The Mathematical Meaninglessness of the NASA Task Load Index: A Level of Measurement Analysis

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

***Abstract*—Human mental workload can profoundly impact human performance and is thus an important consideration in the design and operation of many systems. The standard method for assessing human mental workload is the NASA Task Load Index (NASA-TLX). This involves a human operator subjectively rating a task based on six dimensions. These dimensions are combined into a single workload score using one of two methods: scaling and summing the dimensions (where scales are derived from a paired comparisons procedure) or averaging dimensions together. Despite its widespread use, the level of measurement of NASA-TLX's dimensions and its computed workload score has not been investigated. Additionally, nobody has researched whether NASA-TLX's two approaches for computing overall workload are mathematically meaningful with respect to the constituent dimensions' levels of measurement. This is a serious deficiency. Knowing what the level of measurement is for NASA-TLX scores will determine what mathematics can be meaningfully applied to them. Furthermore, if NASA-TLX workload syntheses are mathematically meaningless, then the measure lacks construct validity. The research presented in this article used a previously developed method to evaluate the level of measurement of NASA-TLX workload and its dimensions. Results show that the dimensions can, in most situations, be treated as interval in population analyses and ordinal for individuals. Our results also suggest that the methods for combining dimensions into workload scores are meaningless. We recommend that analysts evaluate the dimensions of NASA-TLX without combining them.**

***Index Terms*—Human performance assessment, psychometrics and testing, workload.**

Mental workload describes the human mental resource demands placed on a person at a given time. Mental workload is regarded as an important metric when assessing human work because levels that are too high or too low can adversely affect human performance: reducing efficiency, increasing the likelihood of human error, and creating undesirable conditions for human workers. Mental workload can be measured in multiple ways. The most direct methods tend to use subjective self-assessments. In these, humans perform representative work tasks and then rate their workload (either directly or along constituent dimensions) on a numerical scale. The most widely used scale is the NASA Task Load Index (NASA-TLX) [1]. In the NASA-TLX, workload is first assessed along the six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. These are then synthesized into a final, overall workload score.

When employing subjective ratings scientifically (like those used for assessing the constituent dimensions of NASA-TLX), there are multiple criteria that must be satisfied [2], [3]. Measurement scales must be reliable, meaning they produce consistent results across multiple observations. Scales must be valid, meaning they correlate with phenomena associated with the thing being measured. Scales should be selective in that they should be sensitive to the quality being measured, but not other variables. Scales should be diagnostic: They should be able to diagnose the reasons for changes in the measured phenomena. Scales should also not be intrusive. This means that the act of collecting the measure should not interfere with task performance in a way that impacts what is being measured. To its credit, the NASA-TLX has generally shown itself to satisfy these criteria (caveats are discussed in Section V-C) and has become the de facto standard for measuring mental workload [1], [3], [4], [5], [6], [7], [8], [9]. It has been used across analysis domains, adapted to more than a dozen languages, and generally shown itself to be at least as sensitive and useful as alternative measures [5]. However, a consideration that has received little empirical attention when assessing psychometric scales (including the NASA-TLX) is level of measurement [10]. Level of measurement relates to the meaning of the numbers, and differences between numbers, on a scale. Thus, level of measurement is important because it determines how measures can be meaningfully, mathematically synthesized into others (such as the constituent workload dimension being converted into overall workload), as well as the types of statistics that can be meaningfully applied to measures [10]. Thus, despite the success of the NASA-TLX, nobody has examined its level of measurement, nor has anybody evaluated whether the syntheses of its dimensions into workload scores are consistent with the dimensions' levels of measurement. This is a critical gap because it means that NASA-TLX workload computations and/or statistics used for analyzing NASA-TLX results may not be, strictly speaking, mathematically meaningful. With critical design and safety decisions being made based on NASA-TLX results, addressing this concern is important.

Manuscript received 20 April 2022; revised 11 March 2023; accepted 26 March 2023. Date of publication 23 May 2023; date of current version 8 June 2023. This work was supported by the Air Force Research Lab/Universal Technology Corporation/ARCTOS Technology Solutions under Grant FA8650-1.6-C-2642 and Grant 18-S8401-13-C1. This article was recommended by Associate Editor C. Wu. *(Corresponding author: Matthew L. Bolton.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the University at Buffalo's IRB under Application No. STUDY00002118.

Matthew L. Bolton is with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22903 USA (e-mail: mlb4b@virginia.edu).

Elliot Biltekoff is with the Department of Systems and Industrial Engineering, University at Buffalo, Buffalo, NY 14260 USA (e-mail: elbiltek@buffalo.edu).

Laura Humphrey is with the Air Force Research Lab, WPAFB, 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.2023.3263482.

Digital Object Identifier 10.1109/THMS.2023.3263482

Wei et al. [11] recently introduced a method for identifying the level of measurement of psychological phenomena assessed with subjective measures. This has been used to assess the level of measurement of trust in automation [11], [12] and situation awareness (evaluated using the subjective situation awareness rating technique) [13].

In this research, we used the method introduced in these previous studies [11], [12], [13] to assess the level of measurement of NASA-TLX workload, the constituent dimensions of the NASA-TLX, and the validity of the methods for synthesizing workload using the dimensions.

I. Background

A. Level of Measurement

The level of measurement determines what numbers on a given scale mean, the meaning of differences between numbers measured on that scale, and thus the meaningfulness of mathematical operations on that scale. While many different levels exist in measurement theory, psychological measurement generally has the following four levels as posited by Stevens [10]. 1) Nominal scales represent identity or category (e.g., student identification number). 2) Ordinal scales capture order (e.g., the place one finishes in a race). 3) On interval scales (e.g., temperature in Celsius), the distances between numbers are meaningful. However, because there is no meaningful zero on an interval scale (zero does not indicate that none of the measured quantity exists), ratios between numbers are not meaningful. 4) Finally, ratio scales (e.g., length) have meaningful zeros and, thus, there is meaning between ratios of numbers on these scales.

As stated above, a scale's level determines what mathematics and statistics can be meaningfully used with values measured on that scale [10]. Nominal scales are compatible with equalities/inequalities, counts, modes, and contingency correlations. On ordinal scales, comparisons can account for greater-than and less-than relationships, percentiles, and medians; rank-order statistics are also meaningful. Numbers measured on interval scales are compatible with means, standard deviations, product moment correlations, and the majority of parametric statistics. Finally, ratio scales allow for the meaningful application of percent changes, geometric means, and coefficients of variation. It is important to note that mathematical power increases with the level of measurement in the order that they are presented above (nominal, ordinal, interval, and ratio). This means that meaningful operations at a lower level can be applied to all higher levels. Thus, all meaningful operations on nominal scales are meaningful for all other scales. For this reason, practitioners will want to treat measures at the highest possible level to allow for the most analytical power.

One way of reasoning about level of measurement is with permissible transformations. Permissible transformations describe how numbers on a given scale are converted to different scales while preserving the level of measurement. On a nominal scale, these can be any one-to-one transformation: one that preserves identity. On ordinal scales, any strictly increasing function (i.e., one that preserves element order) is a permissible transformation. For interval scales, the permissible transformation is any linear function $f_{\text{interval}}(X) = a \cdot X + b$, where a (a scaling factor) and b (a reposition of the arbitrary zero) are constants. Finally, ratio scales have permissible transformations of the form $f_{\text{ratio}}(X) = a \cdot X$, where the original number is only scaled by a constant (a).

