On the Level of Measurement of Subjective Psychometric Ratings

Matthew L. Bolton, University of Virginia
Elliot Biltekoff, University at Buffalo
Jiajun Wei, Intuitive
Laura Humphrey, Air Force Research Lab

Subjective psychometric ratings are critical tools in human factors engineering. Despite widespread use, there is controversy about whether such ratings can be treated at cardinal levels of measurement. Answering this question is important given that a measure's level determines what mathematical and statistical operations can be meaningfully applied to them. This paper synthesizes results from an effort (published over several papers) that developed a method for assessing the level of measurement of subjective ratings and used it to evaluate common trust, situation awareness, and workload metrics. This paper synthesizes these results to determine that subjective measures should rarely, if ever, be treated as ratio. It also found that, in a population analysis, some measures (like all the ones considered) can be treated as interval. In all other situations, and especially in analyses or modeling efforts for individuals or small sample sizes, ordinal is the safest default.

INTRODUCTION

For subjective 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 product review scores, to disease diagnoses, to scientific research, to the engineering of safety-critical systems. Such measures often constitute the most direct measure of psychological concepts. They are thus a critical tool in the human factors engineering toolbox.

For use in science, subjective ratings are expected to have satisfied several criteria (Eggemeier, Wilson, Kramer, & Damos, 1991; Eignor, 2013). Scales should be reliable (produce consistent results across observations and studies), valid (correlate with phenomena associated with what is being measured), selective (sensitive to the quality being assessed), diagnostic (indicate the reason for changes), and unintrusive (not interfere with the task being performed or phenomena being measured).

A consideration that has received less attention is level of measurement. Level of measurement indicates what numbers on a scale mean relative to each other (Stevens, 1946). Level of measurement is important because it determines what mathematics and statistics can be meaningfully applied to measurement transformation (like in modeling) and analyses (Stevens, 1946). Despite the importance of this issue and the widespread use of subjective psychometrics, there is no clear consensus on how such measures should be treated with respect to the levels of measurement (Annett, 2002; Furr & Bacharach, 2013; Guilford, 1954; Michell, 2008; Velleman & Wilkinson, 1993).

To fill this gap, we developed a method [introduced by Wei, Bolton, and Humphrey (2019, 2020) and refined by Biltekoff, Bolton, and Humphrey (n.d.); Bolton, Biltekoff, and Humphrey (n.d.)] that enables researchers to empirically evaluate the level of measurement of subjective psychometric ratings scales. This method was used to assess multiple ratings scales that are commonly used in human factors engineering: trust (Wei et al., 2020); the three dimensions of the situation awareness rating technique (SART) along with situation awareness itself (Bolton et al., n.d.); and the six dimensions of the NASA task load index (NASA-TLX) (Biltekoff et al., n.d.).

This paper seeks to examine the large number of measures evaluated across these studies in an attempt to draw broader conclusions about the level of measurement of subjective ratings. To accomplish this, we first present background on level of measurement, our method, and the studies we conducted to evaluate the measures listed above. We then present a unique aggregation of the results from across these studies. We discuss these results and their implications for how subjective psychometrics are used.

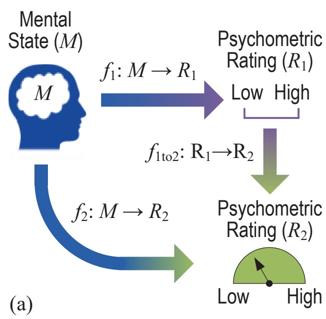
BACKGROUND

Levels of Measurement

Psychological research usually focuses on four levels of measurement (Stevens, 1946). Nominal-level scales represent identity or category (e.g., player number on a sports team). Ordinal scales capture order (e.g., class rank). On interval scales (e.g., temperature in Fahrenheit), the distances between numbers are meaningful. However, interval scales lack a zero that indicates the absence of the measured quantity. Hence, ratios between numbers are not meaningful. Finally, numbers from ratio scales (e.g. distance) have a meaningful zero and thus meaningfully capture ratios between numbers.

