Multimodal Feature Selection to Unobtrusively Model Trust, Workload, and Situation Awareness

by

Savannah L. Buchner

B.S., University of California Davis, 2019

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science

Department of Aerospace Engineering Sciences

2022

Committee Members:

Allison Anderson, Chair

Torin Clark

David Klaus

Buchner, Savannah L. (M.S., Aerospace Engineering Sciences)

Multimodal Feature Selection to Unobtrusively Model Trust, Workload, and Situation Awareness

Thesis directed by Assistant Professor Allison Anderson

Effective human-autonomy teaming is an increasingly important challenge to ensure mission success in operational environments, such as future deep space missions. Modeling an operator's cognitive state, such as trust, situation awareness, and mental workload, can improve performance by informing the autonomous system about the humans that they interact with. Subjective questionnaires are often used to measure these states; however, they are obtrusive and impractical for operational contexts. The use of embedded measures (e.g., measures from actions that naturally occur while completing a task) and physiological measures (e.g., heart rate monitoring, respiration, or eye-based measures) may be an effective way to unobtrusively understand cognitive states. However, these signals may be ambiguous as they can be tied to multiple cognitive states simultaneously.

Models were developed to estimate trust from a variety of predictor variables, as well as models to simultaneously estimate situation awareness and workload. In order to generate these models, data from 15 subjects was collected during a spacecraft piloting and docking simulation. A LASSO-based algorithm was developed to select the important features for these models. From these features, multivariate regression models were generated, and predictive capabilities were assessed through cross validation.

Having the use of multiple categories of features (e.g., embedded measures and physiological measures) allowed for increase model performance and a viable way to estimate cognitive states over models without multiple categories available. Additionally, simultaneously fitting situation awareness and workload generated more parsimonious and operationally feasible models than models that were fit to a single cognitive state. The developed algorithm, use of multiple features, and understanding of importance of simultaneously fitting can be used for generating models for future

human-autonomy teaming research.

Acknowledgements

I am very grateful to my advisor Dr. Allie Anderson for this opportunity and all her help and guidance along the way. Thank you also to Dr. Torin Clark and Jacob Kintz for the advice and help, this project would not have been possible without them.

I'd also like to thank Jacob, Johnny Zhang, and Neil Banerjee for designing the experiment and processing data, as well as Josh Seedorf and Evie Clarke for their help running the experiment. Thank you to the many Bioastronautics students who provide insights and advice, as well as Dr. David Klaus for serving on my thesis committee. Finally thank you to my friends and family.

Contents

Chapter

Appendix

Table

Figure

Chapter 1: Introduction

1.1 Background and Motivation

Effective human-autonomy teaming is an increasingly important challenge to ensure mission success in operational environments, such as future deep space missions. It is desirable for autonomy to have information about the cognitive states of its operators, including trust, situation awareness (SA), and mental workload (WL) to enable adaptive autonomous systems, which are systems that can change their mode of operation or level of transparency to maintain ideal levels of cognitive states [1]. Trust is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [2]. SA is "the perception of critical elements in the environment, the comprehension of their meaning, and the projection of their status into the future" [3]. Finally, WL refers to "the perceived relationship between the amount of mental processing capability or resources and the amount required by the task" [4]. While there are different kinds of workload, this research focuses only on mental workload. This perceived workload is a result of many factors, such as task load, required performance, operator skills, strategy, fatigue, and environmental conditions [5].

Having an operator working at outside of the optimal levels of trust, SA, or WL can lead to mission failure due to issues like over-reliance on autonomous systems, fatigue, and vigilance errors [6] [7] [8]. Thus, to achieve adaptive autonomy that can maintain optimal cognitive state levels, a means by which to unobtrusively estimate these cognitive states is needed to close the loop on human-autonomy teaming by allowing autonomous systems to adapt to or alter the cognitive states of their human operators. This thesis focuses on building models that can estimate trust, SA, and WL using various unobtrusive measures.

Cognitive states can be assessed through subjective, physiological, or embedded measures [9]. Subjective surveys, while imperfect, represent a gold-standard for cognitive state estimation. However, surveys are obtrusive and cannot be used in real-time as they require either pausing a task or waiting for task completion to administer. In addition, humans may not always be accurate

in judgements of their own cognitive states and survey responses can be influenced from user's characteristics like limits of memory, which introduces sources of bias [10–12]. For future operational use, unobtrusive measures will be required in order to understand an operator's cognitive state as it changes over time, as a function of both internal and external factors [13].

Physiological signals and embedded measures may be able to effectively be used as features in a model to estimate someone's cognitive state. Psychophysiology is the study of the relationship between physiological signals, such as those based on electrocardiogram (ECG), electrodermal activity (EDA), respiration (RSP), and eye-tracking, to mental processes like cognitive states. Physiological measures have some important advantages over subjective surveys including objectivity, unobtrusiveness, implicitness, continuity, and responsiveness [12]. Physiological signals can be measured in real-time in a continuous manner, and do not rely on a user's perception of their own cognitive states. Additionally, they are unobtrusive in the sense that they do not directly interfere with the core objective of a person, such as monitoring spacecraft docking, like surveys would. However, they can be obtrusive in the sense that may be uncomfortable to wear. In addition, physiological sensors create data acquisition and interpretation issues due to the low signal levels and ability to be confounded with things like body movement, interference from the power grid, or environmental factors [14].

