Estimation of Subjectively Reported Trust, Mental Workload, and Situation Awareness Using Unobtrusive Measures

Jacob R. Kintz, Neil T. Banerjee, Johnny Y. Zhang, Allison P. Anderson and Torin K. Clark, Smead Department of Aerospace Engineering Sciences, University of Colorado–Boulder, Boulder, Colorado, USA

Objective: We use a set of unobtrusive measures to estimate subjectively reported trust, mental workload, and situation awareness (henceforth "TWSA").

Background: Subjective questionnaires are commonly used to assess human cognitive states. However, they are obtrusive and usually impractical to administer during operations. Measures derived from actions operators take while working (which we call "embedded measures") have been proposed as an unobtrusive way to obtain TWSA estimates. Embedded measures have not been systematically investigated for each of TWSA, which prevents their operational utility.

Methods: Fifteen participants completed twelve trials of spaceflight-relevant tasks while using a simulated autonomous system. Embedded measures of TWSA were obtained during each trial and participants completed TWSA questionnaires after each trial. Statistical models incorporating our embedded measures were fit with various formulations, interaction effects, and levels of personalization to understand their benefits and improve model accuracy.

Results: The stepwise algorithm for building statistical models usually included embedded measures, which frequently corresponded to an intuitive increase or decrease in reported TWSA. Embedded measures alone could not accurately capture an operator's cognitive state, but combining the measures with readily observable task information or information about participants' backgrounds enabled the models to achieve good descriptive fit and accurate prediction of TWSA.

Conclusion: Statistical models leveraging embedded measures of TWSA can be used to accurately estimate responses on subjective questionnaires that measure TWSA.

Application: Our systematic approach to investigating embedded measures and fitting models allows for cognitive state estimation without disrupting tasks when administering questionnaires would be impractical.

Keywords: long-term missions, human-systems integration, adaptive automation, pilot, crew behavior, compliance, and reliance

Address correspondence to Jacob R. Kintz, Bioastronautics Laboratory, University of Colorado–Boulder, 3775 Discovery Dr AERO 340, Boulder, CO, 80303, USA; e-mail: jacob.kintz@colorado.edu

HUMAN FACTORS
2023, Vol. 65(6) 1142–1160
DOI:10.1177/00187208221129371
Article reuse guidelines: sagepub.com/journals-permissions

BACKGROUND AND MOTIVATION

Human operators' cognitive states (such as trust, mental workload, and situation awareness—TWSA) change as they work with autonomous systems due to many factors (Parasuraman et al., 2008; Yang et al., 2021). Here, we use "workload" to refer only to mental workload, as opposed to physical workload. Adaptive autonomous systems, which can change their own behavior in response to an operator or an environment, have been proposed as self-reliant teammates for working with humans (Anderson et al., 2020; Feigh et al., 2012). Providing adaptive autonomous systems with information about their human teammates' cognitive states in real-time remains a gap and active area of research (Feigh et al., 2012; Schwarz & Fuchs, 2018). Furthermore, keeping TWSA at ideal levels is of particular interest in performance-critical and safety-critical settings such as aerospace environments (Parasuraman et al., 2008; Stanton et al., 2001). As an example of an ideal level of TWSA, "calibrated trust" prevents both disuse and over-use of an autonomous system (Dzindolet et al., 2002; Parasuraman & Riley, 1997). Ideal workload prevents the operator from being underloaded which can cause disengagement, vigilance errors, and potentially decreased performance as discussed by Young & Stanton (2002). Ideal workload also prevents overload which can cause fatigue, mistakes, and decreased performance (Hancock & Matthews, 2018; Van Acker et al., 2018; Yerkes & Dodson, 1908). Sufficient SA is needed to complete a task with adequate safety and performance (Endsley, 1988b). Given this interest in providing autonomous systems with estimates of operator TWSA and maintaining ideal TWSA, it is critical to develop accurate methods for determining an operator's TWSA.

Subjective questionnaires are commonly used in experiments to measure TWSA. Questionnaires may be considered a "gold standard" for TWSA measurement as they have strong face validity for querying the cognitive states of an operator. Jian et al. (2000) developed a widely used subjective questionnaire to measure human trust in automated systems (Craig et al., 2019; Ghazali et al., 2018; Gombolay et al., 2018; Gutzwiller et al., 2019; Jian et al., 2000; Sanders et al., 2019; Spain et al., 2008; You & Robert Jr., 2018). The Bedford workload scale (Roscoe, 1979, 1984; Roscoe & Ellis, 1990) is a 10-point scale known to be specific and sensitive in measuring subjective workload (Heard et al., 2018). The Situation Awareness Rating Technique (SART) (Selcon & Taylor, 1990; Taylor, 1990) has been used as a measure of SA when other measures requiring system freezes were not feasible (Lin & Lu, 2017; Liu et al., 2014a, 2014b; Petersen et al., 2019). Subjective questionnaires are accepted as effective approaches to measuring TWSA, but they are obtrusive and require that tasks be paused or completed. They also only provide a single estimate of cognitive state, which prevents their use in providing continuous real-time estimates.

To address these limitations, previous work has put forth what we refer to as "embedded measures" of TWSA. Embedded measures are derived from natural actions that operators take while performing a task. Embedded measures do not disrupt tasks and do not require operators to divert attention as subjective questionnaires do. We define embedded measures as distinct from measures based on observable information alone (e.g., an operator's gaze) (de Winter et al., 2019; Kok & Soh, 2020), since embedded measures relate directly to actions operators are already taking as part of task completion. They are also distinct from performance measures (e.g., how closely operators track a target trajectory) in that they do not require knowledge of task goals or performance criteria to provide useful estimates of TWSA. Examples of embedded measures for trust under this definition include the time participants wait before taking over control from an autonomous system (Kunze et al., 2019; Petersen et al., 2019) and actions participants take to check or override advice from an autonomous system (Akash et al., 2018, 2020; Wickens et al., 2020; Yang et al., 2017).

Workload and spare mental capacity have been assessed using secondary tasks (Casali & Wierwille, 1983; Heard et al., 2018; Hicks & Wierwille, 1979; Knowles, 1963; Wierwille & Eggemeier, 1993; Young et al., 2015). Research on the degree to which secondary task completion or performance correlate with subjectively reported workload is varied, in some instances being correlated (Besson, Dousset, et al., 2012; Besson, Maïano, et al., 2012), while not in others (Hancock et al., 1990). We note that an unrelated secondary task (e.g., performing mental arithmetic) or one requiring known performance goals would not qualify as an embedded measure under our definition. However, operator engagement on a required low-priority subtask can serve as an embedded measure of workload. Similarly, verbal callouts in which operators report vehicle state changes were introduced as an embedded measure of SA in previous studies (Hainley et al., 2013; Karasinski et al., 2016, 2017). While this measure likely only evaluates Level 1 SA (Endsley, 1995), in some environments, callouts integrate naturally into existing tasks or are already standard tasks that operators complete (e.g., aircraft cockpits).

