The Level of Measurement of Subjective Situation Awareness and Its Dimensions in the Situation Awareness Rating Technique (SART)

Matthew L. Bolton, *Senior Member, IEEE*, Elliot Biltekoff, and Laura Humphrey, *Member, IEEE*

***Abstract*—Situation awareness (SA), a measure of how well a person understands the situation, is frequently used to evaluate the safety and effectiveness of critical systems that depend on human behavior. While there are objective ways of measuring SA, subjective assessments, such as the SA rating technique (SART), are still widely used. However, it is not clear what the level of measurement is for SART-measured SA or its constituent dimensions This is a significant gap because the level of measurement determines what mathematics and statistics can be meaningfully used to synthesize and evaluate measures. This research uses a previously developed method for determining the level of measurement of psychometric ratings to evaluate the level of measurement of SART and its elements. Results show that all of the dimensions of SA can be treated as interval in most situations, but that each is on a separate interval scale. This result casts doubt on the validity of the formula SART uses to compute SA from its subcomponents. We ultimately discuss our results and explore future research directions.**

***Index Terms*—Human performance assessment, psychometrics and testing, situation awareness (SA).**

I. Introduction

SITUATION awareness (SA) is a psychological concept that represents how well a person understands what is currently going on. More formally, SA is "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" [1]. For systems that rely on human behavior for correct and efficient operation, SA has become a critical concept because a human who does not understand the current situation will likely do things that are detrimental to safety and performance. In scientific analyses, measurement of SA is typically performed with objective measures, such as those offered by SA global assessment technique (SAGAT) [2]. However, the challenges of developing and administering SAGAT evaluations (identifying appropriate objective questions and pausing system operations to administer tests) can make it inconvenient in some situations. For this reason, subjective assessments of SA, the most popular of which is the SA rating technique (SART) [3], [4], are still regularly used [5]–[12].

When using subjective ratings in scientific research, the scales must be "reliable" (produce consistent results) and "valid" (correlated with things associated with what is being measured) [13]. They must also be handled with respect to the levels of measurement [14]. This last consideration is important as it determines what types of statistics and mathematics can be meaningfully applied to measurements. Until recently, there was no way of assessing the level of measurement people use for subjective scales. This deficiency was addressed by Wei *et al.* [15], who used a new method to assess the level of measurement of trust in automation.

In this research, we used the method introduced in [15] and [16] to assess the level of measurement of subjective SA using the dimensions of SART .

II. Background

A. Level of Measurement

The level of measurement of a scale determines what a number means in relation to other numbers measured on the same scale. Psychological measurement usually uses the four levels originally identified by Stevens [14]. Nominal numbers indicate category or identity (e.g., player number on a team). Ordinal numbers only indicate order (e.g., class rank). Interval numbers (e.g., temperature in Fahrenheit or Celsius) are the first where the distances between numbers have meaning. However, interval numbers have a nonmeaningful zero (a zero that does not constitute the absence of the measured quantity) and, thus, ratios between numbers are meaningless. Finally, ratio numbers (e.g., distance) have a meaningful zero and, thus, ratios between numbers have meaning.

The level of a scale determines what mathematical (and statistical) operations can be meaningfully applied to values measured on a scale [14]. Equalities, inequalities, counts, modes, set membership, and contingency correlation can be meaningfully computed on numbers from nominal scales. Comparisons of

Manuscript received 25 April 2021; revised 2 July 2021; accepted 19 September 2021. Date of publication 15 November 2021; date of current version 15 November 2022. This work was supported by the Air Force Research Lab /Universal Technology Corporation / ARCTOS Technology Solutions under Prime Contract FA8650-1.6-C-2642 / Subcontract18-S8401-13-C1. This article was recommended by Associate Editor Ming Hou. *(Corresponding author: Matthew L. Bolton.)*

Matthew L. Bolton and Elliot Biltekoff are with the Department of Industrial and System Engineering, University at Buffalo, The State University of New York, Amherst, NY 14260 USA (e-mail: mbolton@buffalo.edu; elbiltek@buffalo.edu).

Laura Humphrey is with the Air Force Research Laboratory, Wright-Patterson Air Force Base, OH 45433 USA (e-mail: laura.humphrey@us.af.mil).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/THMS.2021.3121960.