Assessment of cognitive states using physiological signals is still an open area of research. Most of the research has been focused on a single cognitive state and different studies often report different correlations between physiological signals and states [9, 14–22]. This research aims to correlate multiple types of physiological signals from the same data set to different cognitive states simultaneously. This will help increase understanding of which signals are closely tied to, or even possibly specific to each cognitive states; prior efforts may have confounded results as each state was not estimated as a construct simultaneously. For example, some research may have found different correlations between heart rate and SA because the experimental design leads to them measuring heart rate and WL [15, 16, 23–25]. This could also be due to differences in experimental tasks, participants, or measuring different aspects of SA. However, without concurrent estimation there

is no way of knowing the true cause of these differences. It will also allow for comparisons of what physiological signals are most useful for predicting certain cognitive states, potentially allowing for the reduction of the number of sensors needed for future experiments.

In addition to physiological based measures, embedded measures may also be another useful way of unobtrusively measuring an operator's cognitive state. Embedded measures are based on actions or behaviors that operators are naturally conducting during task completion. They have the advantage of being collectable in real-time, which would allow for use in future real-time cognitive state models. It has been shown that some of the common embedded measures may be insufficient on their own in capturing cognitive states [26]. However, they may be useful real-time measures in the models for cognitive states, especially in combination with other performance, demographics, or physiological measures. Additionally, embedded measures may have the same confounds that physiological measures do, and there is evidence that some embedded measures can be correlated to multiple states [27]. This further highlights the need for simultaneous estimation of cognitive states.

This research will use the same data set that Kintz [26] and Zhang [28] both used in building models to estimate cognitive states. These prior research efforts only included embedded measures or physiological models and focused on correlating them to a single cognitive state at a time [26,28]. This work aims to build upon these and model cognitive states with both embedded measures and physiological signals to create stronger models, with better fit and predictive capabilities. Additionally, the simultaneous fitting of SA and WL may be able to reduce the ambiguity in different signals and reduce the future number of sensors required, as the predictors that are selected for these models would be the same [26]. Finally, this thesis will use more robust statistical modeling methods to select the most useful predictor variables than the previous research used [26, 28].

1.2 Research Objectives

The goals of this research are as follows:

(1) Develop a model to predict operator's trust based on unobtrusive physiological measures and embedded measures.

(2) Develop a model to simultaneously predict operator's situation awareness and mental workload based on unobtrusive physiological measures and embedded measures.

The hypothesis are:

(1) The inclusion of both physiological and embedded measures will allow for improved model performance than unimodal models

(2) Simultaneous fitting of situation awareness and workload will allow for more parsimonious models than two independently fit models.

1.3 Research Approach

Chapter 2 of this document will focus on the methods used for the study and analysis. The experimental protocol and simulation environment will be defined. Additionally, the statistical methods used to select the included predictor variables and create the models will be defined.

Chapter 3 is focused on the Trust model. First, a literature review of previous efforts in modeling trust will be presented. Then, the chapter includes the final trust models and performance characteristics. These will be compared to models that only used embedded measures or physiological measures.

Chapter 4 is on the SA and WL models. This is like the trust chapter, except the comparisons also include models that were only fit to SA and only to WL.

Finally, chapter 5 is the conclusion. This will discuss the limitations of these models, and future work for how these models can be used to advance real-time modeling of cognitive states.

Chapter 2: Methods

2.1 Simulation Environment and Tasks

This study was approved by the University of Colorado Institutional Review Board (Protocol #19-0434). Fifteen subjects voluntarily participated in the study. The participants performed a simulated spacecraft docking inside CU Boulder's Aerospace Research Simulator (AReS), a spacecraft cockpit mockup, as seen in Fig. 2.1. Participants completed 12 trials; each trial consisted of a 50 second piloting task, followed by a 20 second docking task, survey completion, and a feedback assessment of the trial. During the trials they were wearing physiological sensors to measure ECG, RSP, EDA, and eye movement. More detailed methods and experimental protocols have been described previously in the theses of Kintz [26] and Zhang [28]. This thesis focuses on using the same data set with more robust modeling methods, co-estimates of SA and WL, and including both embedded and physiological measures.

Figure 2.1: Example of participant inside the AReS mockup during a trial. The primary flight display (top), secondary display (left) and reward display (center) are visible. [26]

The 50 second piloting task was used to understand SA and WL. It consisted of a 2D tracking task in which participants used a joystick to keep a space station target centered in the crosshairs of their spacecraft's 'docking camera' (Fig. 2.2), with a primary goal of a final docking with

no horizontal or vertical offset of the spacecrafts. At the beginning of each trial the spacecraft began with 100% fuel, and it approached the target at a constant rate. During the approach, the spacecraft experienced randomly applied perturbations in the horizontal and vertical directions, requiring participants to use throttle inputs to correct. Fuel decreased based on the magnitude of throttle inputs used. Three different task load levels (low, medium, high) were used in this study, corresponding to varying magnitude of the perturbations; each subject saw each task load level four times in a randomized order. This was done to try to induce a broad range of SA and workload to be assessed across subjects.