Despite the face validity of embedded measures, it remains unclear if they can be utilized to inform accurate estimates of TWSA. Previous studies using embedded measures of SA have employed them based upon their construct validity and have not assessed their accuracy in estimating cognitive states. Some work has investigated embedded measures of trust, but the basis for assessment was not a commonly used formal questionnaire (Xu & Dudek, 2015). For workload, others have noted that more evidence is needed to demonstrate the accuracy of task-based measures (Heard et al., 2018). Some literature has investigated measures that are sensitive to only *task load*, despite referring to them as measures of "workload" (Ding et al., 2020; Heard et al., 2020; Heard & Adams, 2019). Since workload is a function of other elements (e.g., environmental conditions, operator experience, strategy, fatigue) *in addition to task load* (Hooey et al., 2018), it is important to validate embedded

measures against workload questionnaires. Finally, none of the reviewed literature sought to implement embedded measures of TWSA simultaneously and compare them to commonly used questionnaires. Embedded measures' construct validity is accepted, but it is critical to determine their accuracy before they can be used to estimate cognitive states in real-time operations.

The goal of our experiment was to implement three embedded measures of TWSA into a spaceflight-relevant scenario and determine if those embedded measures could be used to inform estimates of operator TWSA, as determined by questionnaires. In our analysis, we aimed to address the relative benefit of increasingly personalized and complex models (e.g., only using an embedded measure vs. incorporating other observable information vs. more information about each individual participant).

METHODS

Scenario and Tasks

The scenario used in this experiment was designed to elicit a wide range of TWSA from each participant. It was presented in the Aerospace Research Simulator (ARES) at the University of Colorado–Boulder using Simulink (MATLAB version R2019b). Participants were given four tasks to complete during each trial of the experiment: a primary tracking task, a pushbutton lighting/response task, a verbal callout task, and a decision task aided by an autonomous system. Participants were seated in the ARES cockpit's left seat and used a joystick with their right hand and buttons on a throttle with their left hand, as shown in Figure 1.

The primary tracking task (upper middle screen, Figure 1) required participants to keep a virtual space station centered in the crosshairs of a "docking camera" as their spacecraft approached the station and experienced random perturbations. Participants were provided with "Offset" indicators for the up/down (Y) and left/right (X) directions they needed to null to track the station, using a joystick to make velocity command control inputs. All control inputs consumed "RCS (reaction control system) fuel" from a limited supply indicated on the primary display. For a given trial, one of three different levels of task load (low, medium, or high) was achieved through different gain settings for random perturbations. The magnitude and frequency of these perturbations were selected based on pilot testing conducted prior to the experiment; the perturbations selected induced a range of task loads for both participants unfamiliar with tracking tasks and those who had expert aircraft piloting backgrounds. The distance to the space station was indicated on the primary display and closed at a constant rate over 50 seconds.

The pushbutton lighting/response task (Hainley et al., 2013; Karasinski et al., 2016, 2017) was presented on a secondary display below and to the left of the primary display, at an angle of approximately 25° from the center of the primary display in the yaw plane and 21° from the center of the primary display in the pitch plane (see Figure 1). Participants were asked to press a corresponding green or blue button using their left hand when the outer ring of a "Data Link" light on the secondary display differed in hue (blue or green) from the inner circle of the light (see Figure 2). Participants were told to only monitor the Data Link light when not occupied with the primary task. We included a distractor "Docking Status" light on the secondary display to discourage participants from using their peripheral vision to respond to the Data Link light while still focusing on the primary task (which would have made the lighting/response task useless as a measure of spare mental capacity). Lighting events occurred unpredictably, two to seven seconds after the previous event either timed out or was acknowledged by the participant. Once lit, the light remained lit for ten seconds. If a participant acknowledged the light it was turned off, but if ten seconds passed with no acknowledgement then the light was turned off automatically.

The verbal callout task was modeled after recent experiments that used verbal callouts as an embedded measure of SA (Hainley et al., 2013; Karasinski et al., 2016, 2017). Participants were instructed to verbally report every 10% of RCS fuel consumption and every whole number value of distance to the space station crossed (e.g., 5, 4, 3, etc.) (see Figure 1). Distance decreased at a constant rate, but RCS fuel

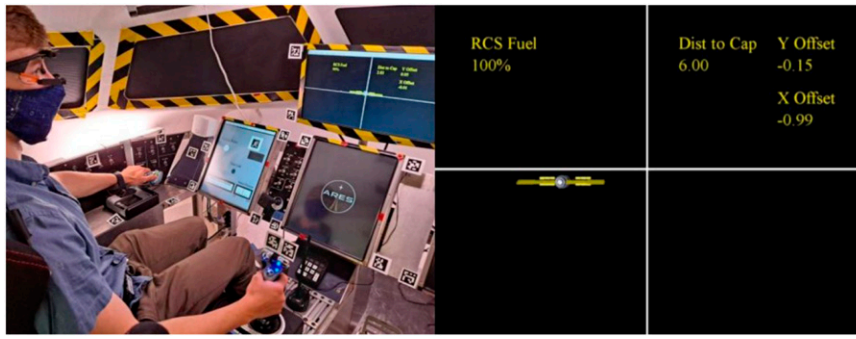

Figure 1. Experiment Environment and Primary Display. *Note.* Example participant seating, hand positioning, and displays during a trial are shown at left, and an enlarged example of the primary display is shown at right. On the primary display the virtual space station, camera crosshairs, offset indicators, reaction control system (RCS) fuel indicator, and distance ("Dist") to capture ("Cap") indicator can be seen. See Figure 2 for enlarged examples of the secondary display.

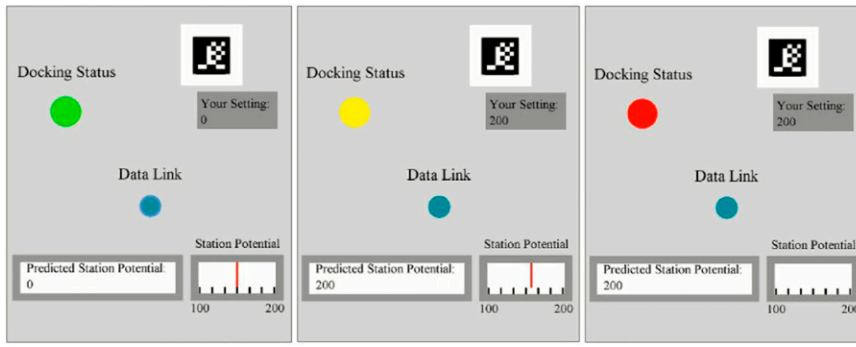

Figure 2. Progression of Secondary Display Information. *Note.* These images show the progression of the secondary display during the piloting phase (left), to the voltage setting phase (center), to after a voltage setting was confirmed (right). In the left panel the outer ring of the "Data Link" light differs slightly in hue from the center of the light. The "Docking Status" light was included only as a distractor.

consumption depended only upon participants' joystick inputs. These states were chosen for the callout task as they were critical to maintaining SA during the piloting phase. Importantly, participants were instructed to only make callouts when they were not occupied with any other tasks. Callouts were deemed successful if participants made them two seconds before or two seconds after the event occurred. Otherwise, callouts were considered missed. An experimenter marked callouts as they occurred, and callouts were verified post-experiment using audio recordings.