Digital Object Identifier 10.1109/THMS.2021.3121960

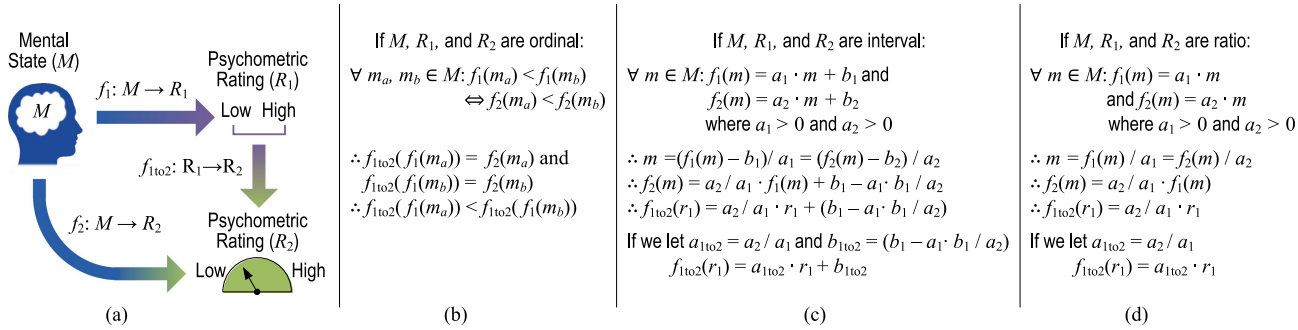


Fig. 1. Illustration of the relationships exploited by the method for assessing the level of measurement of psychological concepts (adapted from Wei *et al.* [15]). (a) Transformations between mental state M and scales R_1 and R_2. (b)–(d) If M, R_1, and R_2 are (b) ordinal, (c) interval, or (d) ratio, then $f_{1\text{to}2}$ is ordinal, interval, or ratio, respectively. In all of these, a_1, a_2, b_1, and b_2 are constants.

greater-than or less-than, percentiles, medians, rank-order statistics, and all nominal operations can be meaningfully calculated for numbers on ordinal scales. Numbers from interval scales can be used to meaningfully compute all nominal quantities, means, standard deviations, product moment correlations, and most parametric statistics. Finally, percent changes, geometric means, coefficients of variation, the full range of parametric statistics, and operations that can be performed on all other levels can be meaningfully calculated for numbers from ratio scales. Clearly, it is extremely advantageous to treat measures at the highest possible level (with ratio at the top and nominal at the bottom) because it enables many more possibilities for meaningful analyses.

Another important concept from level of measurement, which is used by the method for assessing the level of psychological qualities measured with subjective scales [15], is permissible transformation. A permissible transformation describes how numbers on one scale can be converted to another scale that is on the same level of measurement, so that mathematical power is retained. Permissible transformations on nominal scales are any one-to-one transformation: any transformation that preserves identity. For ordinal scales, permissible transformations are any that are strictly increasing: any function that preserves element order. An interval scale can be permissibly transformed into any other interval scale via a linear transformation $f_{\text{interval}}(X) = a \cdot X + b$: that is, a function that scales the original number by a positive constant a and repositions the zero with b. Finally, permissible transformations on ratio scales are ratio transformations $f_{\text{ratio}}(X) = a \cdot X$, where the original number is only scaled by a positive constant factor.

B. Method for Assessing a Scale's Level of Measurement

The method introduced in [15] and [16] for assessing the level of measurement of psychometrics uses meaningful transformations as its basis. Fig. 1 illustrates the concept behind the method. In this, assume that there are two psychometric scales R_1 and R_2 that both measure the same psychological quality M [see Fig. 1(a)]. When a human is asked to transform a specific value or state of M onto the two scales, he or she will implicitly apply transformations $f_1 : M \rightarrow R_1$ and $f_2 : M \rightarrow R_2$. As Fig. 1(b)–(d) shows, as long as the fidelity of R_1 and R_2 are sufficient to capture the level of M, M's level of measurement will determine the form of f_1 and f_2 based on the permissible transformations of the level. This, in turn, determines the form of the transformation for converting values on R_1 to values on R_2 ($f_{1\text{to}2}$), which will assume a form of the permissible transformation as dictated by M's level.

The forms of $f_{1\text{to}2}$ provide indirect means of determining M's level of measurement. By collecting ratings of M for the same conditions on two scales (R_1 and R_2), the level of measurement is revealed by the transformation that converts between them ($f_{1\text{to}2}$). In the assessment method [15], it is assumed that, as long as observations on R_1 and R_2 are distinct, there is enough evidence to assume nominality. A nonparametric (Spearman's ρ) is used to assess the strength of the ordinal relationship. Because both ratio and interval transformations are linear, linear regression can determine if there is evidence of interval or ratio relationships. Because error is possible in measures on both R_1 and R_2, the method prescribes Deming regression [17]. The form of the regression model, based on whether 0 is in the confidence interval around the intercept, shows whether the relationship indicates intervality (0 not in the interval) or ratio (0 in the interval). Because Deming regression does not use ordinary least squares for fitting, R^2 is not computed. Thus, Pearson's correlation coefficient (r) is used to assess the linear relationship ("fit") between the measures (the standard for Deming regression).

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, the chance of incorrect conclusions is reduced. Thus, Wei *et al.* [15] suggest the use of three scales with the heuristics in Table I for assessing the strength of evidence for each level.

