design, and evaluation of safety-critical systems.

Laura R. Humphrey received the B.S., M.S., and PhD. in electrical and computer engineering from the Ohio State University in 2004, 2006, and 2009, respectively. She is a Senior Research Engineer in the Autonomous Controls branch of the Aerospace Systems Directorate of the Air Force Research Laboratory. Her research focuses on formal methods for design and verification of autonomous and human-automation systems.

References

Abe, Genya, and John Richardson. 2006. "Alarm timing, trust and driver expectation for forward collision warning systems." *Applied ergonomics* 37 (5): 577–586. doi:10.1016/j.apergo.2005.11.001.

Annett, John. 2002. "Subjective rating scales: Science or art?" *Ergonomics* 45 (14): 966–987. doi:10.1080/00140130210166951.

Bagheri, Nasrine, Greg A. Jamieson. 2004. "Considering subjective trust and monitoring behavior in assessing automation-induced "complacency." *Human performance, situation Awareness, and Automation: Current Research and Trends*: 54–59.

Bailey, Nathan R., and Mark W. Scerbo. 2007. "Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust." *Theoretical Issues in Ergonomics Science* 8 (4): 321–348. doi:10.1080/14639220500535301.

Bagheri, Nasrine, and Greg A. Jamieson. 2004. "Considering subjective trust and monitoring behavior in assessing automation-induced 'complacency." In *Human performance, situation Awareness, and Automation: Current Research and Trends,* edited by Dennis A. Vincenzi, Mustapha Mouloua, Patter A. Hancock. 54-59. Fort Detrick, MD: US Army Medical Research and Material Command.

Barrett, Paul. 2003. "Beyond psychometrics: Measurement, non-quantitative structure, and applied numerics." *Journal of Managerial Psychology* 18 (5): 421–439. doi:10.1108/02683940310484026.

Bass, Ellen J., Leigh A. Baumgart, and Kathryn Klein Shepley. 2013. "The effect of information analysis automation display content on human judgment performance in noisy environments." *Journal of cognitive engineering and decision making* 7 (1): 49–65. doi:10.1177/1555343412453461.

Biros, David P., Mark Daly, and Gregg Gunsch. 2004. "The influence of task load and automation trust on deception detection." *Group Decision and Negotiation* 13 (2): 173–189. doi:10.1023/B:GRUP.0000021840.85686.57.

Bisantz, Ann M., and Younho Seong. 2001. "Assessment of operator trust in and utilization of automated decision-aids under different framing conditions." *International Journal of Industrial Ergonomics* 28 (2): 85–97. doi:10.1016/S0169-8141(01)00015-4.

Bolton, Matthew L. 2008. "Modeling human perception: Could Stevens' Power Law be an emergent feature?." In *IEEE International Conference on Systems, Man and Cybernetics*, 1073–1078. IEEE.

Clare, Andrew S., Mary L. Cummings, and Nelson P. Repenning. 2015. "Influencing trust for Human-Automation Collaborative Scheduling of Multiple Unmanned Vehicles." *Human factors* 57 (7): 1208–1218. doi:10.1177/0018720815587803.

Cliff, Norman, and JohnA. Keats. 2003. *Ordinal measurement in the behavioral sciences*. London: Psychology Press.

Cohen, Barry H. 2013. *Explaining psychological statistics*. Hoboken, NJ: John Wiley & Sons.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd. ed. New Jersey: Lawrence Erlbaum.

Cramer, Henriette S.M., Vanessa Evers, Maarten W. Van Someren, 2009. and, and BobJ. Wielinga. "Awareness, training and trust in interaction with adaptive spam filters." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 909–912. ACM. doi:10.1145/1518701.1518839.

Davenport, Randy B., and Ernesto A. Bustamante. 2010. "Effects of false-alarm vs. missprone automation and likelihood alarm technology on trust, reliance, and compliance in a miss-prone task." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1513–1517. SAGE Publications Sage CA: Los Angeles, CA. doi:10.1177/154193121005401933.

De Vries, Peter, and Cees Midden. 2008. "Effect of indirect information on system trust and control allocation." *Behaviour & information technology* 27 (1): 17–29. doi:10.1080/01449290600874956.

Deming, William Edwards. 1943. *Statistical adjustment of data*. Hoboken, NJ: Wiley.

Dzindolet, Mary T., Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. "The role of trust in automation reliance." *International journal of humancomputer studies* 58 (6): 697–718. doi:10.1016/S1071-5819(03)00038-7.

Eignor, Daniel R. 2013. *The standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

Ellermeier, Wolfgang, and Günther Faulhammer. 2000. "Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production." *Perception & psychophysics* 62 (8): 1505–1511. doi:10.3758/bf03212151.

Furr, RMichael, and VerneR. Bacharach. 2013. *Psychometrics: An introduction*. 2nd. ed. Los Angeles: Sage.