A 20 second "voltage setting phase" followed the 50 second piloting phase. Participants were asked to pick one of two voltage settings for their own spacecraft (either 100 V or 200 V) to match the voltage potential of the space station as their spacecraft docked. Participants were provided with two independent methods for identifying the voltage of the space station (Figure 2). The first was a discrete recommendation (either 100 V or 200 V) from an autonomous system based on its prediction of the station's voltage. Participants were told the autonomous system (1) used information they

could not access to make its recommendations, (2) was always trying to aid them (even when incorrect), and (3) was the same system across all trials. Pre-experiment training was provided to help participants understand how the autonomous system's recommendations worked. The second source of information about the station's voltage was an "analog" voltage gauge with limits of 100 V and 200 V and a starting needle position of 150 V. The dynamics of the analog gauge were simulated so that noise in the gauge diminished over time while the range of possible values increased over time, as the space station grew closer to the spacecraft. The final value the gauge settled on for each trial was selected from an equally spaced range of twelve values between 100 V and 200 V, so on a given trial the final value could have been obvious (e.g., 100 V) or much less obvious (e.g., 155 V). Once participants made and confirmed a voltage setting (either 100 V or 200 V) using buttons on the joystick base, the analog gauge needle disappeared. This was intended to prevent revealing more information about the autonomous system's reliability before participants saw a feedback screen summarizing the outcome of their decision. The analog gauge allowed participants to have some knowledge and environmental context to aid in making their decision and assessing the autonomous system's recommendation. This approach is more relevant to real-world scenarios than participants' blind reliance on an autonomous system's recommendation alone, as has been used in some previous experiments (Akash et al., 2018; Hu et al., 2016).

Measures of Trust, Workload, and SA

Participants completed the Trust in Automated Systems questionnaire developed by Jian et al. (Jian et al., 2000) after each trial to report their trust in the autonomous system. While biases have been identified with the questionnaire (Gutzwiller et al., 2019), it is one of the most commonly used measures of trust. We administered a modified Bedford workload scale to query participants' workload (Roscoe, 1979, 1984; Roscoe & Ellis, 1990). The modified scale employs slightly different graphics and language than the original Bedford workload scale but retains the same Cooper-Harper Rating Scale format as the original. Other multidimensional scales may capture different aspects of workload than the Bedford scale, which is unidimensional (Estes, 2015; Hancock & Matthews, 2018; Hart & Staveland, 1988). However, we chose a unidimensional scale for this work because it was quick to administer (helping to limit survey fatigue among participants) and because it provided intuitive descriptions of different workload levels for participants. The "14D" SART with visual-analog style ratings was presented to participants to assess SA (Selcon & Taylor, 1990; Taylor, 1990). While participants completed all 14 questions on the SART questionnaire, ultimately only the first 10 questions (the "10D" SART) were used in our analyses for consistency with how previous studies scored the SART (Petersen et al., 2019). The questionnaires used to assess TWSA and their scoring formulas can be found in the provided Supplementary Materials.

We defined three embedded measures (one for each of TWSA) that we sought to investigate with this experiment. While there may have been other embedded measures that could have been defined, these three were selected based on their use in previous literature. The trust embedded measure was the time in seconds participants took to confirm a voltage setting in the voltage setting phase. The workload embedded measure was the percentage of time the Data Link outer ring was lit out of the maximum possible time it could have been lit during the piloting phase. A lower percentage value meant participants quickly responded to lighting events, turning the outer ring of the light back to its original color. This metric is similar to the one originally proposed for lighting/response tasks (Knowles, 1963). The SA embedded measure was the percentage of successful callouts made from the total available callouts on each trial. The number of available callouts depended on the magnitude of control inputs made by the participant. The minimum number of available callouts for any participant was six and the maximum observed across all participants was fourteen (six distance callouts and eight fuel callouts).

To quantify the usefulness of the analog gauge information provided on each trial

(termed the "Expectation"), we first integrated the area under the curve of the analog gauge needle's position with respect to time until participants confirmed their setting, with the central 150 V mark as 0 for integration. We multiplied the magnitude of this value (which describes how strongly the gauge needle moved to one side of the gauge versus the other) by either a positive 1 when the system's recommendation was correct or a negative 1 when the system's recommendation was incorrect. This piece of observable information is not considered an embedded measure as it is not a direct measure of participant's actions or inactions during a task. Instead, it is additional information relating to the context of the task.

We also collected relevant background and demographic information from participants before the experiment as potential predictors of TWSA. Participants completed a questionnaire about their handedness, their experience with aerospace-relevant displays, their experience with robots, and their experience with navigational aids (e.g., Google Maps). Participants also completed a visual-analog scale version of the Automation Induced Complacency Potential (AICP) questionnaire regarding their attitudes towards automation (Merritt et al., 2019) and a simple five trial reaction test (Reaction Time Test, n.d.) similar to the Psychomotor Vigilance Test (PVT) for alertness (Basner et al., 2015). Participants were given five trials to practice the reaction test before completing five trials that were counted for average reaction time. The pre-experiment demographic questionnaire and AICP questionnaire can be found in the provided Supplementary Materials. AICP questionnaires were scored by inspection by experimenters; all other questionnaires were scored using image processing scripts written by experimenters in MATLAB (MATLAB version R2019b).

Experiment Protocol

This research complied with the American Psychological Association Code of Ethics and was approved by the University of Colorado–Boulder Institutional Review Board. Informed consent was obtained from each participant. We enrolled 15 participants (9 males, 6 females; ages 19 to 32, median age 24 years, 1 left-hand dominant) and all completed the full experiment (12 trials). Participants were aware of the high-level project goal from the informed consent but were left naïve to the exact manipulations and measurements being obtained. Participants were pre-screened for alcohol consumption in the six hours prior to their participation in our experiment, and for a known history of seizures before beginning training.

Participants were briefed and trained on the tasks to encourage a steady-state level of performance. To avoid giving participants insight into the autonomous system's reliability during training, participants never received feedback on if their voltage setting was correct or incorrect. Participants then donned psychophysiological sensors (eye tracking glasses, respiration monitor, electrocardiogram, and electrodermal activity), which were used for a different investigation and are not discussed in this work.

Participants experienced low task load, medium task load, and high task load trials four times each during the experiment, in a randomized order for each participant. The autonomous system provided "correct" advice on nine randomly ordered trials, while on three randomly ordered trials the advice was "incorrect" (i.e., a 75% reliability). This rate aimed to replicate realistic trust dynamics in a short experiment, with a system that was still useful despite its errors (Akash et al., 2020; Kantowitz et al., 1997; Lee & See, 2004; Nunnally, 1978; Petersen et al., 2019; Wickens et al., 2020; Wickens & Dixon, 2007). Out of 180 trials across all participants, 135 voltage settings were correctly recommended and selected, 35 were correctly selected despite an incorrect recommendation from the autonomous system, 10 were incorrectly selected based upon an incorrect recommendation, and participants never made an incorrect decision when given a correct recommendation.

Following each trial, participants completed digital copies of the previously described questionnaires. Participants were then provided with feedback on their performance during the previous trial. Waiting to provide participants with feedback regarding the previous trial until *after* they completed questionnaires prevented biasing of participants' subjective ratings. To

encourage engagement and give consequence to participants' decisions, bonus money was awarded for (1) tracking accurately during the tracking task, (2) choosing the correct voltage setting (either a reward for a correct decision or a penalty for an incorrect decision), and (3) making a quick decision for the voltage setting (even if the incorrect voltage setting was chosen). Participants were incentivized to make a quick decision for their voltage setting so that they had to rely on the autonomous system to some extent. Otherwise, participants could have always waited to decide until just before their spacecraft docked with the space station, obtaining the most accurate information from the analog gauge. Bonuses for a quick voltage setting decision were less than the bonuses/penalties for making a correct/incorrect decision. An example of the feedback screens presented to participants after each trial is shown in Figure 3. After participants finished all trials of the experiment, they completed a questionnaire about their sex, age, race, and ethnicity, which can be found in the provided Supplementary Materials. [insert Figure 3]

Statistical Analysis

We fit linear regression models for each cognitive state (of the general form $Y = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p + \gamma_1 X_1 X_2 + \ldots$) using data collected in our experiment. These models were used to estimate participants' TWSA as reported after each trial on questionnaires. We used fixed-effects, single level models for consistency across all model types and regression methods. We explored fitting mixed-effects models with random effects for participant-specific predictors, but accounting for random effects provided a negligible improvement in model descriptive fit and predictive accuracy. Mixed-effects predictor coefficients were also approximately the same as in the fixed-effects models (between 1 and 10%, much less than the standard error of the coefficients). We excluded a participant's data for a cognitive state if the range of their questionnaire responses was less than 10% of the total range of the questionnaire, since this low variation in questionnaire responses indicated either a misunderstanding of how to use the questionnaire or that our manipulations did not produce changes in TWSA for that participant. This criterion makes our approach less generalizable but was implemented in the interest of simplifying analysis. This criterion excluded two participants' trust data, while all workload and SA data were included.