C. Measuring SA With SART

While there have been multiple versions of SART (including its original formulation with ten different dimensions [4]), the most common form [3] (3-D SART) measures SA by having humans subjectively assess three dimensions based on their experience: *Demand* on attentional resources, *supply* of attentional

TABLE I
HEURISTIC FOR ASSESSING THE LEVEL OF MEASUREMENT FOR A GIVEN PARTICIPANT'S SUBJECTIVE RESPONSES

Level	Evidence Strength: Weak ○	Evidence Strength: Strong ●
Single Model		
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 $\lvert CI \rvert \geq 20$
Across All Three Models		
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

Note: ρ and r are Spearman's and Pearson's correlation coefficients, respectively. Standard methods [18] are used to assess coefficient strength. *CI* is a 95% confidence interval around the intercept of the linear Deming regression model.

resources, and *understanding* of the situation. Overall SA is then computed as

$$\text{SA} = \text{Understanding} - \text{Demand} + \text{Supply}. \quad (1)$$

Like with NASA Task Load Index (NASA-TLX) [19], 3-D SART dimensions are traditionally measured by having human participants mark a position on a 100-mm horizontal line. The rating's value is then measured as the mark's distance from the left edge of the line in millimeter. Conventional approaches use 100 point sliders on electronic displays. Note that in 3-D SART's original evaluation, overall SA was also assessed on such a scale [3]. Thus, while SART does not make explicit claims about the level of measurement of SA or its dimensions, the linear combination in (1) suggests that SA and the dimensions are at least interval. Furthermore, the relationship dictated by (1) suggests that all three SART dimensions are on the same level and scale. Finally, because the majority of research that uses SART analyzes results with parametric statistics (e.g., t-tests and analyses of variance) [5]–[12], there is an implicit assumption that SART measures are interval.

III. OBJECTIVES

With SART being used in the evaluation of safety-critical systems, such as aircraft [6], automobiles [8], and offshore oil drilling platforms [7], it is critical that the level of measurement of SA and its dimensions be well understood. This is because it allows analysts to use the mathematics and statistics that give them the most meaningful operations: avoiding meaningless conclusions and not sacrificing statistical or mathematical power. This research specifically sought to assess the level of measurement of SA as evaluated by SART and each of its three subjectively assessed dimensions. To accomplish this, we conducted an experiment where humans made assessments based on an unmanned aerial system (UAS) performing search tasks and analyzed each dimension of SART as well as overall SA [both subjectively and computed using (1)] using the method developed by Wei *et al.* [15]. We also analyzed our results with respect to the observed levels of measurement to ensure that (1) was consistent with them.

IV. METHODS

This study received approval from the University at Buffalo IRB under STUDY00002118.

A. Procedure

This experiment followed a similar procedure to the one originally used in [15]. Participants arrived at the laboratory and signed an informed consent document. They then observed a PowerPoint presentation that introduced them to the experimental task. Afterward, they performed the experiment, which involved watching simulations of UASs performing search tasks. The same simulations were observed in three blocks, where the participants rated their SA during the simulation using three different judgment methods.

B. Participants

We recruited 36 University at Buffalo student participants to the study. Thirteen of these were female and 23 were male.

C. Materials and Apparatus

The experiment was run in a controlled, quiet, evenly lighted laboratory. It was administered on three PCs, each resting on separate computer desks. Each computer was equipped with a 21-in LCD monitor, optical mouse, and keyboard. The experiment was administered on the computers using software that was specifically created for this project.

During the experiment, the software depicted a video of a UAS flying around a given area and performing search tasks (see Fig. 2). The simulations were created using OpenUxAS and OpenAMASE [20]. The UAS was depicted as a 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.

In simulations, the UAS performed line (the green line in Fig. 2) and point (green squares labeled with numbers in Fig. 2) search tasks. When all tasks were completed, the UAS flew to an end point and loitered there. When the UAS's planned flight path was shown (as in Fig. 2), it was depicted as a blue line. In some trials, the UAS would communicate its search intentions. When this was done (as in Fig. 2), it was shown in green text above the simulation.

During the simulation, participants were required to indicate which of the six displayed and labeled points were searched (using the checkboxes in the upper right of the display in Fig. 2). The UAS could explicitly search points on its flight path (such as points 1, 4, and 5 in Fig. 2) or points that fell into the camera's footprint during a line search (such as point 2 in Fig. 2).

After each simulation, participants provided ratings about their SA during simulations using 3-D SART [3], including overall SA, using (see Fig. 3): (a) a number between 0 and 100,

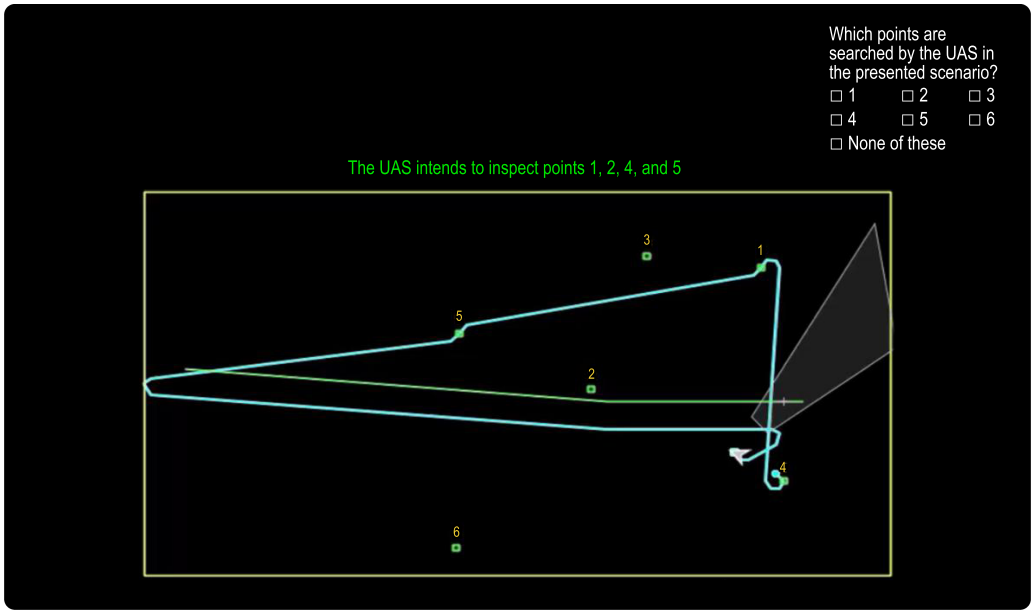


Fig. 2. Screenshot of the UAS simulation used in the experiment.