A scale's level determines what mathematics and statistical methods can be meaningfully used with values measured on that scale (Stevens, 1946). Nominal scales are compatible with equalities/inequalities, counts, modes, and contingency correlations. Ordinal scales are compatible with greater-than and less-than relationships, percentiles, medians, and rank-order statistics. Numbers measured on interval scales are compatible with means, standard deviations, product moment correlations, and most parametric statistics. Finally, ratio scales allow for the meaningful application of percent changes, geometric means, and coefficients of variation. Mathematical power increases with the level of measurement as ordered above (from nominal to ratio). All meaningful operations at a lower level are meaningful at higher levels. However, the opposite is not true. As such, researchers will want to treat measures at the level that offers the most analytical power.

There is no consensus about what the highest possible level subjective ratings scales can assume. Practitioners prefer to treat them at the interval level (Furr & Bacharach, 2013; Guilford, 1954) to enable computation of means, standard deviations, and most parametric statistics. Many experts do not think subjective scales can be safely treated at levels above ordinal, and especially not ratio (Furr & Bacharach, 2013; Guilford,



If M is ordinal and R_1 and R_2 can be ordinal or better:

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

$\therefore f_{1to2}(f_1(m_a)) = f_2(m_a)$ and $f_{1to2}(f_1(m_b)) = f_2(m_b)$

$\therefore f_{1to2}(f_1(m_a)) < f_{1to2}(f_1(m_b))$

(b)

If M is interval and R_1 and R_2 can be interval or better:

$\forall\, m \in M$: $f_1(m) = a_1 \cdot m + b_1$ 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)

If M is ratio and R_1 and R_2 can be ratio or better:

$\forall\, m \in M$: $f_1(m) = a_1 \cdot m$ 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 $a_{1to2} = a_2 / a_1$

$f_{1to2}(r_1) = a_{1to2} \cdot r_1$

(d)

Figure 1. Figure showing the relationships used by the method for assessing the level of measurement of psychological concepts (adapted from (Bolton et al., n.d.)). (a) Shows possible transformations between mental state M and scales R_1 and R_2. f_1 and f_2 represent the transformation a person would make to measure M on scales R_1 and R_2 respectively. f_{1to2} is a transformation of values from R_1 to R_2. (b), (c), and (d) show how the level of M influences the form that f_{1to2} will take. (b) assumes M is ordinal, (c) assumes it is interval, and (d) presumes it is ratio. a_1, a_2, b_1, and b_2 are constants.

1954). However, common measures such as NASA-TLX (Hart & Staveland, 1988) and ratings of trust (Lee & Moray, 1992) appear to treat psychometrics as if they are ratio.

Method for Assessing Level of Measurement

The method for assessing the level of measurement of subjectively assessed psychological phenomena (Wei et al., 2020) uses permissible transformations as its theoretical base. Permissible transformations offer a way of reasoning about level of measurement. A permissible transformation is one that shows how a number from a given scale can be converted to a different scale while preserving the level. These can be any one-to-one (identity preserving) transformation on a nominal scale; any strictly increasing (order preserving) transformations for ordinal scales; any linear transformation of the form $f_{interval}(X) = a \cdot X + b$ for interval scales, where a and b are constants and a is positive; and any ratio transformation $f_{ratio}(X) = a \cdot X$ for ratio scales, where a is a positive constant.

Figure 1 shows the concept behind the method. First, assume that there are two psychometric scales R_1 and R_2 that measure the same psychological quality M (Figure 1(a)). When a human assesses the state of M and convert this into a rating on R_1 or R_2, he or she must transform the value of M onto these scales by applying the transformations: $f_1 : M \rightarrow R_1$ and $f_2 : M \rightarrow R_2$ respectively. Thus, as is shown in Figure 1(b)–(d), as long as R_1 and R_2 can capture the level of M, M's measurement level permissible transformation will determine the form taken by f_1 and f_2. This, in turn, determines the form that a transformation from R_1 to R_2 (f_{1to2}) will take. This means that while M, f_1, and f_2 are unobservable and cannot be examined to evaluate level of measurement, f_{1to2} is observable. Thus, the form taken by f_{1to2} can indicate M's level of measurement.

In the assessment method (Wei et al., 2020), it is assumed that, as long as observations on R_1 and R_2 are distinct, there is enough evidence that a scale is at least nominal. Evidence that the scale is at least ordinal is measured via a nonparametric Spearman's ρ correlation. Because the permissible transformations for both ratio and interval scales assumes a linear form, a Deming linear regression (Deming, 1943) (one where error is assumed in both predictor and response variables) can determine if there is evidence of interval or ratio relationships. If the produced regression model has a significant intercept (0 is not in a 95% confidence interval around the intercept), then there is evidence for (at least) an interval scale. If the intercept is not significant (0 is in the confidence interval), there is evidence for a ratio level. Note that R^2 is not computed for Deming regression models because Deming regression does not use least squares during model fitting. Because of this, a Pearson's correlation coefficient (r) is used to measure the "fit" (linear relationship) between the measures (this is the standard).