Predictors and pairwise interactions included in trust and SA linear regression models were selected according to the Akaike Information Criterion (AIC) using a stepwise search and fit through MATLAB's "stepwiselm" function with all default settings other than the chosen criterion (MATLAB version R2019b). Workload data were exported from MATLAB and loaded in R (R Core Team, 2019). Predictors and pairwise interactions for the ordinal regression models of workload were then selected according to the AIC using a stepwise search through the "stepAIC" function with all default settings other than specifying "both" for search direction (Venables & Ripley, 2002). Finally, ordinal logistic cumulative link models were fit using the "clm" function with all default settings (Christensen, 2019). The modified Bedford workload scale used in this experiment is known to have unequal intervals between its levels, necessitating ordinal regression analysis methods (Casner & Gore, 2010). The most probable level determined by ordinal regression was used as a workload model's estimate for a given trial. Our data and analysis code can be found at https://osf.io/9xs3r/.

Accounting for differences in participants and relevant characteristics of an individual is critical in human factors research (Szalma, 2009), particularly when investigating cognitive states (which are "internal" to a person and cannot be observed directly). We anticipated it would be necessary to account for the role of personalization in predicting cognitive states. To do this, we conceived of five different model types (shown in Table 1, ordered from least to most personalized) to determine what predictors were available to the stepwise algorithm as it selected terms for each model. For each model type, data from all participants were used in one comprehensive fit (see Supplementary Materials tables). Only predictors with construct validity in predicting a given cognitive state were made available for each trust, workload, or SA model.

Figure 3. Example of Feedback Screens. *Note.* Participants only saw one feedback screen after each trial, depending on if they chose a correct voltage setting (left) or incorrect voltage setting (right). Payment details (as shown on the screens above) were displayed only after 15 seconds had passed since the success/failure feedback screen was presented.

Model descriptive fit was assessed by either adjusted R^2 (Ezekiel, 1930; Raju et al., 1997) for trust and SA, or Nagelkerke pseudo-R^2 (Nagelkerke, 1991) for workload. Adjusted R^2 was selected to evaluate descriptive fit as it is a commonly used metric that accounts for the number of predictors included in a model (Ezekiel, 1930; Raju et al., 1997). This was desirable as our models could have included many predictors which would have spuriously increased a non-adjusted metric of descriptive fit. While Nagelkerke pseudo-R^2 does not account for the number of predictors in a model and should not be compared to the adjusted R^2 calculated for the trust and SA models, it can still provide an indication of the descriptive fit of ordinal regression models (Nagelkerke, 1991). Nagelkerke pseudo-R^2 was also selected for its parallels with the Adjusted R^2 metric used for trust and SA linear regression models. Nagelkerke pseudo-R^2 was calculated using the "nagelkerke" function in R (Mangiafico, 2016).

Each model's predictive accuracy was evaluated with two different approaches for leaving out observations in cross-validation: one to evaluate how effective a given model was at predicting a completely unseen/new participant ("leave one participant out") and one to evaluate the same for a single trial/observation ("leave one observation out"). As model types 4 and 5 fit participant-specific parameters, they could not be evaluated with the former approach. We calculated Q^2 (Quan, 1988) and root mean square error (RMSE) to assess predictive accuracy for linear regression models of trust and SA. Q^2 was selected as it is analogous to R^2 in that it assesses predictive power as a proportional reduction of error. As such, Q^2 values near 0 correspond to no predictive power compared to taking the mean of the cross-validation data, while 1 corresponds to perfect predictions (Quan, 1988). For the ordinal regression models of workload, we calculated RMSE and three "Accuracy within N" metrics of predictive accuracy (e.g., accuracy within +/− 1 Bedford workload level = ACC1), where exactly correct predictions are labeled "ACC0" (Gaudette & Japkowicz, 2009).

RESULTS

A summary of available predictors, coefficients for the main effects of those predictors,

TABLE 1: Types of Statistical Models and Predictors Available for the Stepwise Search Algorithm

	Model type				
	1	2	3	4	5
Embedded measure	(Available)				
Observable info (e.g., task load)	(Not available)				
Demographic/ background info					
Participant-specific intercept					
Participant-specific predictor coefficients					
Purpose	How embedded measures have been used before	Requires no information about operator	Can account for operator traits before they do any tasks	Once fit to operator's experiment data, more specific	Once fit to operator's experiment data, most specific

Note: Predictors that were available for inclusion in each model type are ordered from least personalized/most general (top) to most personalized/least general (bottom). The last row of the table describes the purpose of each model type (columns) and each model type's level of personalization, which increases from left to right.

fit statistics, and predictive accuracy metrics for all model types are shown in Tables 2, 3, and 4 for trust, workload, and SA, respectively. Predictors are coded by their source: either embedded measures, observable aspects of the environment, additional information relating to performing a task, or information about a participant obtained before the experiment. If a predictor was not available to be included in a model due to its type, corresponding cells in the table are darkened. If a predictor was available to a given model, but was not selected by the stepwise algorithm, that cell contains only a dash. Model types 4 and 5 could include participant-specific intercepts, and model type 5 could include participant-specific coefficients as well. Median coefficient values are reported for participant-specific terms with a preceding "M." While at least one pairwise interaction term was included in each model type 2, 3, 4, and 5, interaction terms are not reported here for brevity. They can be found in the Supplementary Materials along with an example of a linear regression model formulation. However, the reported fit statistics and predictive accuracy metrics for the models *do include* the contributions of the models' pairwise interaction terms. Terms and coefficients for trust models can be found in Supplemental Tables 5-10, for workload models in Supplemental Tables 11-22, and for SA models in Supplemental Tables 23-30. Residual plots for trust and SA models can be found in Supplemental Figures 3 and 4, respectively.