TABLE II
INDEPENDENT VARIABLES

Variable	Description	Levels
Intention	Whether the UAS's intention (the green text that says which points will be searched) is displayed.	Visible, Invisible
Path	Whether the flight path is displayed.	Visible, Invisible
LineSearch	Whether the line search's green line is displayed.	Visible, Invisible
NumPoints	The number of points searched by the UAS.	[0, 6]
Points	The points searched by the UAS.	Random
Radius	The size of the radius around the UAS.	Small, Medium, Large, Infinite

(b) the position of an on-screen knob (controlled using the mouse scroll wheel), or (c) the position of an on-screen slider.

D. Independent Variables

The independent variables are all related to the experimental trials, where trial geometry included the factors in Table II.

These factors were selected because their variation should produce a range of SA (and subdimension) responses. Consistent with SART's dimensions [3], they either supply attentional resources to humans by giving them guidance (displaying intention, the flight path, or the line search), place additional demands on attention (not showing intention, the flight path, or the line search; and using smaller radii), and/or reduce human understanding (not showing intention, the flight path, or the line search; and using smaller radii). These factors are also related to Endsley's three levels of SA [1]. That is, they impact the ability of the person to identify relevant objects in the environment (path visibility, line search visibility, and radius), comprehend their meaning (path visibility, line search visibility, and intention), and project that meaning into the future (path visibility, line search visibility, intention, and radius).

E. Dependent Measures

The dependent measures were the dimensions of 3-D SART made using each of the three judgment modalities (see Fig. 3): *Demand* represented how much demand was placed on human attention during a simulation; *Supply* represented how much spare attention and mental ability was available to the human during the simulation; and *Understanding* represented how well the human understood the situation during simulations. Because overall SART SA can be directly subjectively measured [3], this (SA_{Rated}) was also collected concurrently with the other subjective dimensions. Beyond this, we computed overall SA using the first three dimensions in accordance with (1)

$$SA_{Computed} = \text{Understanding} - \text{Demand} + \text{Supply}. \quad (2)$$

All five of these measures were collected for each simulation after it was shown to participants using one of the three judgment modalities: ask, knob, and slider. For the ask judgment modality [see Fig. 3(a)], dimensions were measured as a floating point number between 0 and 100 based on the values entered by people in text boxes. With the knob [see Fig. 3(b)], dimensions were measured as a floating-point number from 0 to 100 (with precision down to 0.02) based on the position of a knob between its minimum (0°) and maximum (300°) positions. With the slider modality [see Fig. 3(c)], each dimension was measured as a floating-point number from 0 to 100 (with precision down to 0.05) based on the left-to-right position of a slider.

Finally, for every trial participants viewed, they indicated which points were searched by the UAS. The identified points along with the correct answers for each trial were recorded.