B. Method for Assessing Level of Measurement

The method for determining the level of measurement of subjectively assessed psychological phenomena [11] uses meaningful transformations as its theoretical base. Fig. 1 shows the concept behind the method. First, assume that there are two psychometric scales R_1 and R_2 that both measure a given psychological quality M [Fig. 1(a)]. When a human psychologically assesses the state of M and attempts to convey this as a rating on R_1 or R_2, he or she must transform the value of M onto these scales by applying the respective transformations: $f_1 : M \to R_1$ and $f_2 : M \to R_2$. Thus, as is shown in Fig. 1(b)–(d), as long as R_1 and R_2 can capture the level of M, M's level will dictate the form taken by f_1 and f_2 based on the level's permissible transformations. This, in turn, determines the form that a transformation from R_1 to R_2 ($f_{1\text{to}2}$) will take. While M, f_1, and f_2 are unobservable, $f_{1\text{to}2}$ is observable. Thus, the form of $f_{1\text{to}2}$ can be used as a means of determining M's level of measurement.

The method [11] assumes that, as long as observations on R_1 and R_2 are distinct, there is sufficient evidence that a scale is at least nominal. Evidence for ordinality is measured via a Spearman's ρ correlation (which is nonparametric). The permissible transformations for both ratio and interval scales assume a linear form. Thus, a linear regression can determine if there is evidence of interval or ratio relationships. Deming regression [14] is appropriate for characterizing this relationship because error can occur on both R_1 and R_2. If the produced regression model has a significant intercept (0 is not in the confidence interval around the intercept), then there is evidence for 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 given that ordinary least squares is not used in its fitting process. Because of this, a Pearson's correlation coefficient (r) is used to measure the "fit" (linear relationship) between the measures. This is standard practice for Deming regression.

While the method only requires human judgments on two scales to assess a psychological phenomenon's level of measurement, the use of more measures reduces the likelihood of an incorrect conclusion. Thus, all preceding applications of the method [11], [12], [13] have used three scales. In this situation, 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 shown in Table I is then used to assess the strength of evidence of the psychological phenomenon being at least a given level of measurement.

C. NASA Task Load Index

As covered in the introduction, NASA-TLX [1] is the leading subjective workload assessment tool. It measures workload for a given task by obtaining subjective ratings from 0 to 100 on six subscales/dimensions and then synthesizing them into an overall workload value. These dimensions are as follows.

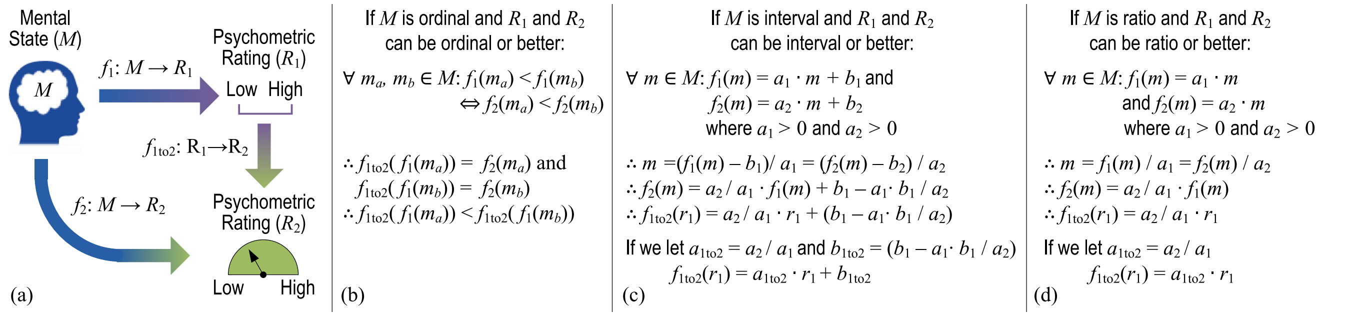

Mental State (M)
M
f_1: $M \rightarrow R_1$
Psychometric Rating (R_1)
Low High
$f_{1\text{to}2}$: $R_1 \rightarrow R_2$
f_2: $M \rightarrow R_2$
Psychometric Rating (R_2)
Low High
(a)
If M is ordinal and R_1 and R_2 can be ordinal or better:
$\forall\ m_a, m_b \in M$: $f_1(m_a) < f_1(m_b) \Leftrightarrow f_2(m_a) < f_2(m_b)$
$\therefore f_{1\text{to}2}(f_1(m_a)) = f_2(m_a)$ and $f_{1\text{to}2}(f_1(m_b)) = f_2(m_b)$
$\therefore f_{1\text{to}2}(f_1(m_a)) < f_{1\text{to}2}(f_1(m_b))$
(b)
If 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 $a_2 > 0$
$\therefore m = (f_1(m) - b_1)/\ a_1 = (f_2(m) - b_2)\ /\ a_2$
$\therefore f_2(m) = a_2\ /\ a_1 \cdot f_1(m) + b_1 - a_1 \cdot b_1\ /\ a_2$
$\therefore f_{1\text{to}2}(r_1) = a_2\ /\ a_1 \cdot r_1 + (b_1 - a_1 \cdot b_1\ /\ a_2)$
If we let $a_{1\text{to}2} = a_2\ /\ a_1$ and $b_{1\text{to}2} = (b_1 - a_1 \cdot b_1\ /\ a_2)$
$f_{1\text{to}2}(r_1) = a_{1\text{to}2} \cdot r_1 + b_{1\text{to}2}$
(c)
If 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 $a_2 > 0$
$\therefore m = f_1(m)\ /\ a_1 = f_2(m)\ /\ a_2$
$\therefore f_2(m) = a_2\ /\ a_1 \cdot f_1(m)$
$\therefore f_{1\text{to}2}(r_1) = a_2\ /\ a_1 \cdot r_1$
If we let $a_{1\text{to}2} = a_2\ /\ a_1$
$f_{1\text{to}2}(r_1) = a_{1\text{to}2} \cdot r_1$
(d)

Fig. 1. Demonstration of the concepts used by the level of measurement assessment method (adapted from [13]). (a) Shows possible transformations between mental state M onto two 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. $f_{1\text{to}2}$ 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 $f_{1\text{to}2}$ 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.

TABLE I
Heuristic (Reproduced From [13]) for Assessing the Level of Measurement for a Given Participant's 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 $\lvert CI \rvert \leq 20$
	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 heuristic assumes three judgment scales (and thus three transformations between judgments) are used in the method. ρ is a Spearman's correlation coefficient. r is a Pearson's correlation coefficient. Standard methods [15] are used to assess coefficient strength. CI is a 95% confidence interval around the linear Deming regression model's intercept.

1) *Mental* demand: How mentally demanding the human perceived the task.
2) *Physical* demand: How physically demanding the human perceived the task.
3) *Temporal* demand: How temporally demanding the human perceived the task.
4) *Performance:* How successful the human felt he or she was at accomplishing the task goals.
5) *Effort:* How hard the human felt he or she worked to accomplish his or her level of performance.
6) *Frustration:* How insecure, discouraged, irritated, stressed, and annoyed the person was during the task.

Rating on these scales are collected via questionnaires that are administered after completion of the task being evaluated.

In its original form (the "paper and pencil" version) [16], all six dimensions are measured by having people place a mark on a 12-cm line that is divided into 20 even intervals by 21 vertical lines (see Fig. 2). The actual score assigned to the dimension is then measured as the distance of the mark from the left-end side of the line, scaled to between 0 and 100. In the traditional formulation, all of the scales are measured from low (on the left) to high (on the right), with the exception of *Performance*. This is measured with high on the left and low on the right. *Performance* can be measured in the same direction as the other dimensions, but this requires additional processing when computing workload (discussed subsequently).