DISCUSSION

This research explored a set of embedded measures to inform statistical models for

TABLE 2: Summary of Trust Models

Predictor	Type 1	Type 2	Type 3	Type 4	Type 5
Intercept	67***	73***	30***	M 64	M 58
Time to confirm voltage setting (0 - 20 seconds)	-2.1***	-3.2*	-1.6***	-1.9***	M -0.1
Number of times participant received ''wrong voltage setting'' feedback prior to trial (0 - 11)		13***	41*	3.5	—
Number of trials completed by participant prior to trial (0 - 11)		-2.0*	-2.6*	—	—
Expectation (-0.18 - 0.18, with negative values as dissonance and positive values as agreement)		38*	294***	81**	M 150
AICP monitoring score (1 - 25, with 1 being more likely to monitor automation)			2.6***		
"Robot/autonomous system user" demographic category (0 or 1)			—		
"Navigation aid user" demographic category (0 or 1)			—		
Adjusted R^2	0.08	0.30	0.65	0.69	0.78
Q^2 (leave out participant)	-0.02[a]	0.18	0.48		
Q^2 (leave out observation)	0.07	0.26	0.55	0.39	0.22
RMSE (leave out participant)	15.6	14.0	11.1		
RMSE (leave out observation)	14.3	12.7	10.0	11.5	13.0

Note: Horizontal stripes = embedded measure, grid checkering = observable aspect of environment, vertical stripes = additional information relating to performing a task, and diagonal stripes = participant info obtained before experiment. Median coefficient values for personalized predictors are noted with "M." Coefficient values not reported here can be found in the Supplementary Materials. If a predictor was not available to be included in a given model, corresponding table cells are empty. If a predictor was available to a given model, but was not selected by the stepwise algorithm, that cell has a dash.
[a]A negative Q^2 corresponds to a model that reduces predictive power assessed by cross-validation, as compared to the mean of the observations.
***: $p < 0.0005$, **: $p < 0.005$, *: $p < 0.05$.

accurately estimating human operator cognitive states (TWSA). The embedded measures we implemented were shown to be related to their relevant cognitive states, with effects in intuitive directions for each model type 1 (longer "Time to confirm voltage setting" corresponded to lower trust; higher "Data Link lighting %" with higher workload; higher "% of available callouts made successfully" with higher SA). However, we also found embedded measures alone were insufficient to accurately describe or predict TWSA (i.e., low R^2, low Q^2 or ACC, and higher RMSE for model type 1). Statistical models of cognitive states increased in predictive accuracy when additional predictors were included and improved in descriptive fit when personalized parameters were fit to each predictor. For all the cognitive states, the best descriptive fit was achieved by a model type 5, which is the most personalized of the model types. However, model type 5 did not achieve the best predictive accuracy for any of the cognitive states when

TABLE 3: Summary of Workload Models

Predictor		Type 1	Type 2	Type 3	Type 4	Type 5
Data link lighting % (0 - 100)		0.03***	0.07***	-0.24	0.07**	M 0.05
Task load setting (-1, 0, or 1)			1.5*	14***	4.1**	M 9.8
Number of trials completed by participant prior to trial (0 - 11)			0.14	-1.1	—	M -0.7
Summed magnitude of joystick control inputs (0 - 1500)			0.000008	-0.02	-0.003	0.002
Root mean square (RMS) of tracking error (min = 0.148, max = 1.68)			6.6	45**	6.6	M 15
"Video game use" demographic category (0, 1, 2, or 3)				-6.9*		
"Aerospace information display skill" demographic category (0, 1, 2, or 3)				3.3**		
Sleep rating (-2, -1, 0, 1, or 2)				4.2		
Sleep hours (rounded to half hour, min = 5, max = 9)				-8.8***		
Reaction test average score (min = 211 ms, max = 333 ms)				-0.04***		
Handedness (0 or 1)				5.6		
Nagelkerke Pseudo-R^2		0.21	0.57	0.83	0.76	0.92
ACC0 (leave out participant)		24%	27%	37%		

(Continued)

TABLE 3: (Continued)

Predictor	Type 1	Type 2	Type 3	Type 4	Type 5
ACC1 (leave out participant)	56%	71%	75%		
ACC2 (leave out participant)	82%	92%	93%		
ACC0 (leave out observation)	24%	32%	41%	38%	36%
ACC1 (leave out observation)	58%	71%	84%	82%	78%
ACC2 (leave out observation)	83%	92%	97%	96%	95%
RMSE (leave out participant)	1.96	1.49	1.38		
RMSE (leave out observation)	1.92	1.46	1.17	1.26	1.31

Note: Horizontal stripes = embedded measure, grid checkering = observable aspect of environment, vertical stripes = additional information relating to performing a task, and diagonal stripes = participant info obtained before experiment. Median coefficient values for personalized predictors are noted with "M." Coefficient values not reported here can be found in the Supplementary Materials. If a predictor was not available to be included in a given model, corresponding table cells are empty. If a predictor was available to a given model, but was not selected by the stepwise algorithm, that cell has a dash.
***: $p < 0.0005$, **: $p < 0.005$, *: $p < 0.05$.

TABLE 4: Summary of Situation Awareness Models

Predictor	Type 1	Type 2	Type 3	Type 4	Type 5
Intercept	18***	28***	6.2	M 21	M 20
% Of available callouts made successfully (0 - 100)	0.07***	—	—	0.05*	M 0.01
Task load setting (-1, 0, or 1)		-6.9***	-0.33	-3.3***	M -3.9
Number of trials completed by participant prior to trial (0 - 11)		0.23*	0.27**	0.58*	M 0.3
Summed magnitude of joystick control inputs (0 - 1500)		-0.004*	-0.009***	-0.01***	-0.008*
Root mean square (RMS) of tracking error (min = 0.148, max = 1.68)		-16***	-3.6	—	M -4.3
"Video game use" demographic category (0, 1, 2, or 3)			—		
"Aerospace information display skill" demographic category (0, 1, 2, or 3)			11**		
Sleep rating (-2, -1, 0, 1, or 2)			—		
Sleep hours (rounded to half hour, min = 5, max = 9)			—		
Reaction test average score (min = 211 ms, max = 333 ms)			0.09***		
Handedness (0 or 1)			-5.1**		
Adjusted R^2	0.06	0.54	0.65	0.72	0.81
Q^2 (leave out participant)	0.01	0.51	0.23		
Q^2 (leave out observation)	0.06	0.53	0.63	0.69	0.61
RMSE (leave out participant)	6.35	4.48	5.59		
RMSE (leave out observation)	6.13	4.32	3.86	3.53	3.95

Note: Horizontal stripes = embedded measure, grid checkering = observable aspect of environment, vertical stripes = additional information relating to performing a task, and diagonal stripes = participant info obtained before experiment. Median coefficient values for personalized predictors are noted with "M." Coefficient values not reported here can be found in the Supplementary Materials. If a predictor was not available to be included in a given model, corresponding table cells are empty. If a predictor was available to a given model, but was not selected by the stepwise algorithm, that cell has a dash. ***: $p < 0.0005$, **: $p < 0.005$, *: $p < 0.05$.

assessed using our cross-validation methods, possibly indicating overfitting for those highly personalized models. At least one predictor based on participants' demographics or predispositions was selected to predict each cognitive state when it was available to be included (e.g., participants with higher "Display skill rating" had higher SA).

Predictive accuracy metrics for trust models suggest the importance of considering participants' predispositions in addition to their evolving relationship with a specific system when estimating trust. This is reflected in the large increase in Q^2 and reduction in RMSE once demographic predictors were added to trust models. For example, Q^2 (observation) for the trust model type 2 was 0.26, but this value increased to 0.55 for model type 3, which includes demographics. This is also reflected in the relatively low trust model type 2 Q^2 values when compared to SA model type 2 Q^2 values (e.g., Q^2 (observation) was already 0.53 for SA model type 2, which lacked demographic predictors). Trust model types 4 and 5, which were the most personalized of the models, achieved the best descriptive fits of the trust models as measured by Adjusted R^2. Despite the challenge of accounting for participants' predispositions, the embedded measure for trust and other included predictors were useful in predicting participants' trust.