Table 1. *Heuristic (Bolton et al., n.d.) for Assessing the Level of Measurement for a Given Participant's (or collection of Participants') Subjective Responses*

Level	Evidence Strength: Weak ○	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

This assumes three judgment scales (thus three transformations/models between judgments). ρ and r are Spearman's and Pearson's correlation coefficients respectively. Standard thresholds (Cohen, 1988) assess coefficient strength. *CI* is a 95% confidence interval around the regression model's intercept.

While the method only requires human judgments on two scales to assess a scale's level of measurement, using additional measures reduces the likelihood of an incorrect conclusion. Thus, all preceding applications of the method have used three scales (Biltekoff et al., n.d.; Bolton et al., n.d.; Wei et al., 2020). When using three scales, participants make judgments for randomly ordered, identical experimental conditions in three blocks, one for each judgment scale. Analyses are then performed (using the statistics described above) to understand the transformation/models between each pair of scales.

The heuristic in Table 1 assesses the strength of evidence of the measures being at least at a given level of measurement. These are based on standard interpretations of the strength of correlations (Cohen, 1988) and the number of comparisons that showed evidence at different levels of strength.

Method Application Experiments

This method was used to evaluate a number of subjective psychometric scales across three different experiments. The scales assessed were chosen because they are the most common subjective scales used for assessing trust, situation awareness, and mental workload (three of the most important cognitive engineering methods). Thus, the first experiment evaluated subjective trust in automation (Wei et al., 2020). The second evaluated the three dimensions of SART (demand on attentional resources, supply of attentional resources, understanding) as well as overall situation awareness (Bolton et al., n.d.). The third evaluated the six dimensions of the NASA-TLX for assessing human mental workload (mental demand, physical demand, temporal demand, performance, effort, and frustration) (Biltekoff et al., n.d.). All three received approval from the University at Buffalo IRB under STUDY00002118. They also all followed a similar experimental procedure and design.

Procedure. In all experiments, participants watched simulations of Unmanned Aerial Systems (UASs) performing various (line, point, area) search tasks. In the SART and NASA-TLX experiments, participants were asked to dynamically indicate which (if any) of six marked points were searched as the simulation ran. In all experiments, the same set of experiment-specific simulations were observed in three blocks, where the participants provided ratings about the simulation using the scales being evaluated in the experiment. For the trust experiment, this was to "indicate how much you would trust the observed UAS to perform search tasks." In the other two, these were to rate the dimensions of SART or NASA-TLX for the monitoring and target identification tasks. Standard language was used to administrating these techniques (Hart & Staveland, 1988; Selcon & Taylor, 1990; Taylor, 1989).

Participants. In each experiment, 36 University at Buffalo graduate engineering student participants were recruited. This resulted in 108 participants across all the studies.

Materials and Apparatus. The experiment was run on personal computers with a mouse and keyboard via software that was specifically created to administer the experiment.

During a given experimental trial, the software showed a video of a UAS simulation (Figure 2). Simulation videos were created using UxAS and AMASE (Rasmussen, Kingston, & Humphrey, 2018) with video capture software.

Independent Variables. In the experiments, variables were varied along dimensions specifically identified to influence the ratings that were being collected (see Biltekoff et al. n.d.; Bolton et al. n.d.; Wei et al. 2020 for details).

Dependent Measures. The dependent measures were defined by the specific ratings being evaluated in each experiment: trust in the first experiment; demand, supply, understanding, and situation awareness in the second; and mental demand, physical demand, temporal demand, performance, effort, and frustration in the third. All of an experiment's ratings were measured for each simulation that was shown to participants. After a given simulation, the measures were collected with one of the experiment's three judgment modalities. These ultimately constitute the separate scales used by the method, like R_1 and R_2 from Figure 1. In the so-called "ask" modality, each dimension was measured as a floating-point number from 0 to 100 that was entered into a text box. With the knob, dimensions were measured as a floating-point number from 0 to 100 based on the onscreen position of a knob between its minimum (0°) and maximum (300°) positions (controlled via a physical knob / scroll wheel connected to the computer). With the slider modality, dimensions were measured as a floating-point number from 0 to 100 based on the left-to-right position of a slider.