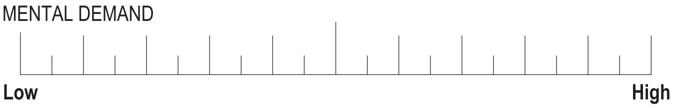


Fig. 2. Example of a "paper and pencil" scale used for measuring dimensions of mental workload with NASA-TLX.

In this traditional approach, measured values for each dimension are synthesized into an aggregate workload rating by eliciting subjective weightings for each dimension. These are computed by having participants make pairwise comparisons between the possible 15 unique dimension pairs to indicate which is the most important to mental workload. The number of times a given dimension is selected divided by 15 creates a weighting between 0 and 1 that constitutes the given dimension's contribution to workload. This results in a formula of the form (see [1] and [16])

$$\begin{aligned} Workload = {} & w_{\text{Mental}} \cdot Mental \\ & + w_{\text{Physical}} \cdot Physical \\ & + w_{\text{Temporal}} \cdot Temporal \\ & + w_{\text{Performance}} \cdot Performance \\ & + w_{\text{Effort}} \cdot Effort \\ & + w_{\text{Frustration}} \cdot Frustration \end{aligned} \tag{1}$$

where each w_x represents the weighting for a given dimension x.

Most conventional applications of the NASA-TLX do two things to simplify application. First, scales are administered on computers, tablets, or phones to avoid having to measure dimension ratings from a sheet of paper [17]. Second, each of the six dimensions are treated as if they have equal weights. This reduces the amount of data that need to be collected from participants and simplifies the equation for computing workload to [18]

$$\begin{aligned} Workload = (& Mental + Physical \\ & + Temporal + Performance \\ & + Effort + Frustration)/6. \end{aligned} \tag{2}$$

This is sometimes called the "raw" workload score. Both of these approaches were found to have no statistically significant impact on results for several datasets [17], [18], [19] (though broader applicability has shown sensitivity between both synthesis approaches to vary depending on the domain [5]).

Note that if *Performance* is measured from low to high, then *Performance* in (1) and (2) is replaced with

$$(100 - Performance). \tag{3}$$

D. NASA-TLX and Level of Measurement

While there is no explicit declaration of the level of measurement assumptions of the NASA-TLX, these can be inferred from its design and use. First, across the scientific literature, NASA-TLX scores are used with parametric statistics, especially t-tests and analyses of variance. Thus, clearly the scientific community views NASA-TLX-measured workload and its constituent dimensions as if they are at least interval. Second, the traditional scaling approach which results in the dimension synthesis from (1) treats all the constituent dimensions as if they can be ratio transformed to the workload scale. This suggests that all constituent dimensions, and thus workload itself, are assumed to be at the ratio level. Third, the revised method that produces the synthesis in (2), because it clearly assumes the dimensions can be averaged, is treating workload and the dimensions as if they are at least interval. Many psychometrics experts do not think psychometric scales are capable of providing ratio measures [20], [21]. Many even doubt that subjective ratings can be treated as anything higher than ordinal [22], [23], [24], [25], [26], [27], [28], including Stevens [29] himself (the person who originally defined the levels of psychological measurement). Thus, there is good reason to doubt whether the NASA-TLX approaches to computing workload are valid with respect to the level of measurement.

II. Objectives

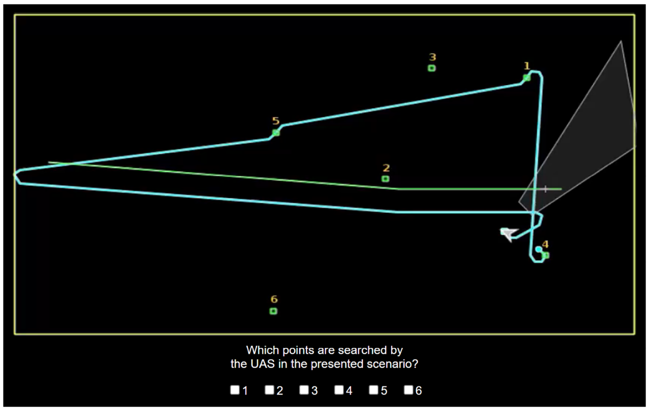
In this research, we sought to evaluate the level of measurement of mental workload as measured by NASA-TLX (along with its constituent dimensions) using the method from [12]. In doing this, we also sought to evaluate the validity of the level of measurement assumptions inherent in the two approaches [characterized by (1) and (2)] to synthesizing the dimensions of workload into a single workload score. In what follows, we describe how this was achieved using a human subjects experiment and analyses consistent with the approaches discussed in previous method applications [11], [12], [13].

III. Methods

This research was approved by the University at Buffalo's IRB under STUDY00002118.

A. Procedure

This experiment's procedure was based on the ones established in [11], [12], [13], but modified to accommodate social distancing necessitated by the Covid-19 pandemic. Specifically, the experiment was administered remotely using a project website and the Zoom conferencing system. Participants signed into a prescheduled zoom meeting after signing an electronic informed consent. They then observed a prerecorded video that introduced them to the experiment and its tasks. They performed the experiment via a website. This began when they entered an ID given to them by the experimenter.



Which points are searched by the UAS in the presented scenario?

□1 □2 □3 □4 □5 □6

Fig. 3. UAS simulation screenshot.

The experiment involved watching simulations of unmanned aerial systems (UASs; air vehicles that fly autonomously or are piloted remotely) performing search tasks and dynamically indicating which points were searched as the simulation ran. The same set of simulations were presented three time: Once in each of three blocks, where participants rated their workload using NASA-TLX using the three different judgment methods (modalities) from the previous experiment.

This task and application were chosen for the following three reasons: 1) it offered enough variables around which to create different demands on NASA-TLX dimensions (see Section III-D); 2) the application domain (human interaction with UAS search tasks) was of interest to the research sponsor. 3) this task allowed us to create multiple, independent, and short scenarios to facilitate the economic collection of the all the measurements required by our method.

B. Participants

Thirty-six University at Buffalo graduate engineering students (19 males and 17 females with ages between 22–30 years) participated in this study.

C. Materials and Apparatus

The experiment was run on the participants' computers via a web application that was created for this experiment. Before the experiment, participants were instructed to use a computer with a working internet connection, keyboard, mouse, and compatible web browser (Google Chrome, Mozilla Firefox, and Microsoft Edge).

