Subjective Measurement of Trust: Is It on the Level?

Jiajun Wei
University at Buffalo, SUNY

Matthew L. Bolton
University at Buffalo, SUNY

Laura Humphrey
Air Force Research Lab

Psychometrics are increasingly being used to evaluate trust in the automation of safety-critical systems. There is no consensus on what the highest level of measurement is for psychometric 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 psychometric ratings. We use this to assess the level of measurement of trust in automation using human ratings about the behavior 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 ratings are ratio. We discuss these results, their implications, and future research.

INTRODUCTION

With the rise of autonomous systems, researchers have become increasingly interested in designing automation that humans will trust. There are different ways of measuring trust in automation (Hoff & Bashir, 2015). Because trust is psychological, it is typically measured using psychometric rating scales. For these, humans use introspection to convert their psychological state into a number on a predetermined scale. Unfortunately, it is not clear what the level of measurement is for trust.

Psychometric scales have one of four levels of measurement: nominal (where numbers only indicate name); ordinal (where numbers only indicate order); interval (where the distances between numbers have meaning but there is no meaningful zero); or ratio (where ratios between numbers have meaning by virtue of there being a meaningful zero). A scale's level determines what transformations, mathematical comparisons, and statistical operations can be meaningfully employed on measures made on the scale (B. H. Cohen, 2013; Stevens, 1946). Nominal scales are compatible with counts, mode, and contingency correlation; ordinal scales support medians and percentiles; interval scales allow for the computation of means, standard deviations, rank-order and product moment correlations, and most parametric statistics; and ratio scales are compatible with percent changes, logarithms, geometric means, and coefficients of variation (Stevens, 1946). The levels of measurement are ordered from nominal (lowest), to ordinal, to interval, and to ratio (highest). Mathematical operations that can be performed meaningfully at lower levels can be meaningfully applied to higher ones. The reverse is not true. Thus, more powerful mathematics and statistics can be performed on scales at a higher level of measurement. For this reason, practitioners prefer to treat most psychometric ratings as being at least interval (Furr & Bacharach, 2013; Guilford, 1954).

This topic is controversial. Most psychometrics experts do not think psychometric scales should be treated as ratios (Furr & Bacharach, 2013; Guilford, 1954). Furthermore, many ergonomists and measurement theorists (Annett, 2002; Michell, 2008; Trendler, 2009) doubt that subjective ratings should be used as anything more than ordinal. Despite this, researchers have handled human trust in automation at levels ranging from ordinal to ratio (Lee & Moray, 1992; Muir, 1987). This is concerning because subjective trust ratings are being used in the design and execution of complex systems. If these measures are processed at a higher level than they should be, meaningless assessments are being made. In safety-critical applications, this could mean the difference between life and death.

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. There are techniques for eliciting interval-level measures for any given continuum (see McGrath et al. 1996). However, designing a scale to create data at a level does not necessarily imply that the phenomenon being measured is on that level.

We set out to fill this gap. We introduce a new method for determining what the maximum level of measurement is for psychometrics, and we use this method to assess trust in automation. Below, we provide background on trust measurement 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 preliminary results of this analysis and discuss their implications for the measurement and modeling of trust.

BACKGROUND

Measuring Trust in Automation

Trust in automation is subjectively measured on scales from 1 to n points, where n can range from 5 to 100 (Kessler, Larios, Walker, Yerdon, & Hancock, 2017). Across the literature, trust has been treated as being at ordinal, interval (evidenced by the prolific use of parametric statistics), and ratio (evidenced by analysts computing percent changes in trust values; Lee and Moray 1992) levels. The interval level is the most popular because it allows researchers to use parametric statistics to analyze data. Thus, there is no clear consensus about the level of measurement of trust.

Psychophysical Levels of Measurement

While we have been not been able to identify any specific studies that investigate the maximum level of measurement of 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 in physical units on a ratio scale. Prominent researchers have expressed skepticism that humans are capable of making true ratio judgments (Laming, 1997). Attempts have been made to check this (Ellermeier & Faulhammer, 2000; Zimmer, 2005) by assessing whether judged ratio differences between measured physical stimuli levels follow multiplicative and commu-

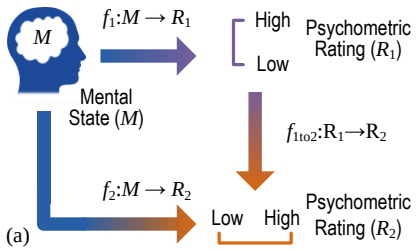

Figure 1. (a) Shows transformations between mental state M and scales R_1 and R_2. (b) and (c) show that if M, R_1, and R_2 are ratio (b) or interval (c), then f_{1to2} is ratio or interval respectively.