The workload model type 3 achieved a noteworthy 75% accuracy in predicting reported workload within one level on the Bedford scale for all twelve of an unseen participant's trials. This shows the workload model type 3, which could be fit in future experiments using demographic data before a participant completes any tasks, was quite accurate even though it was not a personalized model like model types 4 and 5. However, as previously noted, the intervals between different levels of the Bedford scale are not the same, and our accuracy metric does not account for the specific descriptors associated with each level of the scale (Casner & Gore, 2010). Additionally, the workload model type 1 had poor predictive accuracy and descriptive fit (e.g., Nagelkerke Pseudo-R^2 was 0.21). While the embedded measure of workload is arguably the embedded measure with the most precedence from previous literature (and was shown to be useful in all workload models), this poor accuracy shows the embedded measure still does not capture the "whole picture" of a participant's workload. Our results indicate more predictors and information were needed to accurately predict workload.

The verbal callout measure of SA was the embedded measure that was most frequently excluded from models by the stepwise algorithm, as only model types 1, 4, and 5 included the embedded measure of SA. However, when the measure was included in models, its coefficient had an intuitive positive sign (indicating that more successful callouts reflected a higher level of SA). Scores on the SART can range from a maximum of 46 to a minimum of −14 (Selcon & Taylor, 1990; Taylor, 1990). Thus, as an example, the SA model type 1 predicts that a participant who made 100% of their available callouts would score seven points higher on the SART than a participant who made none of the available callouts. The Q^2 values for SA models were generally high, over 0.5 for Q^2 (observation) for model types 2, 3, 4, and 5. This suggests that even if the embedded measure for SA needs improvement (i.e., it was not universally included by the stepwise algorithm), the other observable predictors used in SA models provided an accurate estimate of SA.

We are confident that participants prioritized the primary task above other tasks, and that participants completed questionnaires in a way that reflected their TWSA. Experiment data showed participants continuously made joystick inputs but sometimes missed Data Link lighting events or verbal callouts. The short durations of trials in this experiment mitigated the effect of participants forgetting or biasing their questionnaire responses regarding earlier parts of a trial, as can happen with SART ratings (Salmon et al., 2009).

Our use of the SART to measure SA was a limitation of our work. SART ratings are known to be confounded by participants' confidence, task performance, and workload; the SART is often not the best means of querying SA (Endsley et al., 1998; Lin and Lu, 2017). Other measurement methods without these limitations such as the SPAM (Durso et al., 1998) or the SAGAT (Endsley, 1988a, 1988b, 2021) were not used because it was desirable to measure SA without interrupting trials in this experiment. The SPAM and the SAGAT require question "probes" and task freezes, respectively, and individual trials were too short for such methods to be feasible. Furthermore, aspects of workload and SA are known to be correlated when measured by the SART (Selcon et al., 1991). Since the predictors used to estimate workload and SA were the same

except for the embedded measures, it is possible our predictions for workload and SA are correlated. Future work could investigate the specificity of cognitive state estimates produced from embedded measures or employ multivariate statistics to simultaneously estimate workload and SA using the same predictors.

Additional limitations of this study include our focus on one embedded measure per cognitive state (using multiple measures may improve cognitive state estimates), the use of a single subjective measure for each cognitive state (different questionnaires can produce different outcomes; Casner & Gore, 2010), and our sample size of only fifteen participants. Further, our multiple regression modeling approach assumes the TWSA outcomes on each trial are independent. However, cognitive states may have temporal dynamics, like an operator developing trust in an autonomous decision aid (Guo & Yang, 2020; Yang et al., 2021), that violate this assumption. For this reason, we do not emphasize the statistical significance of coefficients included in our modeling results. Finally, the embedded measures implemented in this work are specific to the tasks in our experiment and would not apply to different tasks in another experiment. Yet, comparable measures (such as response times, actions taken, or verbal callouts of relevant environment states) likely exist for other human operator tasks and could be implemented in a manner similar to our experiment. Future work should take our results showing the importance of task information and participants' predispositions into consideration when defining unobtrusive measures of cognitive states.

CONCLUSION

This work represents the first study to rigorously investigate a set of embedded measures of trust, workload, and SA by comparing those measures to commonly used subjective questionnaires. We also implemented three embedded measures (one each for TWSA) in a spaceflight-relevant scenario; previous studies had only implemented one or two embedded measures (Akash et al., 2018, 2020; Hainley et al., 2013; Karasinski et al., 2016, 2017; Petersen et al., 2019). Our results are consistent with studies showing the general utility of embedded measures of cognitive states (Hainley et al., 2013; Karasinski et al., 2016, 2017; Kunze et al., 2019; Petersen et al., 2019). We improved upon those previous studies by collecting participants' responses on subjective questionnaires and then using cross-validation methods to assess embedded measures' predictive capabilities, developing a systematic approach to creating models that can estimate TWSA. Critically, we found embedded measures should not be equated with TWSA (as was assumed in several previous studies) because our embedded measures were inaccurate in capturing cognitive states when used on their own. However, when embedded measures were combined with other readily observed data (e.g., an individual's reported predisposition towards automation), they could estimate TWSA unobtrusively and accurately. Accurate cognitive state estimates enable adaptive autonomous systems (systems that can change their mode of operation or level of transparency to maintain ideal levels of operator cognitive states) in human-autonomy teams.

ACKNOWLEDGMENTS

This material is based upon work supported by NASA under grant or cooperative agreement award number 80NSSC19K1052. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Aeronautics and Space Administration (NASA). We appreciate the input we received from our reviewers and the help of Joshua Seedorf and Carlos Pinedo in the development of our scenario. Thank you to Evelyn Clarke for her assistance with data analysis and discussions, and to the "Habitats Optimized for Missions of Exploration" (HOME) research team for ideas that guided our experiment. We would also like to thank everyone who participated in our experiment.

KEY POINTS

- A set of unobtrusive measures of trust, mental workload, and situation awareness were implemented into a spaceflight-relevant scenario

- The measures were compared to commonly used subjective questionnaires using statistical models created with varied degrees of personalization
- Statistical models combining the measures with observable task information could accurately predict operators' trust, mental workload, and situation awareness as indicated on the subjective questionnaires
- Unobtrusive, accurate cognitive state estimation enables autonomous systems that adapt according to operator cognitive states in human-autonomy teams

ORCID iDs

Jacob R. Kintz https://orcid.org/0000-0001-9444-7409

Neil T. Banerjee https://orcid.org/0000-0002-6825-917X

Allison P. Anderson https://orcid.org/0000-0001-7808-8557

Torin K. Clark https://orcid.org/0000-0002-9345-9712

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

Anderson, A. P., Clark, T. K., & Kong, Z. (2020). Adaptive Autonomy for Future Spacecraft Habitats. *Human Robot Interaction for Space Robotics*, *4*. https://doi.org/10.1109/ICHMS53169.2021.9582622

Akash, K., Hu, W.-L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems*, *8*(4), 1–20. https://doi.org/10.1145/3132743

Akash, K., McMahon, G., Reid, T., & Jain, N. (2020). Human trust-based feedback control: dynamically varying automation transparency to optimize human-machine interactions. *IEEE Control Systems Magazine*, *40*(6), 98–116. https://doi.org/10.1109/MCS.2020.3019151

Basner, M., Savitt, A., Moore, T. M., Port, A. M., McGuire, S., Ecker, A. J., Nasrini, J., Mollicone, D. J., Mott, C. M., McCann, T., Dinges, D. F., & Gur, R. C. (2015). Development and validation of the cognition test battery for spaceflight. *Aerospace Medicine and Human Performance*, *86*(11), 942–952. https://doi.org/10.3357/AMHP.4343.2015

Besson, P., Dousset, E., Bourdin, C., Bringoux, L., Marqueste, T., Mestre, D. R., & Vercher, J. L. (2012). Bayesian network classifiers inferring workload from physiological features: Compared performance. *2012 IEEE Intelligent Vehicles Symposium*, *1*(1), 282–287. https://doi.org/10.1109/IVS.2012.6232134

Besson, P., Maïano, C., Bringoux, L., Marqueste, T., Mestre, D. R., Bourdin, C., Dousset, E., Durand, M., & Vercher, J. (2012). Cognitive workload and affective state: A computational study using Bayesian networks. *2012 6th IEEE International Conference Intelligent Systems*, *1*(1), 140–145. https://doi.org/10.1109/IS.2012.6335127

Casali, J. G., & Wierwille, W. W. (1983). A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load. *Human Factors*, *25*(6), 623–641. https://doi.org/10.1177/001872088302500602

Casner, S. M., & Gore, B. F. (2010). *Measuring and evaluating workload: A primer*. NASA Technical Memorandum, 216395, 2010.