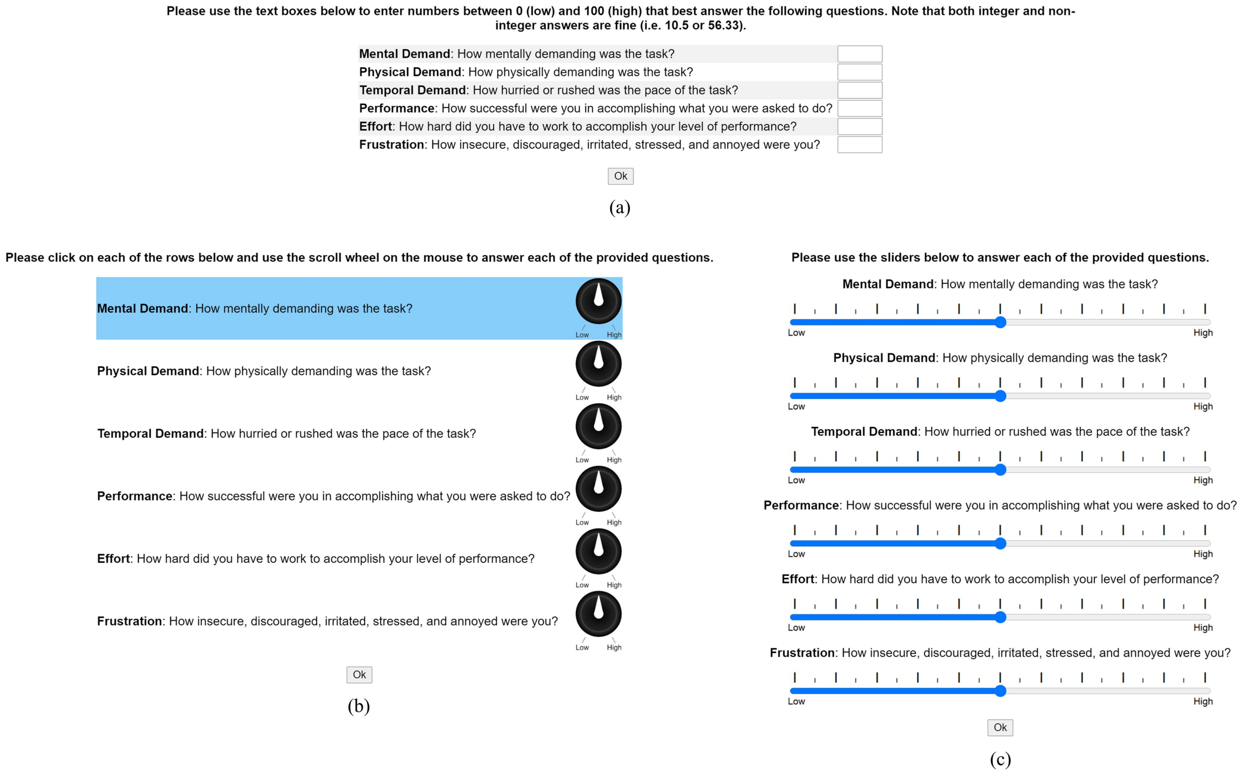
During each given trial, the web application showed a video of a UAS performing search tasks within a set region while flying (Fig. 3). Simulations were created using unmanned systems autonomy services (UxAS) and the aerospace multi-agent simulation environment (AMASE) [30]. The UAS was presented as a chevron moving through the area. A "footprint" of the UAS's camera (a gray-lined shape) conveyed the ground area captured by the camera. A cross in the footprint showed the camera's center view.

While the simulation was playing, the UAS performed point (greens squares labeled with numbers; Fig. 3) and line (the green line in Fig. 3) searches. After all were finished, the UAS flew to an end point and loitered. When the flight path was visible (as in Fig. 3), it was depicted as a blue line.

TABLE II
INDEPENDENT VARIABLES AND THEIR LEVELS

Variable	Description	Levels
CursorNoise	The mouse cursor could move randomly with varying levels of noise*	None, medium (max of 6 pixels), high (max of 12 pixels)
NumPoints	The number of points searched by the UAS	0, 1, ..., 6
Path	The simulation may or may not display the flight path	Visible, invisible
PlaySpeed	The simulation may play at regular speed or fast speed†	Regular, fast (1.6x)
Points	The specific points searched during the simulation	Randomly assigned
Radius	The simulations limit the radius of what you can see around the UAS	Small (25 pixels), medium (50 pixels), large (75 pixels)

All simulations were presented as videos that were 748 pixels wide and 438 pixels high.
*Mouse cursor position noise was computed every 70 ms. This re-positioned the mouse cursor away from its genuine position (according to the computer) randomly (based on a uniform distribution) up to a maximum level of pixels (dictated by the levels). Deviations were computed independently for both the horizontal and vertical and could be positive or negative.
† Regular speed had the UAS move at 90 pixels per second. Thus, fast speed was 1.6 times faster at 144 pixels per second.



Please use the text boxes below to enter numbers between 0 (low) and 100 (high) that best answer the following questions. Note that both integer and non-integer answers are fine (i.e. 10.5 or 56.33).

Mental Demand: How mentally demanding was the task?
Physical Demand: How physically demanding was the task?
Temporal Demand: How hurried or rushed was the pace of the task?
Performance: How successful were you in accomplishing what you were asked to do?
Effort: How hard did you have to work to accomplish your level of performance?
Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?

(a)

Please click on each of the rows below and use the scroll wheel on the mouse to answer each of the provided questions.

Mental Demand: How mentally demanding was the task?
Physical Demand: How physically demanding was the task?
Temporal Demand: How hurried or rushed was the pace of the task?
Performance: How successful were you in accomplishing what you were asked to do?
Effort: How hard did you have to work to accomplish your level of performance?
Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?

(b)

Please use the sliders below to answer each of the provided questions.

Mental Demand: How mentally demanding was the task?
Physical Demand: How physically demanding was the task?
Temporal Demand: How hurried or rushed was the pace of the task?
Performance: How successful were you in accomplishing what you were asked to do?
Effort: How hard did you have to work to accomplish your level of performance?
Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?

(c)

Fig. 4. Screenshots of the web-based dialog boxes used for collecting ratings. (a) Participants use a keyboard to enter numbers ranging from 0 to 100. (b) On-screen knobs are turned to positions using the computer mouse's scroll wheel. (c) Sliders are moved with the computer's mouse.

During simulation playback, participants were tasked with indicating which of six displayed and labeled points (green boxes with numbers above them) were searched. There were always six points present in a simulation. Participants would indicate if they thought a point was searched by clicking on the correspondingly labeled checkbox below the simulation (see Fig. 3). The UAS would search any points that fell on the flight path (like points 1, 4, and 5 in Fig. 3) or points that passed through the camera's footprint during line searches (like point 2 in Fig. 3).

After each simulation, participants were asked to provide ratings about their workload via NASA-TLX's six dimensions using the interfaces in Fig. 4: a) a ratio number between 0 and 100 entered into a text box; b) the position of an on-screen knob (adjusted by clicking on a measure's area to select it and changing the knob's position using the mouse's scroll wheel); c) the position of on-screen sliders (adjusted using the mouse). Note that to ensure the polarity of *Performance* was consistent between the judgment modalities (you cannot measure *Performance* using the ask modality with a traditional inverted scale), *Performance* was always measured from low to high in the same direction as the other dimensions.

D. Independent Variables

The independent variables and their associated levels are shown in Table II. Each of these variables were selected because their values were designed to vary task difficulty/workload along the dimensions of NASA-TLX.

Mental demand should be impacted by all the factors, but particularly *NumPoints*, *Path*, and *Radius* because these affect how closely the person must attend to the simulation to identify searched points. Increasing *NumPoints* increases the number of identifications a person must make and simultaneously document as the simulation unfolds. Given, the quickness of the simulation and the time it takes to document search, this can provide varying degrees of load on human attention and working memory. When the search *Path* is visible, participants can see the route the UAS will take through the environment and, thus, be able to more easily anticipate if a displayed point will be searched. The absence of this path, in turn, will make this projection more difficult and place more load on attention. Finally, the display radius varyingly shows and limits the area the person can observe and thus determine what points are or could potentially be searched. Thus, smaller radii will inherently place higher demands on human attention.