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 $a_2 > 0$

$\therefore m = f_1(m) / a_1 = f_2(m) / a_2$
$\therefore f_2(m) = a_2 / a_1 \cdot f_1(m)$
$\therefore f_{1to2}(r_1) = a_2 / a_1 \cdot r_1$

If we let $a_{1to2} = a_2 / a_1$
$f_{1to2}(r_1) = a_{1to2} \cdot r_1$ (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 $a_2 > 0$

$\therefore m = (f_1(m) - b_1) / a_1 = (f_2(m) - b_2) / a_2$
$\therefore f_2(m) = a_2 / a_1 \cdot f_1(m) + b_1 - a_1 \cdot b_1 / a_2$
$\therefore f_{1to2}(r_1) = a_2 / a_1 \cdot r_1 + (b_1 - a_1 \cdot b_1 / a_2)$

If we let $a_{1to2} = a_2 / a_1$ and $b_{1to2} = (b_1 - a_1 \cdot b_1 / a_2)$
$f_{1to2}(r_1) = a_{1to2} \cdot r_1 + b_{1to2}$ (c)

tative 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.

Transformations and Levels of Measurement

Scales that fall within a level of measurement can be converted to other scales at the same level through specific transformations. Let X and Y represent sets of numbers in two different scales at the same level and $f : X \to Y$ be a function that converts X to Y. If we assume X and Y are nominal, f can be any one-to-one function (where each element of X maps to exactly one element of Y, preserving the identity of each element). If X and Y are ordinal, f can be any strictly increasing function (thus preserving order). If X and Y are interval, f is a linear transformation $f(x) = a \cdot x + b$, where a (a positive scaling factor) and b (a repositioning of the relative zero) are constants. Finally, if X and Y are ratio, $f(x) = a \cdot x$, where a is a constant and $a > 0$.

These transformations determine whether mathematical operations are meaningful. For a comparison to be meaningful, it must hold when the numbers are permissibly transformed to different scales at the same level. See Table 1 for examples.

Table 1. *Meaningful and Meaningless Expressions Based on Transformations*

Expression	If X and Y are interval with $f(x)=a\,x+b$	If X and Y are ratio with $f(x)=a\,x$
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**	$f(x_1)-f(x_2)=k(f(x_3)-f(x_4))$ $\therefore (ax_1)-(ax_2)$ $=k((ax_3)-(ax_4))$ $\therefore x_1-x_2=k(x_3-x_4)$ $\therefore$ 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 $f:X \to Y$; 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.

OBJECTIVE

We developed a method for assessing the maximum level of measurement of psychometrics. This 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, 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 on these scales, the human will implicitly apply transformations $f_1 : M \to R_1$ and $f_2 : M \to R_2$ respectively. If M is best represented at a ratio level, R_1 measures can be converted to R_2 using a ratio transformation of the form $f_{1to2}(r_1) = a_{1to2} \cdot r_1$ (Figure 1(b)), where a_{1to2} is a constant. If M is interval, the conversion from R_1 to R_2 will be the interval transformation $f_{1to2}(r_1) = a_{1to2} \cdot r_1 + b_{1to2}$ (Figure 1(c)), where a_{1to2} and b_{1to2} are constants.

The forms of f_{1to2} give us an indirect means of determining the level of measurement most appropriate for measuring M. 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_{1to2} in Figure 1). Because both ratio and interval transformations are in a linear form, characterizing 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.

If the measure on one scale (R_1) is treated as the X variable and the comparable measure on the other (R_2) is the Y, a regression will result in a model $f_{1to2}(r_1) = a_{1to2} \cdot r_1 + b_{1to2}$. The statistics produced by this analysis will give us the means to identify the measurement level of M: If there is a 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 the regression model has a significant intercept (b_{1to2}), then M will be **interval**. If there is a strong linear relationship between R_1 and R_2 and the regression model has 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.

We used this approach to evaluate the level of measurement of trust using a human subjects experiment.

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.

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. 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.

Participants

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

Materials and Apparatus

The experiment was run in a controlled, quit, 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, & Humphrey, 2018). This enabled simulations to represent realistic UAS dynamics and route planning. The UAS was depicted as a blue chevron shape 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 center of the camera's view. The smaller the footprint, the more focused 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, (b) the position of a physical knob, or (c) the position of an on-screen slider.