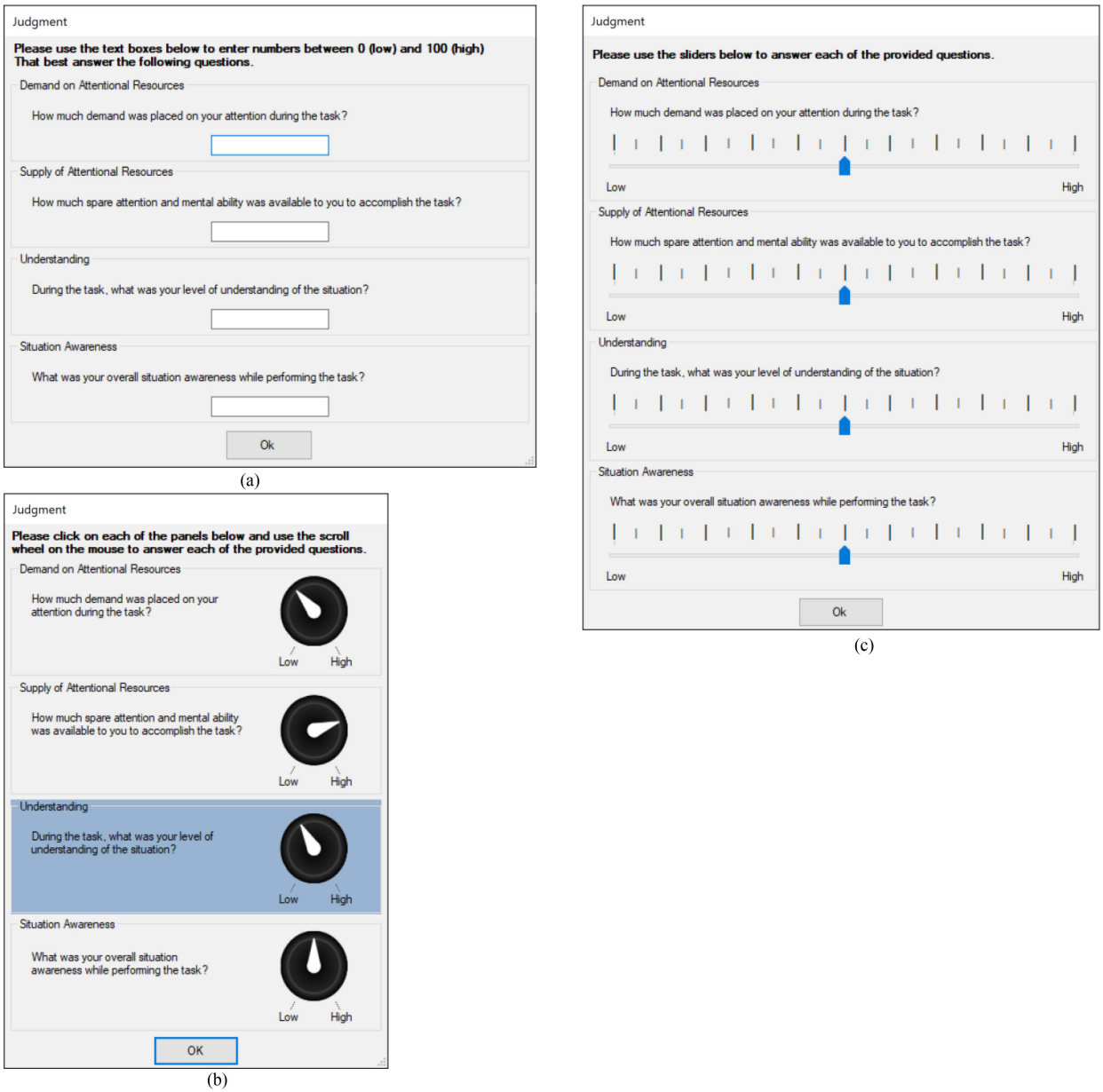


Fig. 3. Software dialog boxes that were used for collecting human SA ratings. (a) Participants enter numbers between 0 and 100. (b) Participants use the computer's mouse scroll wheel to turn an on-screen knob. (c) Participants use the computer's mouse to move a slider.

F. Experimental Design

We created 32 trials: one for each possible combination of the levels of *Intention*, *Path*, *LineSearch*, and *Radius* ($2 \cdot 2 \cdot 2 \cdot 4 = 32$). Within these, *NumPoints* and searched *Points* were assigned randomly. Four additional training trials were also created that exhibited variation along all independent variable dimensions.

When the experiment was run, each participant was assigned three unique random orders of the 32 experimental trials, one for each of the three judgment modalities. Trials for a given modality were presented in blocks. Block order, and thus judgment modality, was counterbalanced between participants.

Training trials were presented in a consistent order. At the beginning of the experiment, participants saw all four training trials to introduce them to the experimental task and first judgment modality. Subsequent training blocks of two trials were presented between judgment modalities to introduce participants to a new modality. Training trial and presentation orders were consistent between participants regardless of the given judgment modality order.

G. Data Analysis

Data analysis in this experiment followed the same general process established in [15]. However, in this experiment, the method was applied to the four measures collected from human participants (*Demand*, *Supply*, *Understand*, and SA_{Rated}). As was done in [15], the method was applied for each individual

participant as well as across all participants, with the heuristics from Table I being used to determine if weak or strong evidence was observed for each level of measurement.

Furthermore, because of the relationship predicted by (2), we intended to determine whether this is appropriate for the level of measurement for each of the components. We accomplished this in three ways.

First, we calculated $SA_{Computed}$ using (2) for each set of collected ratings. We then determined its level of measurement using the same technique used for the other measures. We compared these results with those found for SA_{Rated} to see if they were consistent.

Second, (2) suggests that *Understanding*, *Demand*, and *Supply* are on the same scale given that they can be combined together without any sort of transformation. If this were true, then we would expect the Deming regression models for converting between judgment modality pairs to be the same for all four measures. That is, the model for converting a participants *Demand* from an ask judgment to a knob judgment should be the same as converting *Supply*, *Understanding*, and SA_{Rated} between the same judgment type pairs. To test this, we used repeated measure analyses of variance. In these, the slopes and intercepts from the Deming regression models were the response variables. The measure type was the independent factor and the combination of the participant and judgment modality pair were the "subject" factor.

Third, if $SA_{Computed}$ were capturing SA, we would expect to be able to scale SA_{Rated} to the range of $SA_{Computed}$ and fit a regression that produces (2). That is, a regression model with regression coefficients that are effectively 1 and an intercept of 0. Thus, we performed this scaling procedure and regression fitting to perform this comparison.