Physical demand is impacted by *CursorNoise*, *NumPoints*, and *PlaySpeed*. *CursorNoise* influences how easily one can select searched points physically with the mouse. *NumPoints* determines how many physical point identification/documentation activities must be undertaken. Finally, *PlaySpeed* influences how physically quickly point identification will need to occur.

Temporal demand is potentially impacted by *PlaySpeed*, *Radius*, *Path*, and *CursorNoise*. All three influence how much time people have to identify and select points. *PlaySpeed* does this by limiting the length of the scenario and decreasing the amount of time between searched points. *Radius* and *Path* do this by reducing the amount of information people have for anticipating whether a point will be searched. Finally, the amount of *CursorNoise* increases the amount of time a human will need to identify a searched point.

Performance and effort should be impacted by all the factors.

Finally, frustration is particularly impacted by *CursorNoise*. This parameter was specifically included to make it difficult to select a searched point and thus make the task frustrating. *PlaySpeed* and *Radius* could also impact frustration, most likely due to their associated impact on temporal demand.

In combination, these factors created scenarios that were expected to elicit a range of workload responses from low (no cursor noise, zero searched points, a visible flight path, a regular play speed, and a larger radius) to high (high cursor noise, six searched points, an invisible flight path, a fast play speed, and a small radius).

E. Dependent Measures

The dependent measures were six ratings aligning with the six NASA-TLX dimensions, made using each of the three judgment modalities (Fig. 4): *Mental* demand; *Physical* demand; *Temporal* demand; *Performance*; *Effort*; *Frustration*.

All six of these dimensions 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. In the so-called "ask" modality [Fig. 4(a)], each dimension was measured as a floating-point number from 0 to 100 that was entered into a text box. With the knob [Fig. 4(b)], 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. With the slider modality [Fig. 4(c)], dimensions were measured as a floating-point number from 0 to 100 based on its position (left-to-right) between the controls "low" and 'high" labels. These judgment modalities were selected because, as per the requirements of our method (see Section I-B), they offer enough expressive power to capture numbers that fall at any level of measurement, up to and including ratio.

F. Experimental Design

We created a set of 36 trials, one for each of the possible combination of the levels of the *CursorNoise*, *Path*, *PlaySpeed*, and *Radius* independent variables ($3 \cdot 2 \cdot 2 \cdot 3 = 36$). Within these, *NumPoints* and *Points* values were assigned randomly. Four training trials were also created. These varied across all trial geometry dimensions.

At the start of the experiment, a participant was assigned three random (unique) orders of the 36 experimental trials: one for each judgment modality. Trials for a given modality (i.e., ask, knob, and slider, with their associated judgment interfaces; see Fig. 4) were presented in a block. The order of judgment modality blocks was counterbalanced between participants.

Training trials were used to introduce participants to the experimental task as well as the different judgment modalities. When the experiment started, participants saw all of the four training trials. All subsequent judgment modality trial blocks were introduced with two training trials. The order of training trials was consistently ordered for all participants regardless of judgment modality order.

G. Data Analysis

The level of measurement was assessed for each participant and across all participants for all six dependent measures using the method in [11] and [13], and Section I-B as follows.

1) Computing Spearman's (ρ) and Pearson's (r) correlation coefficients as well as Deming regression models for paired ratings made between judgment modalities.
2) Heuristically assessing the strength of evidence (Table I).

Overall "raw" workload (*Workload*; the modern standard) for each participant's ratings was computed as the arithmetic average of all six workload dimensions using (2), with the *Performance* substitution from (3). The level of measurement was also assessed for this measure (*Workload*) for each participant and across all participants.

As discussed previously, the original workload computation formulation from (1) appears to assume that the scales of the individual dimensions are at the ratio level. Thus, the level of measurement assessments of these dimensions should be sufficient to assess the validity of (1). However, some additional analyses are required to test the assumption made by the averaging approach from (2): that all the workload dimensions are on the same scale. If this is true, then we would expect the Deming regression models that convert between judgment modality pairs to be the same for all six dimensions. For example, the model for converting a participant's *Mental* demand from an ask judgment to a knob judgement should be the same as converting *Physical* demand, *Temporal* demand, *Performance*, *Effort*, and *Frustration* between the same judgment modality pairs. We used repeated measure analyses of variance (ANOVAs) to evaluate this. In these analyses, the slopes and intercepts from the Deming regression models were the response variables. Dimension was the independent factor and the combination of the participant and judgment modality pair was the "subject" factor.

IV. Results

A full listing of results and computed statistics are reported in the article's supplementary materials (supplement Figs. 1 to 7 show scatter plots and fitted Deming regression lines of all individual data for each workload dimension and judgment modality, Tables I to VII show the statistics computed for each participant in accordance with our method; Fig. 8 and Table VIII show this same information for the combined analysis). Due to space constraints, a summary of results are presented. Across all the participants, a full range of ratings were produced for each of the workload dimensions for each of the different judgment modalities (see supplement Figs. 1 to 6), suggesting that our experiment did affect all the dimensions of NASA-TLX. Most importantly, Fig. 5 (synthesized from the statistics reported across supplement Tables I to VII) shows the total number of individual participants that displayed both weak and strong evidence (as per Table I) that each of the evaluated measures was at a given level of measurement. When all the participants were considered in aggregate (where the results of all were pooled and evaluated together; see supplement Fig. 8 and Table VIII), strong evidence was observed for nominal and ordinal levels for all the measures (including *Workload*). Strong evidence of intervality was seen for *Physical* demand, *Temporal* demand, *Frustration*, and *Workload*. Only weak evidence of intervality was seen for the others. Strong evidence for a ratio scale was observed for *Physical* demand; no other measures showed any evidence of the ratio level.

The repeated measures' ANOVAs, which checked whether the slopes and intercepts were the same for converting between the measures collected from participants, both showed significant differences ($\alpha = 0.05$) for regression model slopes [$F_{5,510} = 2.956$, $p = 0.012$, $\eta_p^2 = 0.028$; Fig. 6(a)] and intercepts [$F_{5,510} = 5.540$, $p < 0.001$, $\eta_p^2 = 0.052$; Fig. 6(b)] between measures. However, Mauchly's tests indicated violations of sphericity ($\chi^2(5) = 0.026$, $p < 0.001$; and $\chi^2(5) = 0.002$, $p < 0.001$). Thus, Greenhouse–Geisser corrections were applied ($\epsilon = 0.405$ and $\epsilon = 0.268$). With these, slope was not significant ($F_{2.068,210.954} = 2.956$, $p = 0.052$) but intercept was ($F_{1.341,136.793} = 5.540$, $p = 0.012$, $\eta_p^2 = 0.052$).

To both limit alpha inflation and limit loss of power, a Hsu's multiple comparisons with the best [31] was performed to test for differences between levels of intercept. That is, comparing the dimension with the maximum average intercept (which was *Effort*) to the five others using one-tailed paired t-tests with a Holm step-down correction [32]. This showed that *Effort* had a significantly larger intercept from Physical demand ($p = 0.005$), *Frustration* ($p = 0.005$), *Performance* ($p = 0.007$), Mental demand ($p = 0.012$), and *Temporal* demand ($p = 0.010$) (see Fig. 6).