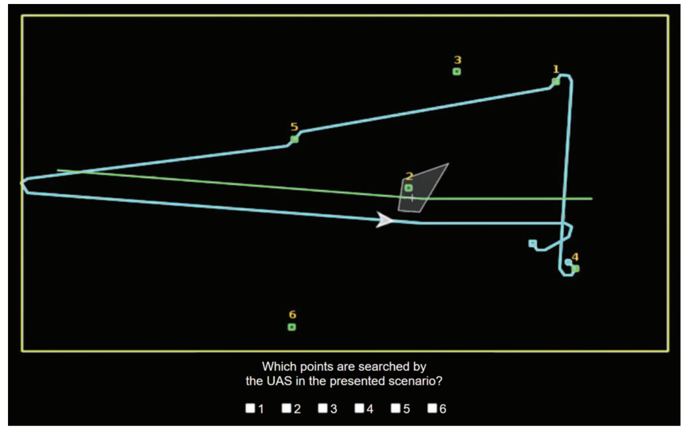


Figure 2. A screenshot of the type of UAS simulation used in experiments. The UAS was depicted as a chevron shape moving through a set area. A "footprint" of the UAS's camera (an area surrounded by a gray line) showed the ground area the camera was capturing. A cross in the footprint indicated the center of the camera's view. Check boxes on the perimeter of the simulation appeared in the SART and NASA-TLX experiments to allow participants to indicate which points were searched by the UAS during the simulation.

Experimental Design. Each experiment created a set of trials constituting a full factorial design of its independent variables' levels. Four additional training trials were created that varied across all variable dimensions. At the start of the experiment, a participant was assigned three random orders of the experimental trials: one unique order for each judgment modality. Trials for a given modality were presented in blocks and the order of these blocks was counterbalanced between participants.

Training trials introduced participants to experimental tasks and judgment modalities. At the beginning of an experiment, participants saw four training trials. All subsequent judgment modality trial blocks were introduced with two training trials. Within an experiment, training trials was consistently ordered for all participants regardless of judgment modality order.

Data Analysis. For all three experiments, the level of measurement was assessed for each participant and across all participants for all collected scales both for each individual participant and across all participants using the method and heuristic (Table 1) discussed previously in this section.

AGGREGATED RESULTS

Figure 3 visualizes the results of each scales' level of measurement for individuals. Figure 3(a)–(c) show the total number of participants who exhibited evidence (both weak and strong; see Table 1) for the scales from the (a) trust, (b) SART, and (c) NASA-TLX experiments. Figure 3(d) shows the results of aggregating the data from (a)–(c) across all assessed ratings.

Table 2 shows the evidence strength found when all participants for each scale were considered in aggregate.

DISCUSSION AND CONCLUSIONS

According to Stevens' (1975), psychological measurement is "the assignment of numerals to objects and events according to a rule." With this definition, the measures evaluated across