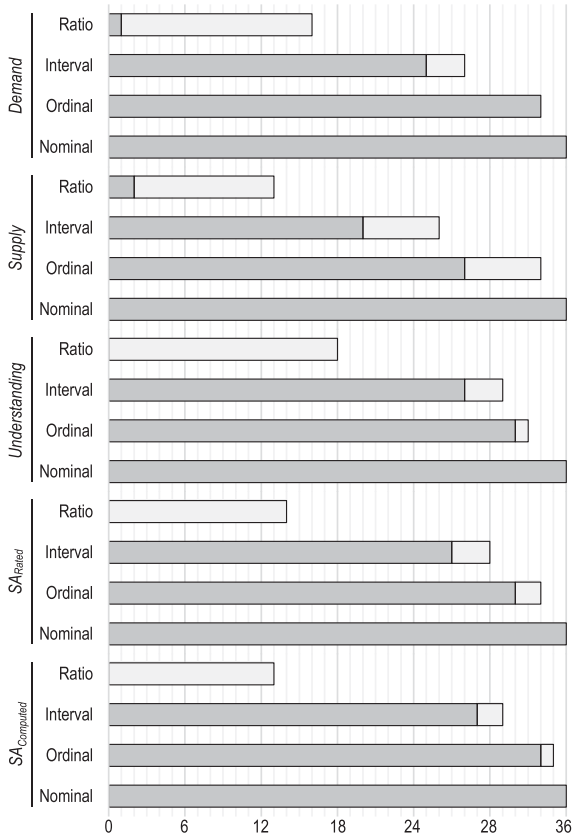


Fig. 4. Stacked bar chart showing the number of participants (out of a maximum of 36) that showed evidence of the different levels of measurement for each dependent measure and $SA_{Computed}$. Dark gray indicates the number of participants with strong evidence. Light gray indicates the number with weak evidence.

V. Results

A full listing of the level of measurement results using the same conventions as the previous trust experiment [15] can be found in this article's supplementary materials. For the sake of brevity, we report the overall results here. Specifically, Fig. 4 shows the total number of individual participants that exhibited weak and strong evidence that each of the evaluated measures was at a given level of measurement. When all of the participants were considered together (in aggregate), strong evidence was observed for nominal, ordinal, and interval levels for all of the measures, but no evidence for ratio.

Based on these results, SA_{Rated} and $SA_{Computed}$ produced similar results, though minor differences were seen in the number of participants that exhibited evidence for each level (see Fig. 4).

The repeated measures ANOVA that checked whether the slopes and intercepts were the same for converting between the measures collected from participants (*Demand*, *Supply*, *Understanding*, and SA_{Rated}) showed that there were no significant differences in regression model intercepts ($F_{2,\,296} = 1.78$, $p = 0.171$) between measures. However, there was a significant difference in model slopes ($F_{2,\,296} = 6.49$, $p = 0.002$).

Finally, SA_{Rated} was linearly rescaled to fit within the range of $SA_{Computed}$, creating SA_{Scaled}. When *Demand*, *Supply*, and *Understanding* were used as independent variables to fit a regression model to SA_{Scaled}, the following was obtained:[1]

$$\begin{aligned} SA_{Scaled} = \; & 2.6120.\text{Understanding} \\ & +0.1412.\text{Demand} \\ & +0.3934.\text{Supply} \\ & -102.51. \end{aligned} \tag{3}$$

This result indicated that the three predictors explained 79% of the variance ($R^2 = .79$, $F_{3,\,3452} = 4327.37$, $p < 0.00$), with all predictors and the intercept being significant ($p < 0.01$).

VI. Discussion

This work constitutes the first we know of to assess the level of measurement of SA. Below we discuss the significance of our findings and suggest avenues of future research.

A. Level of Measurement of SART and Its Dimensions

The results across all of our measured dimensions and across participants (both individually and in aggregate) are fairly consistent. Specifically, significant evidence exists to show that it

[1]Note that because this model required multiple independent variables, Deming regression could not be used.

is safe to treat *Demand*, *Supply*, *Understanding*, SA_{Rated}, and SA_{Computed} at both the ordinal and interval levels of measurement. Little evidence exists that any of these measures can be treated as ratio. This is largely an encouraging finding for the human factors' research given that it creates no conflict when using parametric statistics to analyze SART results. However, this also means that analysts need to be careful not to perform statistical or mathematical operations on SART data that assume a ratio level.

The results also show that there are clear individual differences between participants. This is demonstrated by some participants not being capable of providing evidence that they were thinking about the measures as being interval or even ordinal (e.g., see participants 17, 19, and 21 in the supplement). Thus, individual results should be handled carefully so as not to make assumptions about the level of measurement.

B. Validity of SART

Even with the encouraging results about SA's level of measurement and the similar measures observed for SA_{Rated} and $\text{SA}_{\text{Computed}}$, our results provide evidence that SART may not be an accurate measure of SA. First, the fact that we observed significant differences in model slopes for converting between judgment type pairs between *Demand*, *Supply*, *Understanding*, and SA_{Rated} suggests that these measures are not on the same scale. Second, when we scaled SA_{Rated} to $\text{SA}_{\text{Computed}}$ and fitted a regression model to it, the equation we obtained [see (3)] was clearly different from the original SART formulation [see (2)]. This further suggests that *Demand*, *Supply*, and *Understanding* are on separate interval scales that must be linearly transformed to the scale of SA_{Rated}. All of this provides evidence that the standard SART formulation [see (1)] is not supported by the levels of measurement and scales we observed.