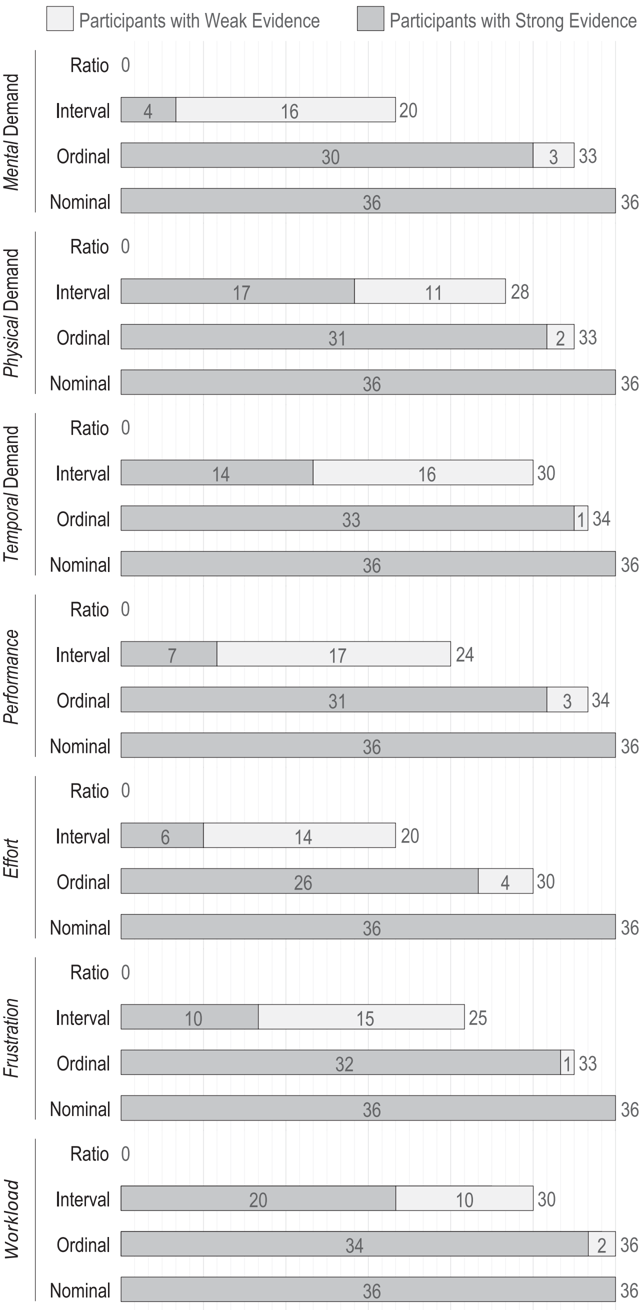


Fig. 5. Stacked bar charts showing the number of participants (out of 36) that showed evidence (see Table I) of the different levels of measurement for each dependent measure and *Workload* computed using (2) [with the *Performance* substitution; (3)]. Dark gray bars represent the number of participants that showed strong evidence. Light gray bars indicate the number that showed weak evidence. Numbers in bars indicate how many participants exhibited the associated strength of evidence for the given level. Numbers following stacked bars indicate the total number that showed any evidence for the given level.

V. Discussion

This work constitutes a unique effort to determine the level of measurement of workload and its constituent dimensions from NASA-TLX. In what follows, we discuss our results, their significance, and outlets for future research.

A. Level of Measurement of Workload Dimensions

The results across the NASA-TLX dimensions and *Workload*, computed using the modern averaging method from (2), were generally consistent. Strong evidence was present that most individuals treat all as being at least ordinal (Fig. 5): each measure had at least 72.22% of participants (26 out of 36) showing strong evidence and 83.33% (30 out of 36) showing at least weak evidence. Evidence for the interval level was present, but much weaker: with only between 11.11% (for *Mental*) and 47.22% (for *Physical*) showing strong evidence and between 55.55% (for *Mental* and *Effort*) and 83.33% (for *Temporal*) showing at least weak evidence. No participants showed any evidence for a ratio level.

When considering the aggregate results (where all participant data were considered together; supplement Fig. 8 and Table VIII), the results become more encouraging. Strong evidence of

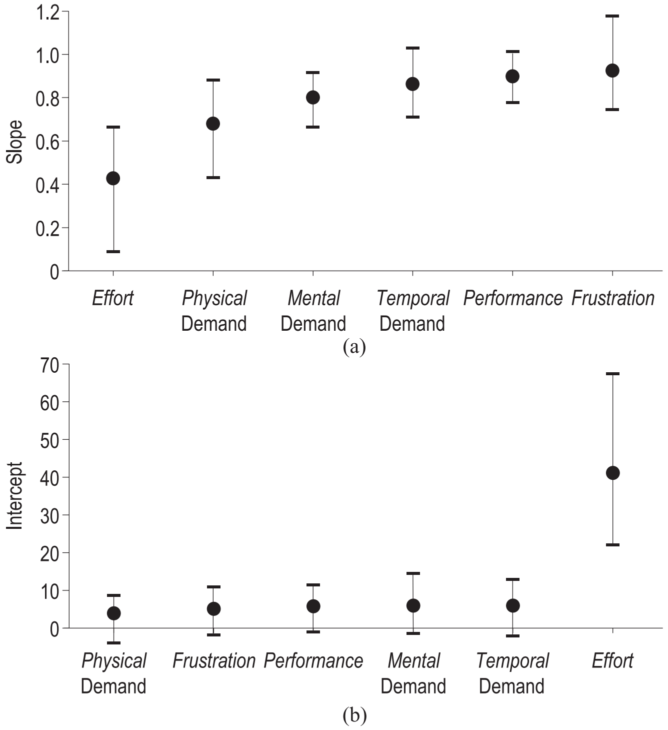


Fig. 6. Graphs showing the within subject 95% confidence intervals around the mean for regression model (a) slopes and (b) intercepts for each of the workload dimensions, sorted from smallest to largest.

at least nominality and ordinality was observed for all measures, and at least weak evidence for all scales being interval. One dimension (*Physical*) showed strong evidence of the ratio level.

Collectively, these results indicate that it is safe to treat *Workload* and its dimensions as if they are interval. This is a positive development because it suggests that the standard parametric statistics that most researchers use to evaluate NASA-TLX are valid. However, when considering measurements from an individual, analysts should be more careful. In this situation, it is probably prudent to treat the dimension and *Workload* as being ordinal unless specific evidence exists that the individual is treating them at a higher level.

B. *Workload Computation*

Despite these positive findings, our results do suggest problems with both approaches for computing workload. First, the overwhelming lack of evidence for the ratio level in individual results suggests that it is invalid to use separate ratio scaling factors to synthesize ratings on each dimension into a workload score. Thus, the original approach to computing workload using (1) is not producing meaningful values based on our result. Furthermore, the ANOVA results of regression model intercepts showed significant differences between dimensions. This indicates that the modern, "raw" method for computing workload as an average [as in (2)] is also not meaningful. These results provide additional support for the practice advocated by Galy et al. [33]: that NASA-TLX dimensions should be considered separately rather than synthesized into a single score due to a lack of independence between dimensions.[1]

Note, our ANOVAs and post hoc were not specifically designed to probe where differences in NASA-TLX scales manifest, but rather to detect that differences exist. That said, our analyses (and Fig. 6) do appear to suggest that the Effort dimension may be on a different scale than the other dimensions. The analysis by Galy et al. [33] may shed light on this discrepancy. These researchers found that, in a regression analysis, *Physical* demand, *Mental* demand, and alertness (an additional factor in their study) all had significant main effects on effort. This suggests that *Effort* may be a more complex measure than the other NASA-TLX ones. Many variations of the TLX (including the original [1]) have people rate Effort using the question: "How hard did you have to work (mentally and physically) to accomplish your level of performance?" Additionally, mental effort load is a singular dimension of the subjective workload assessment technique (SWAT) [34], a popular alternative to the NASA-TLX. All of this suggests that additional research is required to determine if and/or how *Effort* should be accounted for in mental workload computations.