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

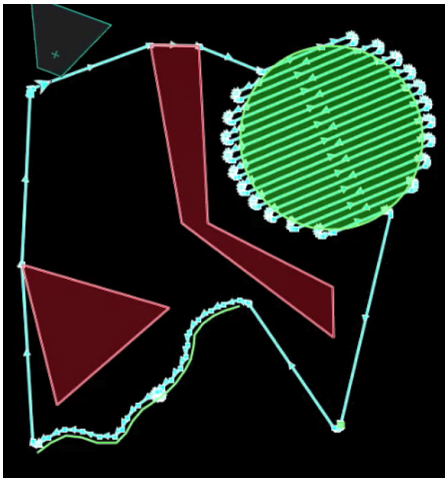

Figure 2. A screen of the UAS simulation.

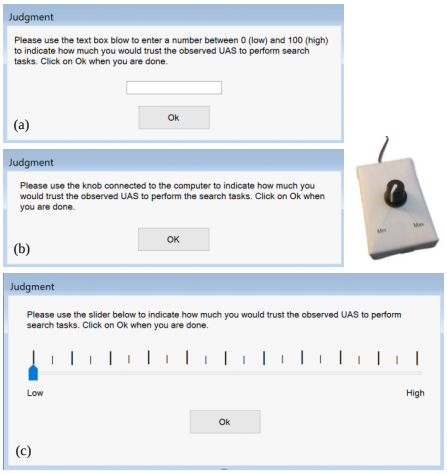

Figure 3. Software dialog boxes used for collecting human trust ratings. (a) Participants would enter a number between 0 and 100. (b) Participants would turn the physical knob connected to the computer. (c) Participants would use the computer's mouse to move a slider.

relate to *purpose*. The *Order*, *Density*, and *Path* relate to *process*. *Error*, *Skip*, and *NoFly* all relate to *performance*.

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.

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.

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 high-

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 the possible orders
Skip	The UAS could skip at most one task or part of the line search task	{None, Area, Point, Line, FirstLine, SecondLine}
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.

est 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.

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 synthesize 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.

Across all three models for each participant: (a) Strong evidence of nominality 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 models exhibited evidence of a ratio level and two or more exhibited strong evidence of this.

RESULTS

Due to page limits and an ongoing data analysis, we only present the results of the first six participants. Analysis results and the synthesis of all three modality comparisons for each participant are reported in Table 3 and Fig. 4. Analyses revealed that no participant exhibited strong evidence of a ratio level of measure for trust. Only participant 6 had weak evidence of a ratio level. Conversely, only participant 3 showed no evidence of an interval level. Five of the six participants showed evidence of an interval level and the evidence for all but one of these was strong. All the participants had evidence of an ordinal level of measurement, and all but one had strong evidence for this.

DISCUSSION AND CONCLUSIONS

This research is the first to identify the maximum level of measurement of the psychometrics of trust in automation. We are still analyzing all of our results. However, there is consistency in those we reported. First, because ordinal is the highest level that all the participants exhibited strong evidence for, and higher levels can always be accommodated by a lower level, it is safest to treat trust as ordinal. However, only one participant did not provide evidence of interval-level trust and only one of the remaining had evidence that was not strong. Thus, given the significant increase in mathematical power offered by the

Table 3. *Preliminary 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	●	●	●	○

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 are shorthand 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.