It is important to note that, based on our results, we are not advocating for reformulating SART around (3): this formula is derived from only this one limited study. Rather, we feel like our results contribute to the growing consensus that there are problems with SART [21], [22]: that it diverges from objective performance (this topic is explored in the next section) and that it confounds with human confidence and mental workload. In fact, our results provide additional evidence for the workload confound in that they suggest that *Supply* and *Demand* are not on the same scale as SA_{Rated}. Furthermore, our results highlight a potential issue of considering elements of mental workload in SA. This is seen in the discrepancy in the sign of the *Demand* coefficient in our fitted model [see (3)] from the original one [see (1)]. As is well known in the human factors' community, workload and/or demand on resource can both improve and degrade performance [23]. Thus, even if mental workload considerations should be accounted for in SA (a position we are not advocating), the negative relationship enumerated in (1) is likely too simplistic.

C. SART Correspondence With Object Performance

While not directly tied to project objectives, the aforementioned results made us curious about how SA scores correlated with human objective performance during target identification. To investigate this post hoc consideration, we computed the total number of errors (*TargetErrors*) made by a participant for each trial based on which points were actually searched in the trial. Errors were then counted for each trial based on whether an unsearched point was selected or if a searched point was not. We then compared *TargetErrors* to both SA_{Rated} and $\text{SA}_{\text{Computed}}$ across all participants and trials with Spearman's ρ correlation. Ultimately, *TargetErrors* was significantly ($p < 0.001$) negatively correlated with both SA_{Rated} and $\text{SA}_{\text{Computed}}$ with $\rho = -0.288$ and $\rho = -0.327$, respectively, suggesting medium strength relationships between the measures [18]. Thus, although there are many reasons to be suspicious of SART's computed value of SA, it does somewhat correlate with objective measures of performance.

Because this experiment was concerned with subject SA, this analysis should not be regarded as a comprehensive comparison of SART to objective scales, such as SAGAT. We refer readers to the literature Endsley *et al.* [21] and Endsley [22] for such evaluations.

D. Additional Experiments

While our results are compelling, it is important to note that they come from just one experiment. Future work should explore whether the results can be replicated under similar conditions, with other populations (possibly those with more domain expertise), and other application domains.

E. Other Subjective Scales

As Endsley [22] notes, there are other subjective measures of SA. Our results suggest that humans generally treat these as if they are interval. This should be the subject of future research. Beyond this, there are subjective measures of other phenomena that are used to make design decisions about safety-critical systems. For example, the NASA-TLX [19] is widely used to assess mental workload. Future research should investigate the level of measurement of additional subjective measures to ensure they are being handled with mathematical meaning.

F. Level of Measurement of Subjective Dimensions

This study and the trust analysis reported in [15] and [16] are the first to determine the level of measurement of psychological concepts assessed using psychometrics. Given that trust and all of the SART dimensions showed good evidence of intervality, it may be that humans inherently think about psychological continua that are not conceptually constrained at a lower level (they do not represent inherent ordinal or nominal qualities) as being interval. This may be why, for example, Jaccard *et al.* [24] found that Likert scale data (which is generally thought to be ordinal) can usually be analyzed accurately with parametric statistics. In any case, the results presented here and in [15] and [16] provide additional evidence that few people are capable of thinking about concepts effectively on ratio scales. Similar conclusions have been reached in psychophysics research [25],

[26], where humans are actually making judgments about physical, ratio quantities. Future work should continue to explore the level of measurement of different psychological qualities to see if these findings holds or if there are specific features of the quality being measured that produces variations in level.

VII. Conclusion

In this research, we found evidence that subjectively assessed SA, demand of attentional resources, supply of attentional resources, and understanding of the situation exhibit enough evidence to be treated as if they are on an interval level of measurement when performing analyses across multiple participants. However, the level of measurement analyses also revealed problems with the equation [see (1)] SART uses to synthesize dimensions into a single rating. Thus, if analysts plan to use subjective SA as part of their analyses, the results of this study suggest that it can be collected directly and not computed from its subdimensions as is done by SART.

References

[1] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems: Situation awareness," *Human Factors*, vol. 37, no. 1, pp. 32–64, 1995.

[2] M. R. Endsley, "Measurement of situation awareness in dynamic systems," *Human Factors*, vol. 37, no. 1, pp. 65–84, 1995.

[3] S. J. Selcon and R. M. Taylor, "Evaluation of the situational awareness rating technique (SART) as a tool for aircrew systems design," in *Proc. AGARD, Situational Awareness Aerosp. Oper.*, 1990, pp. 5-1–5-8.

[4] R. Taylor, "Situational awareness rating technique (SART): The development of a tool for aircrew systems design," in *Proc. AGARD, Situational Awareness Aerosp. Operations*. Seuilly-Sur Seine: NATO AGARD, 1989, pp. 3/1–3/17.

[5] H. Engelbrecht, S. G. Lukosch, and D. Datcu, "Evaluating the impact of technology assisted hotspot policing on situational awareness and task-load," *Proc. ACM Interactive, Mobile, Wearable, Ubiquitous Technol.*, vol. 3, no. 1, pp. 1–18, 2019.