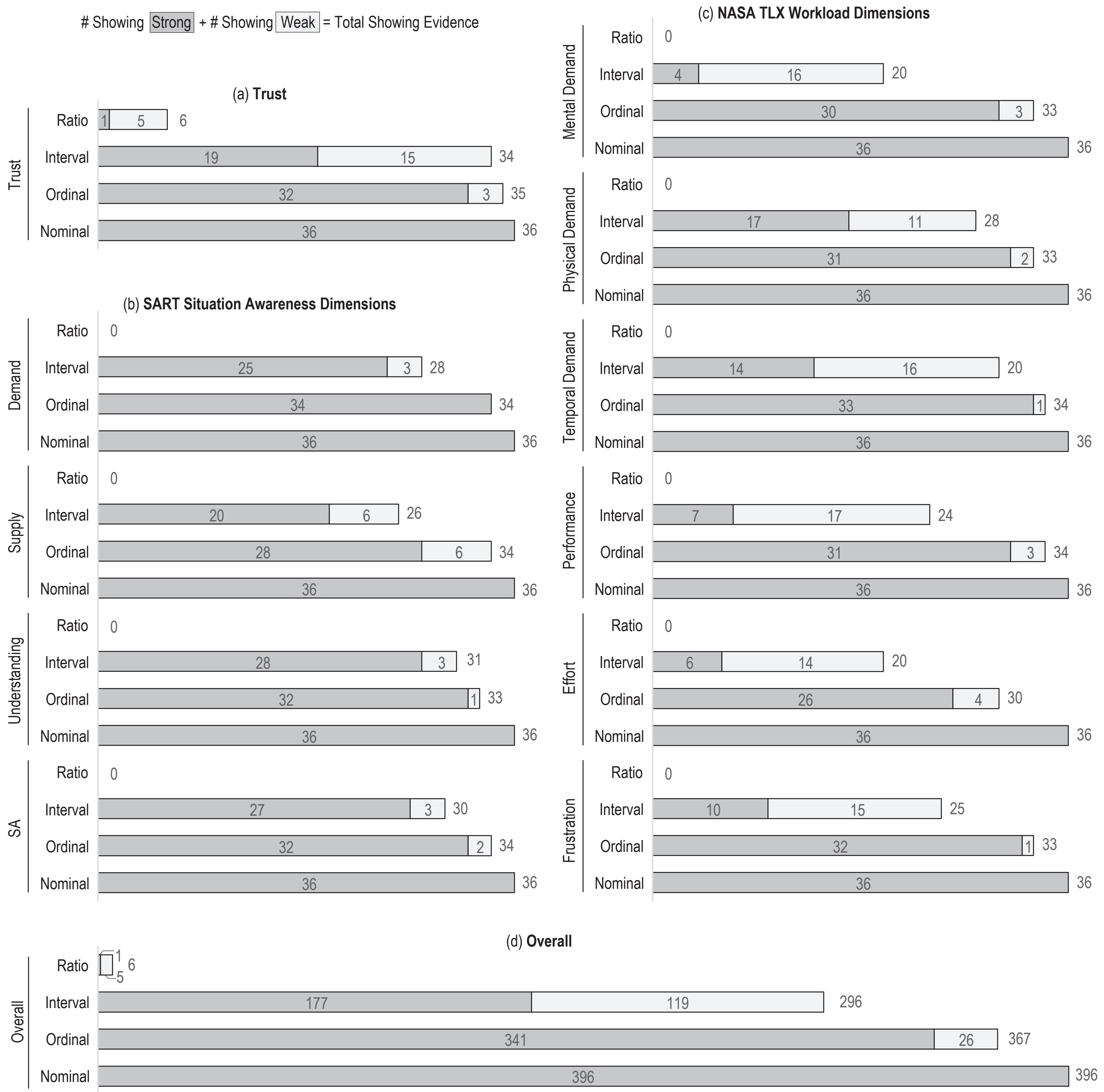


Figure 3. Stacked bar charts showing the number of participants presenting evidence (strong, weak, and combined; see Table 2) evidence of each level of measurement for the three experiments (a)–(c) and aggregated across all measures (d). In (a)–(c), numbers are out of 36. In (d) numbers are out of 396.

these experiments constitutes some of the best candidates for subjective measures being ratio. That is, all of these scales have been measured using the traditional pen and paper technique where: participants place marks on a horizontal line, the rating is taken as the distance of the mark to the left side of the line, and the rating is scaled to a number between 0 and 100 based on the length of line. Thus, the rule for creating the measures is consistent with a ratio level because a physical measurement is at the ratio level (Narens, 2012). Further, all of the modalities used in this experiment (the slider, itself designed to mimic standard electronic analogues of the pen and paper approach; the knob which goes from 0°to 300°; and the ability to enter any number between 0 and 100 in the ask modality) are capable of producing ratio data. Despite this, the studies produced very little evidence that people (individually or in aggregate) treat the measured concepts as ratio: only 6 (1 strong and 5 weak; Figure 3(a)) individuals for trust and, in aggregate (Table 2), only physical demand. This provides compelling evidence that researchers should generally (without explicit evidence to the contrary) assume that subjective ratings are not ratio.