Christensen, R. H. B. (2019). *ordinal—Regression models for ordinal data*.

Craig, S. D., Chiou, E. K., & Schroeder, N. L. (2019). The impact of virtual human voice on learner trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 2272–2276. https://doi.org/10.1177/1071181319631517

de Winter, J. C. F., Eisma, Y. B., Cabrall, C. D. D., Hancock, P. A., & Stanton, N. A. (2019). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology & Work*, *21*(1), 99–111. https://doi.org/10.1007/s10111-018-0527-6

Ding, Y., Cao, Y., Duffy, V. G., Wang, Y., & Xuefeng, Z. (2020). Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning. *Ergonomics*, *1–32*, 896–908, https://doi.org/10.1080/00140139.2020.1759699

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly*, *6*(1), 1–20. https://doi.org/10.2514/atcq.6.1.1

Dzindolet, M., Pierce, L., Beck, H., & Dawe, L. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*, 79–94. https://doi.org/10.1518/0018720024494856

Endsley, M. R. (1988a). Situation awareness global assessment technique (SAGAT). *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, Dayton, OH, USA, 23–27 May 1988, pp. 789–795 vol.3. https://doi.org/10.1109/NAECON.1988.195097

Endsley, M. R. (1988b). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society Annual Meeting*, *32*(2), 97–101. https://doi.org/10.1177/154193128803200221

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*(1), 65–84. https://doi.org/10.1518/001872095779049499

Endsley, M. R. (2021). A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of SAGAT and SPAM. *Human Factors*, *63*(1), 124–150. https://doi.org/10.1177/0018720819875376

Estes, S. (2015). The workload curve: Subjective mental workload. *Human Factors*, *57*(7), 1174–1187. https://doi.org/10.1177/0018720815592752

Ezekiel, M. (1930). *Methods of correlation analysis* (pp. xiv, 427). Wiley.

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems: a framework for researchers and system designers. *Human Factors*, *54*(6), 1008–1024. https://doi.org/10.1177/0018720812443983

Gaudette, L., & Japkowicz, N. (2009). Evaluation methods for ordinal classification. In Y. Gao & N. Japkowicz (Eds.), *Advances in Artificial Intelligence* (pp. 207–210). Springer. https://doi.org/10.1007/978-3-642-01818-3_25

Ghazali, A. S., Ham, J., Barakova, E. I., & Markopoulos, P. (2018). Effects of robot facial characteristics and gender in persuasive human-robot interaction. *Frontiers in Robotics and AI*, *5*. 73, https://doi.org/10.3389/frobt.2018.00073

Gombolay, M., Yang, X. J., Hayes, B., Seo, N., Liu, Z., Wadhwania, S., Yu, T., Shah, N., Golen, T., & Shah, J. (2018). Robotic assistance in the coordination of patient care. *The International Journal of Robotics Research*, *37*(10), 1300–1316. https://doi.org/10.1177/0278364918778344

Guo, Y., & Yang, X. J. (2020). Modeling and predicting trust dynamics in human–robot teaming: A Bayesian inference approach. *International Journal of Social Robotics*, *13*(8), 1899–1909. https://doi.org/10.1007/s12369-020-00703-3

Gutzwiller, R. S., Chiou, E. K., Craig, S. D., Lewis, C. M., Lematta, G. J., & Hsiung, C.-P. (2019). Positive bias in the 'Trust in Automated Systems Survey'? An examination of the Jian et al. (2000) scale. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 217–221. https://doi.org/10.1177/1071181319631201

Hainley, C. J., Duda, K. R., Oman, C. M., & Natapoff, A. (2013). Pilot performance, workload, and situation awareness during lunar landing mode transitions. *Journal of Spacecraft and Rockets*, *50*(4), 793–801. https://doi.org/10.2514/1.A32267

Hancock, P. A., & Matthews, G. (2018). *Workload and performance: Associations, insensitivities, and dissociations:* Human Factors. https://doi.org/10.1177/0018720818809590

Hancock, P. A., Wulf, G., Thom, D., & Fassnacht, P. (1990). Driver workload during differing driving maneuvers. *Accident; Analysis and Prevention*, *22*(3), 281–290. https://doi.org/10.1016/0001-4575(90)90019-h

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Heard, J., & Adams, J. A. (2019). Multi-dimensional human workload assessment for supervisory human–machine teams. *Journal of Cognitive Engineering and Decision Making*, *13*(3), 146–170. https://doi.org/10.1177/1555343419847906

Heard, J., Fortune, J., & Adams, J. A. (2020). SAHRTA: A supervisory-based adaptive human-robot teaming architecture. *2020 IEEE Conference on Cognitive and Computational Aspects of Situation. Management (CogSIMA)*, 1–8. https://doi.org/10.1109/CogSIMA49017.2020.9215996

Heard, J., Harriott, C. E., & Adams, J. A. (2018). A survey of workload assessment algorithms. *IEEE Transactions on Human-Machine Systems*, *48*(5), 434–451. https://doi.org/10.1109/THMS.2017.2782483

Hicks, T. G., & Wierwille, W. W. (1979). Comparison of five mental workload assessment procedures in a moving-base driving simulator. *Human Factors*, *21*(2), 129–143. https://doi.org/10.1177/001872087902100201

Hooey, B. L., Kaber, D. B., Adams, J. A., Fong, T. W., & Gore, B. F. (2018). The underpinnings of workload in unmanned vehicle systems. *IEEE Transactions on Human-Machine Systems*, *48*(5), 452–467. https://doi.org/10.1109/THMS.2017.2759758

Hu, W.-L., Akash, K., Jain, N., & Reid, T. (2016). Real-time sensing of trust in human-machine interactions. *IFAC-PapersOnLine*, *49*(32), 48-53. https://doi.org/10.1016/j.ifacol.2016.12.188

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Kantowitz, B. H., Hanowski, R. J., & Kantowitz, S. C. (1997). Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. *Human Factors*, *39*(2), 164–176. https://doi.org/10.1518/001872097778543831

Karasinski, J. A., Robinson, S. K., Duda, K. R., & Prasov, Z. (2016). Development of real-time performance metrics for manually-guided spacecraft operations. *2016 IEEE Aerospace Conference*, 1–9. https://doi.org/10.1109/AERO.2016.7500734

Karasinski, J. A., Robinson, S. K., Handley, P., & Duda, K. R. (2017). Real-time performance feedback in a manually-controlled spacecraft inspection task. In *AIAA Modeling and Simulation Technologies Conference*. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2017-1314