[6] M. Cover, C. Reichlen, M. Matessa, and T. Schnell, "Analysis of airline pilots subjective feedback to human autonomy teaming in a reduced crew environment," in *Proc. Int. Conf. Human Interface Manage. Inf.*, Cham, Switzerland: Springer, 2018, pp. 359–368.

[7] R. K. Mehta *et al.*, "Operator situation awareness and physiological states during offshore well control scenarios," *J. Loss Prevention Process Ind.*, vol. 55, pp. 332–337, 2018.

[8] W. Zhang *et al.*, "Optimal time intervals in two-stage takeover warning systems with insight into the drivers' neuroticism personality," *Frontiers Psychol.*, vol. 12, 2021, Art. no. 601536.

[9] E. T. Evans *et al.*, "Usability evaluation of indicators of energy-related problems in commercial airline flight decks," in *Proc. IEEE/AIAA 38th Digit. Avionics Syst. Conf.*, 2019, pp. 1–9.

[10] J. Yoo *et al.*, "A real-time autonomous dashboard for the emergency department: 5-year case study," *JMIR mHealth uHealth*, vol. 6, no. 11, 2018, Art. no. e10666.

[11] M. Oberhauser and D. Dreyer, "A virtual reality flight simulator for human factors engineering," *Cogn., Technol. Work*, vol. 19, no. 2, pp. 263–277, 2017.

[12] L. Petersen, L. Robert, J. Yang, and D. Tilbury, "Situational awareness, driver's trust in automated driving systems and secondary task performance," *SAE Int. J. Connected Auton. Veh., Forthcoming* vol. 2, no. 2, pp. 129–141, 2019.

[13] D. R. Eignor, *The Standards for Educational and Psychological Testing*. Washington, DC, USA: American Psychological Association, 2013.

[14] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946.

[15] J. Wei, M. L. Bolton, and L. Humphrey, "The level of measurement of trust in automation," *Theor. Issues Ergonom. Sci.*, vol. 22, no. 3, pp. 274–295, 2021. [Online]. Available: https://doi.org/10.1080/1463922X.2020.1766596

[16] J. Wei, M. L. Bolton, and L. Humphrey, "Subjective measurement of trust: Is it on the level?," in *Proc. Int. Annu. Meeting Human Factors Ergonom. Soc.*, 2019, pp. 212–216.

[17] W. E. Deming, *Statistical Adjustment of Data*. Hoboken, NJ, USA: Wiley, 1943.

[18] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Mahwah, NJ, USA: Lawrence Erlbaum, 1988.

[19] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, 1988.

[20] S. Rasmussen, D. Kingston, and L. Humphrey, "A brief introduction to unmanned systems autonomy services (UxAS)," in *Proc. Int. Conf. Unmanned Aircr. Syst.*, 2018, pp. 257–268.

[21] M. R. Endsley, S. J. Selcon, T. D. Hardiman, and D. G. Croft, "A comparative analysis of SAGAT and SART for evaluations of situation awareness," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, vol. 42. Los Angeles, CA, USA: Sage, 1998, pp. 82–86.

[22] M. R. Endsley, "The divergence of objective and subjective situation awareness: A meta-analysis," *J. Cogn. Eng. Decis. Making*, vol. 14, no. 1, pp. 34–53, 2020.

[23] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *J. Comparative Neurol. Psychol.*, vol. 18, no. 5, pp. 459–482, 1908.

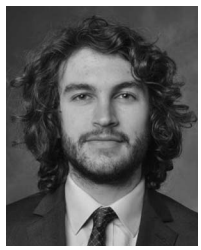
[24] J. Jaccard, C. K. Wan, and J. Jaccard, *LISREL Approaches to Interaction Effects in Multiple Regression*. Newbury Park, CA, USA: Sage, 1996.

[25] D. R. J. Laming, *The Measurement of Sensation*. London, U.K.: Oxford Univ., 1997.

[26] M. L. Bolton, "Modeling human perception: Could Stevens' power law be an emergent feature?," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2008, pp. 1073–1078.

Matthew L. Bolton (Senior Member, IEEE) 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 currently an Associate Professor with the Department of Industrial and Systems Engineering, University at Buffalo, The State University of New York, Amherst, NY, USA. His research focuses on the use of human performance modeling and formal methods in the design and analysis of safety-critical systems.



Elliot Biltekoff received the B.A. degree in cognitive science, and the M.S. degree in human factors engineering in 2018 and 2020, respectively, from the University at Buffalo, The State University of New York, Amherst, NY, USA, where he is currently working toward the Ph.D. degree in industrial engineering.

His research focuses on building computational models of psychophysical phenomena and their associated reasoning processes to understand cognitive mechanisms .

Laura Humphrey (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from The Ohio State University, Columbus, OH, USA, in 2004, 2006, and 2009, respectively.

She is currently a Senior Research Engineer with the Aerospace Systems Directorate, Air Force Research Laboratory, Wright-Patterson Air Force Base, OH, USA. Her research focuses on the use of formal methods for design and implementation of autonomous and human-automation systems.