Table 2. *Evidence Strength (see Table 1) Observed For Each Evaluated Measure when Considering Data From All Participants in a Single Analysis*

	Evidence for Level of Measurement			
Measure	Nominal	Ordinal	Interval	Ratio
Trust	**Strong**	**Strong**	**Strong**	
Demand	**Strong**	**Strong**	**Strong**	
Supply	**Strong**	**Strong**	**Strong**	
Understanding	**Strong**	**Strong**	**Strong**	
SA	**Strong**	**Strong**	**Strong**	
Mental Demand	**Strong**	**Strong**	Weak	
Physical Demand	**Strong**	**Strong**	**Strong**	**Strong**
Temporal Demand	**Strong**	**Strong**	**Strong**	
Performance	**Strong**	**Strong**	Weak	
Effort	**Strong**	**Strong**	Weak	
Frustration	**Strong**	**Strong**	**Strong**	

The aggregate results across all participants (Table 2) show compelling evidence that scales can achieve the interval level. All evaluated scales showed evidence of intervality, with this evidence being strong for all but three scales. This is an encouraging result because it implies that the standard practice of analyzing multiple participants' subjective ratings with parametric statistics (i.e. T-tests and ANOVAs) is valid.

Evidence of intervality also manifested at the individual level: more than 55% of participants showed some evidence of this level (Figure 3(a)–(c)); eight of the twelve scales had more than 70% of participants show evidence of being at least interval; and across all of the measures (overall; Table 2(d)), 74.74% (296 out of 396) showed evidence of intervality. However, this still means significant numbers of participants did not evidence interval-level data for the scales. This ultimately means that researchers need to be cautious when evaluating subjective measures collected from individuals or small sample sizes. In these situations, unless researchers have evidence that the measured humans are using interval levels, they should default to treating data as if it is ordinal (a level for which nearly all participants showed strong evidence).

This last point is extremely important given that modern automation, robotics, and autonomy concepts are attempting to have machines monitor individual humans and adapt their behavior based on predictions or assessments of human situation awareness, workload, and trust. Given that such systems may be making safety and performance critical decision, it is essential that they treat these subjective concepts at a mathematically meaningful level. Based on the results of the presented studies, it appears that this needs to be ordinal.

There has been considerable debate surrounding Stevens' levels of measurement. Part of this has speculated that humans might actually think about subjective concepts measured on scales at a level somewhere between ordinal and interval (Velleman & Wilkinson, 1993). That is, scales that are not quite interval, but have enough numerical structure to be treated as if they are cardinal, enabling their analysis with parametric statistics (Norman, 2010). The weak evidence seen for interval level measurement for the scales reported here (especially those for the workload experiment; Figure 3(c)), appear to add weight to this argument. Future research should further investigate whether this level exists and, if so, characterize its properties.

Finally, to the best of our knowledge, the method introduced in these efforts is the first to actually characterize the level of measurement of subjective psychometrics. There are many more psychometrics that could be analyzed including usability, teaching evaluation, and emotional responses. Other scales may varyingly exhibit different levels. However, given the fidelity of the scales we evaluated, we would be surprised to find a subjective scale that consistently produces ratio levels. This will hopefully be borne out in future research.

ACKNOWLEDGEMENT

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

REFERENCES

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

Biltekoff, E., Bolton, M. L., & Humphrey, L. (n.d.). The level of measurement of the NASA task load index and its constituent dimensions. *IEEE Transactions on Human-Machine Systems*, 9 pages. (Under Review)

Bolton, M. L., Biltekoff, E., & Humphrey, L. (n.d.). The level of measurement of subjective situation awareness and its dimension in the situation awareness rating technique (SART). *IEEE Transactions on Human-Machine Systems*, 8 pages. (In Press) doi: 10.1109/THMS.2021.3121960

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

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

Eggemeier, F. T., Wilson, G. F., Kramer, A. F., & Damos, D. L. (1991). Workload assessment in multi-task environments. In *Multiple task performance* (pp. 207–216). CRC Press.

Eignor, D. R. (2013). *The standards for educational and psychological testing*. American Psychological Association.

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.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in psychology*, *52*, 139–183.

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

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

Narens, L. (2012). *Theories of meaningfulness*. Psychology Press.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, *15*(5), 625–632.

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).

Selcon, S. J., & Taylor, R. M. (1990). Evaluation of the situational awareness rating technique (SART) as a tool for aircrew systems design. In *AGARD, situational awareness in aerospace operations* (pp. 5-1–5-8).

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

Stevens, S. S. (1975). *Psychophysics*. Transaction Publishers.

Taylor, R. (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.

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

Wei, J., Bolton, M. L., & Humphrey, L. (2019). Subjective measurement of trust: Is it on the level? In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 212–216). Los Angeles: Sage.

Wei, J., Bolton, M. L., & Humphrey, L. (2020). The level of measurement of trust in automation. *Theoretical Issues in Ergonomics Science*, *22*(3), 274–295.