Knowles, W. B. (1963). Operator loading tasks. *Human Factors*, *5*(2), 155–161. https://doi.org/10.1177/001872086300500206

Kok, B. C., & Soh, H. (2020). Trust in robots: Challenges and opportunities. *Current Robotics Reports*, *1*(4), 297–309. https://doi.org/10.1007/s43154-020-00029-y

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: Implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, *62*(3), 345–360. https://doi.org/10.1080/00140139.2018.1547842

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Lin, L.-W., & Lu, M.-S. (2017). Empirical research regarding effects of pilot age and flight hours on pilot workload and situation awareness. *Journal of Aeronautics, Astronautics and Aviation*, *49*(1), 31–47. https://doi.org/10.6125/16-1114-910

Liu, S., Wanyan, X., & Zhuang, D. (2014a). Modeling the situation awareness by the analysis of cognitive process. *Bio-Medical Materials and Engineering*, *24*(6), 2311–2318. https://doi.org/10.3233/BME-141044

Liu, S., Wanyan, X., & Zhuang, D. (2014b). A quantitative situational awareness model of pilot. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, *3*(1), 117–122. https://doi.org/10.1177/2327857914031019

Mangiafico, S. S. (2016). *Summary and analysis of extension program evaluation in R*. Rutgers Cooperative Extension: New Brunswick, NJ, USA, 125, 16–22.

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., & Shirase, L. (2019). Automation-induced complacency potential: Development and validation of a new scale. *Frontiers in Psychology*, *10*. 225, https://doi.org/10.3389/fpsyg.2019.00225

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692. https://doi.org/10.1093/biomet/78.3.691

Nunnally, J. C. (1978). *Psychometric Theory/Jum C. Nunnaly.*

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation:

Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 140–160. https://doi.org/10.1518/155534308X284417

Petersen, L., Robert, L., Yang, X. J., & Tilbury, D. (2019). Situational awareness, driver's trust in automated driving systems and secondary task performance. *SAE International Journal of Connected and Automated Vehicles*, *2*(2), 12. https://doi.org/10.4271/12-02-02-0009

Quan, N. T. (1988). The prediction sum of squares as a general measure for regression diagnostics. *Journal of Business & Economic Statistics*, *6*(4), 501–504. https://doi.org/10.2307/1391469

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1997). Methodology review: Estimation of population validity and cross-validity, and the use of equal weights in prediction. *Applied Psychological Measurement*, *21*(4), 291–305. https://doi.org/10.1177/01466216970214001

Reaction Time Test. (n.d.). Human Benchmark. https://www.humanbenchmark.com/tests/reactiontime.

Roscoe, A. H. (1979). Handling qualities, workload and heart rate. In *AGARDograph No. 246; Survey of Methods to Assess Workload*. AGARD.

Roscoe, A. H. (1984). Assessing pilot workload in flight. Flight test techniques (AGARD-CP373). In *Proceedings of NATO Advisory Group for Aerospace Research and Development (AGARD)*. AGARD.

Roscoe, A. H., & Ellis, G. A. (1990). *A subjective rating scale for assessing pilot workload in flight: A decade of practical use*. Royal Aerospace Establishment Farnborough (United Kingdom). https://apps.dtic.mil/sti/citations/ADA227864

Sanders, T., Kaplan, A., Koch, R., Schwartz, M., & Hancock, P. A. (2019). The relationship between trust and use choice in human-robot interaction. *Human Factors*, *61*(4), 614–626. https://doi.org/10.1177/0018720818816838

Schwarz, J., & Fuchs, S. (2018). Validating a "Real-Time Assessment of Multidimensional User State" (RASMUS) for adaptive human-computer interaction. 2018 *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 704–709. https://doi.org/10.1109/SMC.2018.00128

Selcon, S. J., & Taylor, R. (1990). Evaluation of the Situational Awareness Rating Technique (SART) as a tool for aircrew systems design. *Situational Awareness in Aerospace Operations (AGARD-CP-478)*, *478*, 1–8.

Selcon, S. J., Taylor, R. M., & Koritsas, E. (1991). Workload or situational awareness?: TLX vs. SART for aerospace systems design evaluation. *Proceedings of the Human Factors Society Annual Meeting*, *35*(2), 62–66. https://doi.org/10.1518/107118191786755706

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). *Towards an empirically developed scale for system trust: Take two: Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. https://doi.org/10.1177/154193120805201907

Stanton, N. A., Chambers, P. R. G., & Piggott, J. (2001). Situational awareness and safety. *Safety Science*, *39*(3), 189–204. https://doi.org/10.1016/S0925-7535(01)00010-8

Szalma, J. L. (2009). Individual differences in human–technology interaction: Incorporating variation in human characteristics into human factors and ergonomics research and design. *Theoretical Issues in Ergonomics Science*, *10*(5), 381–397. https://doi.org/10.1080/14639220902893613

Taylor, R. M. (1990). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. *Situational Awareness in Aerospace Operations (AGARD-CP-478)*, *478*, 1-17.

Van Acker, B. B., Parmentier, D. D., Vlerick, P., & Saldien, J. (2018). Understanding mental workload: From a clarifying concept analysis toward an implementable framework. *Cognition, Technology & Work*, *20*(3), 351–365. https://doi.org/10.1007/s10111-018-0481-3

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S (Fourth)*. Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212. https://doi.org/10.1080/14639220500370105

Wickens, C. D., Fitzgerald, N. J., Clegg, B. A., Smith, C. A. P., Orth, D., & Kincaid, K. (2020). Decision aiding for nautical collision avoidance: Trust, dependence, and implicit understanding of the decision algorithm. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 1950–1954. https://doi.org/10.1177/1071181320641470

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, *35*(2), 263–281. https://doi.org/10.1177/001872089303500205

Xu, A., & Dudek, G. (2015). OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 221-228. https://doi.org/10.1145/2696454.2696492.

Yang, X. J., Schemanske, C., & Searle, C. (2021). *Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation*. Human Factors, 00187208211034716. https://doi.org/10.1177/00187208211034716

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, *1*(1), 408–416. https://doi.org/10.1145/2909824.3020230

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit rormation. *Journal of Comparative Neurology & Psychology*, *18*(1), 459–482. https://doi.org/10.1002/cne.920180503

You, S., & Robert, L. P., Jr. (2018). Human-robot similarity and willingness to work with a robotic co-worker. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 251–260. https://doi.org/10.1145/3171221.3171281

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17. https://doi.org/10.1080/00140139.2014.956151

Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human Factors*, *44*(3), 365–375. https://doi.org/10.1518/0018720024497709

Jacob R. Kintz is pursuing a doctoral degree in Bioastronautics at the University of Colorado–Boulder. He received a master's degree in Aerospace

Engineering Sciences from the University of Colorado–Boulder, 2021.

Neil T. Banerjee received master's degrees in Aerospace Engineering Sciences and Engineering Management from the University of Colorado–Boulder, 2021.

Johnny Y. Zhang received a master's degree in Aerospace Engineering Sciences from the University of Colorado–Boulder, 2021.

Allison P. Anderson is an Assistant Professor at the University of Colorado–Boulder. She received her PhD in Aerospace Biomedical Engineering from the Massachusetts Institute of Technology, 2014.

Torin K. Clark is an Assistant Professor at the University of Colorado–Boulder. He received his PhD in Humans in Aerospace Engineering from the Massachusetts Institute of Technology, 2013.