Ghiselli, Edwin Ernest, John Paul Campbell, and Sheldon Zedeck. 1981. *Measurement theory for the behavioral sciences*. New York: WH Freeman.

Guilford, JoyPaul. 1954. *Psychometric methods*. New York: McGraw-Hill.

Hart, Sandra G., and Lowell E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research." *Advances in psychology* 52: 139–183.

Hoff, Kevin Anthony, and Masooda Bashir. 2015. "Trust in automation: integrating empirical evidence on factors that influence trust." *Human factors* 57 (3): 407–434. doi:10.1177/0018720814547570.

Hoffman, Robert R., Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. 2013. "Trust in automation." *IEEE Intelligent Systems* 28 (1): 84–88. doi:10.1109/MIS.2013.24.

Jaccard, James, Choi K. Wan, and Jim Jaccard. 1996. *LISREL approaches to interaction effects in multiple regression*. New York: Sage.

Jian, Jiun-Yin, Ann M. Bisantz, and Colin G. Drury. 2000. "Foundations for an empirically determined scale of trust in automated systems." *International Journal of Cognitive Ergonomics* 4 (1): 53–71. doi:10.1207/S15327566IJCE0401_04.

Kazi, Tarannum, Neville A. Stanton, Guy H. Walker, and Mark S. Young. 2007. "Designer driving: Drivers' conceptual models and level of trust in adaptive cruise control." *International Journal of Vehicle Design* 45 (3): 339–360. doi:10.1504/IJVD.2007.014909.

Kazi, Tara A., Neville A. Stanton, Mark S. Young, and D. A. Harrison. 2005. "Assessing drivers' level of trust in Adaptive Cruise Control and their conceptual models of the system: implications for system design." *Driver behaviour and training* 2: 132–142.

Kline, P. 1986. *A Handbook of Test Construction: Introduction to Psychometric Design*. New York, NY: Methuen.

Laming, Donld Richard John. 1997. *The measurement of sensation*. Oxford: Oxford University.

Lee, John D., and Neville Moray. 1992. "Trust, control strategies and allocation of function in human-machine systems." *Ergonomics* 35 (10): 1243–1270. doi:10.1080/00140139208967392.

Lee, John D., and Katrina A. See. 2004. "Trust in automation: Designing for appropriate reliance." *Human factors* 46 (1): 50–80. doi:10.1518/hfes.46.1.50_30392.

Luce, R. Duncan. 1997. "Quantification and symmetry: Commentary on Michell, Quantitative science and the definition of measurement in psychology." *British Journal of Psychology* 88 (3): 395–398. doi:10.1111/j.2044-8295.1997.tb02645.x.

Ma, Ruiqi, and David B. Kaber. 2007. "Effects of in-vehicle navigation assistance and performance on driver trust and vehicle control." *International Journal of Industrial Ergonomics* 37 (8): 665–673. doi:10.1016/j.ergon.2007.04.005.

Madhavan, Poornima, Douglas A. Wiegmann, and Frank C. Lacson. 2006. "Automation failures on tasks easily performed by operators undermine trust in automated aids." *Human factors* 48 (2): 241–256. doi:10.1518/001872006777724408.

Manzey, Dietrich, Juliane Reichenbach, and Linda Onnasch. 2012. "Human performance consequences of automated decision aids: The impact of degree of automation and system experience." *Journal of Cognitive Engineering and Decision Making* 6 (1): 57–87. doi:10.1177/1555343411433844.

Messick, Samuel. 1995. "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning." *American psychologist* 50 (9): 741–749. doi:10.1037/0003-066X.50.9.741.

Michell, Joel. 1997. "Quantitative science and the definition of measurement in psychology." *British Journal of Psychology* 88 (3): 355–383. doi:10.1111/j.2044-8295.1997.tb02641.x.

Michell, Joel. 2008. "Is psychometrics pathological science?" *Measurement* 6 (1-2): 7–24.

Mosteller, Frederick, and JohnWilder Tukey. 1977. *Data analysis and regression: A second course in statistics*. Boston, MA: Addison-Wesley.

Muir, Bonnie M. 1987. "Trust between humans and machines, and the design of decision aids." *International Journal of Man-Machine Studies* 27 (5-6): 527–539. doi:10.1016/S0020-7373(87)80013-5.

Narens, Louis. 1996. "A theory of ratio magnitude estimation." *Journal of Mathematical Psychology* 40 (2): 109–129. doi:10.1006/jmps.1996.0011.

NCSS 2016. "Deming Regression." Chap. 303 in NCSS Statistical Software, 303-1–303-33. NCSS. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Deming_Regression.pdf.