Similarity observed between the interval scales for all but the Effort dimension should be treated with healthy skepticism. Given the supposedly well-founded nature of the concepts being analyzed by the NASA-TLX dimensions, there is little reason to think that level of measurement should change with the domain being analyzed (e.g., if *Frustration* is a consistent concept, and people think about it cardinally in one situation, they should think about it cardinally in others; we discuss this topic more in the next section). The same cannot be said for the actual scale somebody may use. In fact, interval generally being the ceiling on scale level makes a lot of sense based on what we know about the effect of context on workload and other subjective response [35]. Specifically, subjective ratings can be compressed or amplified based on the varying amount of stimuli exhibited during training or over an experiment. This suggests people, generally, adjust the interval scale on which they make workload ratings based on their experience. Thus, the similarity seen between the interval scales of NASA-TLX dimensions may not hold in other task domains and *Effort* may not be as uniquely different as our results suggest. While the study documented in this article was occurring, independent researchers evaluated an alternative method for developing dimension scaling factors for (1) [36]. The produced approach was designed to address limitations of the original scaling process that: prevent direct assignment of equal importance between dimensions, force importance order between all factors, and artificially limit the maximum weight a factor can be given. While this research effort is addressing very real additional problems with the NASA-TLX, their recommendations will not address the problems that come from assuming ratio level scales that are made by (1).

C. *Validity of NASA-TLX*

Level of measurement has not historically been considered in validation efforts of psychometrics. This is likely due to the fact that, until the introduction of our method [11], [12], [13], there was no way of performing such an assessment. Thus, the work here constitutes a cutting-edge validation assessment of

[1] Hart, the primary researcher behind NASA-TLX, casually acknowledges this phenomenon in a literature review of NASA-TLX uses from 2006 [5].

NASA-TLX. Given that this evaluation has revealed problems in something as established as the NASA-TLX, it is our contention that the level of measurement assessment should become standard practice in psychometric validation efforts.

While our analysis found problems, our results do provide some evidence in support of NASA-TLX's validity. Specifically, it has shown that it is safe to treat workload dimensions at a cardinal level (interval) in aggregate analyses and that (with caution) there may be ways forward for synthesizing subdimensions into a workload score (see Section V-D2 for additional discussion). This (along with the measure's reliability) helps provide evidence to address a major criticism of NASA-TLX [37] by showing that the subdimensions constitute real measurable qualities.

This said, the issues we found with synthesizing subdimensions into a workload score constitute a serious violation of construct validity: The conceptual equivalent of averaging different temperatures, some in Fahrenheit and some in Celsius. The literature shows other construct validity problems with NASA-TLX. This includes failure of NASA-TLX to converge with other (objective and performance) measures [7], [38], [39] as well as previously noted correlations between dimensions [5], [33] and artificial constraints in subdimension scaling [35]. All this suggest that there are issues with the NASA-TLX and that, if it is used, caution is warranted until these problems can be resolved or alternative measures established.

D. Areas of Future Research

The advances reported here suggest many directions for future research. We explore some of them below.

1) Other Task Domains and Demographics: The specific task domain evaluated in this experiment is not a standard one for evaluating NASA-TLX. However, we feel it is consistent enough with president for our results to be valid. NASA-TLX has been regularly used in various UAS operation studies (e.g., [40], [41], and [42]) and monitoring has been classed as being in the top six types of tasks NASA-TLX has been used to evaluate [5].

That said, we are fully cognizant that one study in one domain is not necessarily sufficient to reveal systemic deficiencies in something as established as the NASA-TLX. Thus, future research should attempt to apply our method to other task domains (and with more diverse demographics) to see if the results presented here hold. In particular, tasks like those from the multiattribute task battery [43], [44] would be worthy of investigation. Note that an experiment in this domain would likely require a substantial time commitment to support the multiple, independent trials necessitated by our method.

2) Workload Computation Possibilities: The results presented here suggest that existing methods for computing workload from the NASA-TLX dimensions are not mathematically meaningful according to measurement theory. However, given that many people show evidence that the dimensions are interval, it may be possible to develop an overall workload formula that properly accounts for this (as well as potential confounds between dimensions [33]). In fact, our results suggest that only the effort dimension appears to be on a different interval scale than the other dimensions. Future work should investigate whether this holds for other work domains. If it does, effort could be dropped from the "raw" workload computation and possibly even replaced by additional dimensions (such as alertness as explored by Galy et al. [33]). Beyond being able to more accurately compute mental workload, such a development would also enable analyses to evaluate how the problems of (1) and (2) may have impacted conclusions about mental workload across the scientific literature.

3) Other Measures of Workload: Common alternatives to the NASA-TLX include the SWAT [34], [45], workloadprofile [46], and subjective workload dominance [47]. Future work should evaluate the measurement theory assumptions of these approaches and adapt our level of measurement technique to evaluate them.

4) Other Subjective Measures: This study and the trust [11], [12] and situation awareness [13] studies represent the first to evaluate the level of measurement of subjectively assessed psychological concepts (Bolton et al. [48] describes all studies' aggregate implications for subjective measurement). There are many such measures used across society and the scientific literature. This includes usability, product review scores, and teaching evaluations. Future research should investigate the level of measurement of these and other psychometric scales.

VI. Conclusion

In this research, we found evidence that mental workload and its dimensions from NASA-TLX can be, when considered across participants, treated as interval numbers. They should be treated with more care (and possibly only at an ordinal level) when individual scores are under consideration. However, our analyses also showed that the traditional approaches to computing workload from the dimensions violate the levels of measurement of those dimensions or are done on noncomparable scales. This implies that workload as computed for the NASA-TLX is mathematically meaningless. Thus, we conclude that analysts should evaluate the dimensions of mental workload from NASA-TLX separately rather than synthesizing them into a single workload score.

Acknowledgment

The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy of their affiliated institutions or sponsors.

References

[1] 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.

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

[3] F. T. Eggemeier, G. F. Wilson, A. F. Kramer, and D. L. Damos, "Workload assessment in multi-task environments," in *Multiple Task Performance*. Boca Raton, FL, USA: CRC Press, 1991, pp. 207–216.

[4] Y. Xiao, Z. Wang, M. Wang, and Y. Lan, "The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index," *Chinese J. Ind. Hyg. Occup. Dis.*, vol. 23, no. 3, pp. 178–181, 2005.

[5] S. G. Hart, "Nasa-task load index (NASA-TLX); 20 years later," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2006, vol. 50, pp. 904–908.

[6] P. Ø. Braarud, "Investigating the validity of subjective workload rating (NASA TLX) and subjective situation awareness rating (SART) for cognitively complex human–machine work," *Int. J. Ind. Ergonom.*, vol. 86, 2021, Art. no. 103233.

[7] R. D. McKendrick and E. Cherry, "A deeper look at the NASA TLX and where it falls short," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2018, vol. 62, pp. 44–48.

[8] S. Rubio, E. Díaz, J. Martín, and J. M. Puente, "Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods," *Appl. Psychol.*, vol. 53, no. 1, pp. 61–86, 2004.