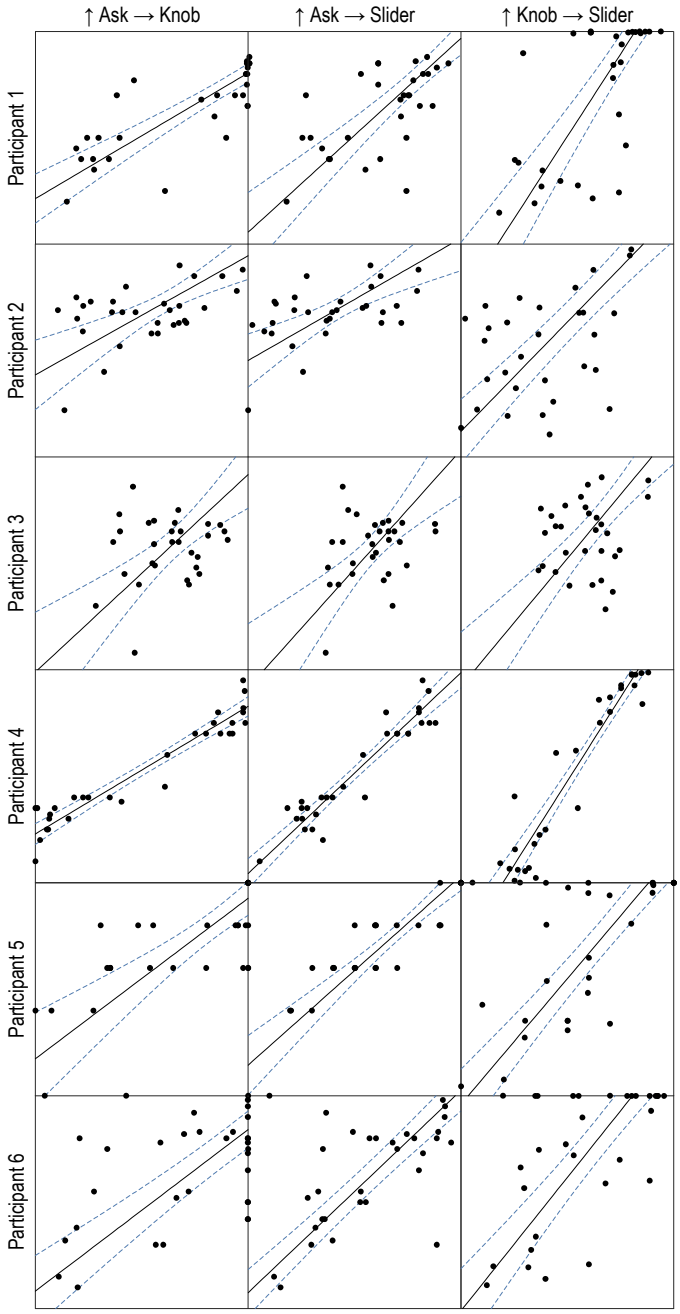

Figure 4. Plots showing the data collected from each participant (points), the fitted Deming regression line (black lines), and 95% confidence interval (blue dotted lines) for each participant for each pair of judgment modalities. All plots go from 0 to 100 on both the x and y axes.

interval level, our results indicate that this is a safe option. Conversely, only one participant had evidence of a ratio level, and this evidence was weak. This suggests that while some people may think about trust at a ratio level, it is not common.

If the full results are consistent with what is presented here, analysts should be extremely careful when handling subjective trust data. This is because some people are treating trust as if it is ordinal and/or ratio. Results that have processed trust as if it is ratio (Lee & Moray, 1992) should be reexamined to see if they still hold with trust being interval. These points will be more deeply explored after analyzing the complete data set.

There are limitations with Stevens' (1946) levels of measurement. For example, percentages constitute a scale that has a meaningful zero but does not support meaningful ratio transformations (Velleman & Wilkinson, 1993). Because of such discrepancies, researchers (Mosteller & Tukey, 1977) have proposed alternative topologies of measurement (though they are rarely used). Future work should investigate how our results and methods could address these other topologies.

Finally, 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. Future work should assess these concepts using our new method.

ACKNOWLEDGEMENT

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.

REFERENCES

Annett, J. (2002). Subjective rating scales: Science or art? *Ergonomics*, *45*(14), 966–987.

Bolton, M. L. (2008). Modeling human perception: Could Stevens' power law be an emergent feature? In *Ieee international conference on systems, man and cybernetics* (pp. 1073–1078).

Cohen, B. H. (2013). *Explaining psychological statistics*. John Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Deming, W. E. (1943). *Statistical adjustment of data*. Wiley.

Ellermeier, W., & Faulhammer, G. (2000). Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Attention, Perception, & Psychophysics*, *62*(8), 1505–1511.

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: An introduction* (2nd ed.). Los Angeles: Sage.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

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

Kessler, T. T., Larios, C., Walker, T., Yerdon, V., & Hancock, P. (2017). A comparison of trust measures in human–robot interaction scenarios. In *Advances in human factors in robots and unmanned systems* (pp. 353–364). Springer.

Laming, D. R. J. (1997). *The measurement of sensation*. Oxford University.

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

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

McGrath, P. A., Seifert, C. E., Speechley, K. N., Booth, J. C., Stitt, L., & Gibson, M. C. (1996). A new analogue scale for assessing children's pain: an initial validation study. *Pain*, *64*(3), 435–443.

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

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: a second course in statistics*. Addison-Wesley.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5-6), 527–539.

Narens, L. (1996). A theory of ratio magnitude estimation. *Journal of Mathematical Psychology*, *40*(2), 109–129.

NCSS. (2016). Deming regression. In *NCSS statistical software* (pp. 303-1–303-33). NCSS, LLC.

Rasmussen, S., Kingston, D., & Humphrey, L. (2018). A brief introduction to unmanned systems autonomy services (UxAS). In *2018 international conference on unmanned aircraft systems (icuas)* (pp. 257–268).

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Stevens, S. S. (1956). The direct estimation of sensory magnitudes: Loudness. *The American journal of psychology*, *69*(1), 1–25.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, *19*(5), 579–599.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, *47*(1), 65–72.

Zimmer, K. (2005). Examining the validity of numerical ratios in loudness fractionation. *Perception & Psychophysics*, *67*(4), 569–579.