Pak, Richard, Ericka Rovira, Anne Collins McLaughlin, and Natalee Baldwin. 2017. "Does the domain of technology impact user trust? Investigating trust in automation across different consumer-oriented domains in young adults, military, and older adults." *Theoretical issues in ergonomics science* 18 (3): 199–220. doi:10.1080/1463922X.2016.1175523.

Perkins, Lee Ann, Janet E. Miller, Ali Hashemi, and Gary Burns. 2010. "Designing for humancentered systems: Situational risk as a factor of trust in automation." In *Proceedings of the human factors and ergonomics society annual meeting*, 2130–2134. SAGE Publications Sage CA: Los Angeles, CA. doi:10.1177/154193121005402502.

Rasmussen, Steven, Derek Kingston, 2018. and, and Laura Humphrey. "A brief introduction to unmanned systems autonomy services (UxAS)." In *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*, 257–268. IEEE. doi:10.1109/ICUAS.2018.8453287.

Rovira, Ericka, Kathleen McGarry, and Raja Parasuraman. 2007. "Effects of imperfect automation on decision making in a simulated command and control task." *Human factors* 49 (1): 76–87. doi:10.1518/001872007779598082.

Rovira, Ericka, Richard Pak, and Anne McLaughlin. 2017. "Effects of individual differences in working memory on performance and trust with various degrees of automation." *Theoretical Issues in Ergonomics Science* 18 (6): 573–591. doi:10.1080/1463922X.2016.1252806.

Rovira, Ericka, and Raja Parasuraman. 2010. "Transitioning to future air traffic management: Effects of imperfect automation on controller attention and performance." *Human factors* 52 (3): 411–425. doi:10.1177/0018720810375692.

Selcon, S. J., and R. M. Taylor. 1990. "Evaluation of the Situational Awareness Rating Technique (SART) as a tool for aircrew systems design." In *AGARD, Situational Awareness in Aerospace Operations* 5-1–5-8.

Seong, Younho, and Ann M. Bisantz. 2008. "The impact of cognitive feedback on judgment performance and trust with decision aids." *International Journal of Industrial Ergonomics* 38 (7-8): 608–625. doi:10.1016/j.ergon.2008.01.007.

Stevens, Stanley Smith. 1946. “On the theory of scales of measurement.” *Science (New York, N.Y.)* 103 (2684): 677–680. doi:10.1126/science.103.2684.677.

Stevens, Stanley Smith. 1956. “The direct estimation of sensory magnitudes: Loudness.” *The American journal of psychology* 69 (1): 1–25. doi:10.2307/1418112.

Stevens, Stanley Smith. 1951. “Mathematics, measurement, and psychophysics.” In *Handbook of experimental psychology*, edited by Stanley Smith Stevens. Hoboken, NJ: Wiley.

Stevens, StanleySmith. 1975. *Psychophysics*. Piscataway, NJ: Transaction Publishers.

Taylor, R. M. 1989. “Situational awareness rating technique (SART): The development of a tool for aircrew systems design.” In *AGARD, Situational Awareness in Aerospace Operations*. Seuilly-sur Seine: NATO AGARD.

Trendler, Günter. 2009. “Measurement theory, psychology and the revolution that cannot happen.” *Theory & Psychology* 19 (5): 579–599. doi:10.1177/0959354309341926.

Velleman, Paul F., and Leland Wilkinson. 1993. “Nominal, ordinal, interval, and ratio typologies are misleading.” *The American Statistician* 47 (1): 65–72. doi:10.2307/2684788.

Visser, Ewart de., and Raja Parasuraman. 2011. “Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload.” *Journal of Cognitive Engineering and Decision Making* 5 (2): 209–231. doi:10.1177/1555343411410160.

Visser, Ewart. J. de., Frank Krueger, Patrick McKnight, Steven Scheid, Melissa Smith, Stephanie Chalk, and Raja Parasuraman. 2012. “The world is not enough: Trust in cognitive agents.” In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 263–267. Sage Publications Sage CA: Los Angeles, CA. doi:10.1177/1071181312561062.

Wang, Lu, Greg A. Jamieson, and Justin G. Hollands. 2009. “Trust and reliance on an automated combat identification system.” *Human factors* 51 (3): 281–291. doi:10.1177/0018720809338842.

Wei, Ting, and Scott Bell. 2012. “Impact of indoor location information reliability on users’ trust of an indoor positioning system.” In *International Conference on Geographic Information Science*, 258–269. Springer.

Wei. Jiajun, Bolton Matthew L., and Humphrey Laura. 2019. “Subjective measurement of trust: Is it on the level?.” In *Proceedings of the 2019 International Annual Meeting of the Human Factors and Ergonomics Society*, 5 pages. In Press. Santa Monica: Human Factors/Ergonomics Society.

Zimmer, Karin. 2005. “Examining the validity of numerical ratios in loudness fractionation.” *Perception & psychophysics* 67 (4): 569–579. doi:10.3758/bf03193515.