[9] L. Longo, C. D. Wickens, G. Hancock, and P. A. Hancock, "Human mental workload: A survey and a novel inclusive definition," *Front. Psychol.*, vol. 13, 2022, Art. no. 883321.

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

[11] 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, 2020.

[12] J. Wei, M. L. Bolton, and L. Humphrey, "Subjective measurement of trust: Is it on the level?," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2019, vol. 63, pp. 212–216.

[13] M. L. Bolton, E. Biltekoff, and L. Humphrey, "The level of measurement of subjective situation awareness and its dimensions in the situation awareness rating technique (SART)," *IEEE Trans. Hum.- Mach. Syst.*, vol. 52, no. 6, pp. 1147–1154, Dec. 2022.

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

[15] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York, NY, USA: Lawrence Erlbaum, 1988.

[16] S. Hart, "NASA task load index (NASA-TLX) v. 1.0 paper and pencil package," *Moffett Field: NASA Ames Res. Center*, 1986.

[17] S. Mach, J. P. Gründling, F. Schmalfuß, and J. F. Krems, "How to assess mental workload quick and easy at work: A method comparison," in *Congress of the International Ergonomics Association*. Berlin, Germany: Springer, 2018, pp. 978–984.

[18] J. C. Byers, A. Bittner, and S. G. Hill, "Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary," *Adv. Ind. Ergonom. Saf.*, vol. 1, pp. 481–485, 1989.

[19] A. Cao, K. K. Chintamani, A. K. Pandya, and R. D. Ellis, "NASA TLX: Software for assessing subjective mental workload," *Behav. Res. Methods*, vol. 41, no. 1, pp. 113–117, 2009.

[20] R. M. Furr and V. R. Bacharach, *Psychometrics: An Introduction*, 2nd ed. Thousand Oaks, CA, USA: Sage, 2013.

[21] J. P. Guilford, *Psychometric Methods*. New York, NY, USA: McGraw-Hill, 1954.

[22] J. Annett, "Subjective rating scales: Science or art?," *Ergonomics*, vol. 45, no. 14, pp. 966–987, 2002.

[23] J. Michell, "Quantitative science and the definition of measurement in psychology," *Brit. J. Psychol.*, vol. 88, no. 3, pp. 355–383, 1997.

[24] J. Michell, "Is psychometrics pathological science?," *Measurement*, vol. 6, no. 1–2, pp. 7–24, 2008.

[25] P. Barrett, "Beyond psychometrics: Measurement, non-quantitative structure, and applied numerics," *J. Managerial Psychol.*, vol. 18, no. 5, pp. 421–439, 2003.

[26] G. Trendler, "Measurement theory, psychology and the revolution that cannot happen," *Theory Psychol.*, vol. 19, no. 5, pp. 579–599, 2009.

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

[28] N. Cliff and J. A. Keats, *Ordinal Measurement in the Behavioral Sciences*. Psychology Press, 2003.

[29] S. S. Stevens, "Mathematics, measurement, and psychophysics," in *Handbook of Experimental Psychology*, S. S. Stevens, Ed. Hoboken, NJ, USA: Wiley, 1951.

[30] 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.

[31] J. Hsu, *Multiple Comparisons: Theory and Methods*. Boca Raton, FL, USA: CRC Press, 1996.

[32] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.

[33] E. Galy, J. Paxion, and C. Berthelon, "Measuring mental workload with the NASA-TLX needs to examine each dimension rather than relying on the global score: An example with driving," *Ergonomics*, vol. 61, no. 4, pp. 517–527, 2018.

[34] G. B. Reid and T. E. Nygren, "The subjective workload assessment technique: A scaling procedure for measuring mental workload," *Adv. Psychol.*, vol. 52, pp. 185–218, 1988.

[35] H. A. Colle and G. B. Reid, "Context effects in subjective mental workload ratings," *Hum. Factors*, vol. 40, no. 4, pp. 591–600, 1998.

[36] K. Virtanen, H. Mansikka, H. Kontio, and D. Harris, "Weight watchers: NASA-TLX weights revisited," *Theor. Issues Ergonom. Sci.*, vol. 23, no. 6, pp. 1–24, 2021.

[37] S. W. Dekker and J. M. Nyce, "From figments to figures: Ontological alchemy in human factors research," *Cogn. Technol. Work*, vol. 17, no. 2, pp. 185–187, 2015.

[38] Y.-Y. Yeh and C. D. Wickens, "Dissociation of performance and subjective measures of workload," *Hum. Factors*, vol. 30, no. 1, pp. 111–120, 1988.

[39] G. Matthews, J. De Winter, and P. A. Hancock, "What do subjective workload scales really measure? operational and representational solutions to divergence of workload measures," *Theor. Issues Ergonom. Sci.*, vol. 21, no. 4, pp. 369–396, 2020.

[40] G. Gabriel, M. A. Ramallo, and E. Cervantes, "Workload perception in drone flight training simulators," *Comput. Hum. Behav.*, vol. 64, pp. 449–454, 2016.

[41] J. Ruiz, A. Viguria, J. Martinez-de Dios, and A. Ollero, "Immersive displays for building spatial knowledge in multi-UAV operations," in *Proc. Int. Conf. Unmanned Aircr. Syst.*, 2015, pp. 1043–1048.

[42] T. M. Schnieders, Z. Wang, R. T. Stone, G. Backous, and E. Danford-Klein, "The effect of human-robot interaction on trust, situational awareness, and performance in drone clearing operations," *Int. J. Hum. Factors Ergonom.*, vol. 6, no. 2, pp. 103–123, 2019.

[43] J. Cegarra, B. Valéry, E. Avril, C. Calmettes, and J. Navarro, "OpenMATB: A multi-attribute task battery promoting task customization, software extensibility and experiment replicability," *Behav. Res. Methods*, vol. 52, no. 5, pp. 1980–1990, 2020.

[44] J. R. Comstock Jr and R. J. Arnegard, "The multi-attribute task battery for human operator workload and strategic behavior research," NASA Langley Research Center, Hampton, Tech. Rep. NASA-TM-104174, 1992.

[45] A. Luximon and R. S. Goonetilleke, "Simplified subjective workload assessment technique," *Ergonomics*, vol. 44, no. 3, pp. 229–243, 2001.

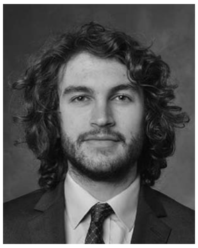
[46] P. S. Tsang and V. L. Velazquez, "Diagnosticity and multidimensional subjective workload ratings," *Ergonomics*, vol. 39, no. 3, pp. 358–381, 1996.

[47] M. A. Vidullch, G. F. Ward, and J. Schueren, "Using the subjective workload dominance (SWORD) technique for projective workload assessment," *Hum. Factors*, vol. 33, no. 6, pp. 677–691, 1991.

[48] M. L. Bolton, E. Biltekoff, J. Wei, and L. Humphrey, "On the level of measurement of subjective psychometric ratings," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2022, vol. 66, pp. 80–84.

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 an Associate Professor with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA. His research interests include 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, the M.S. degree in human factors engineering, in 2018 and 2020, respectively. He is currently working toward the Ph.D. degree with the Department of Industrial and Systems Engineering from the University at Buffalo, The State University of New York, Amherst, NY, USA.

His research interests include 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 Ohio State University, Columbus, OH, USA, in 2004, 2006, and 2009, respectively.

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