During the docking, a secondary push button lighting/response task was used as an embedded measure for WL. This is a standard method of assessing workload as a secondary task [29, 30]. Participants monitored a data-link light (left side of Fig. 2.3); when the light turned on, participants pressed a green or blue button depending on the color of the data-link light. Participants were instructed to treat this task as secondary and perform it when not occupied with the primary tracking and docking task. The location and colors of the light made it difficult to notice with peripheral vision, as thus participants needed to actively monitor the secondary task light to detect changes. For example, someone with high workload may not be able to notice or respond to the changing color of the data-link light and instead need to focus on the task.

A tertiary verbal callout task was used as an embedded measure for SA, which is a standard method of assessing SA [27]. Participants were also asked to perform verbal callouts of 'distance to capture' and 'remaining fuel' at intervals of 1.0 and 10% respectively; these values were located on the primary flight display as seen in Fig. 2.2. Callouts were considered successful if occurring within 2 seconds of the event occurring. Participants were instructed to do this only when not occupied with the primary tracking task and secondary lighting task. For example, a participant with high SA would be aware of these values changing and be able to respond at the appropriate time.

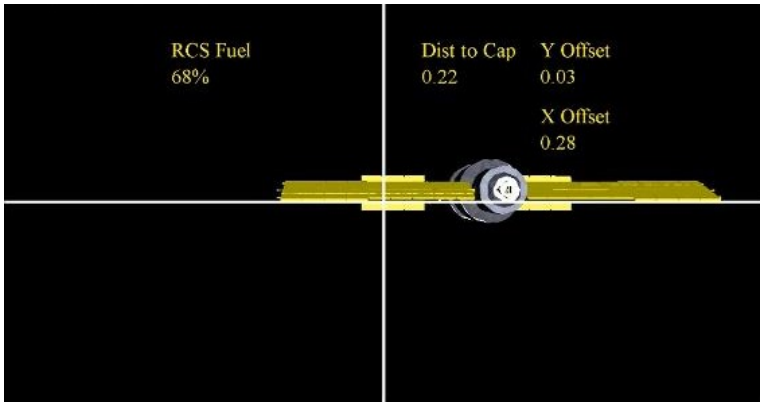


Figure 2.2: The primary flight display during a typical piloting phase. The X, Y offset to be minimized are in the upper right. The "RCS fuel" and "Dist to Cap" represent the tertiary callout measures. [26]

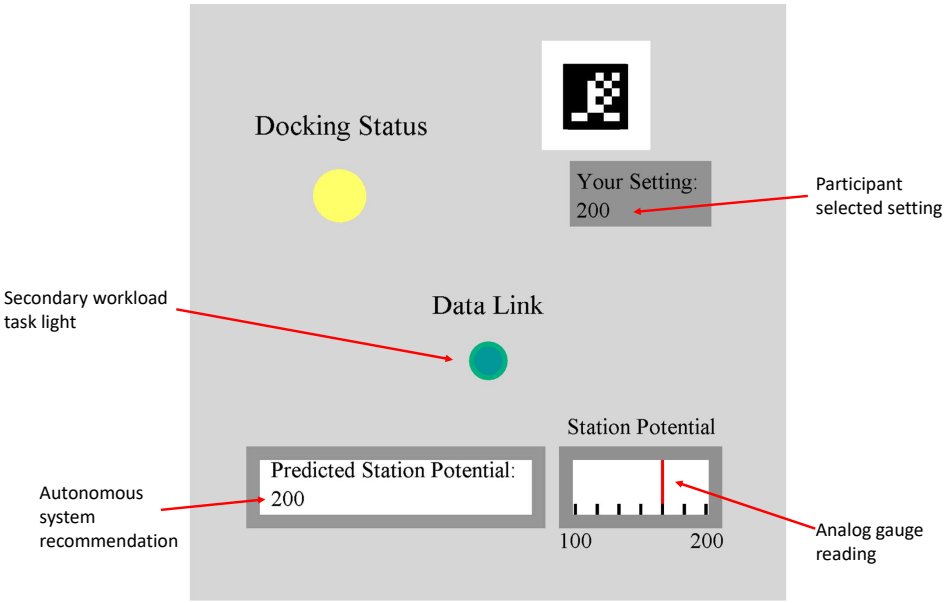


Figure 2.3: An annotated secondary flight display.

Immediately following the pilot task, participants performed a trusting task related to the docking. This trusting task was not included in the SA and WL models, and instead in its own trust model since it occurred over a different time period. All tasks and actions occurred during a 20 second interval; participants were asked to pick one of two (100 or 200 V) 'voltage potential settings' for their spacecraft to match the space station. They were informed that a mismatch in voltage between their spacecraft and the station would cause damage. Participants had two ways of identifying voltage of the spacecraft as seen in Fig. 2.3: an autonomous system that would

make a discrete recommendation (100 or 200 V), and a voltage gauge with noise and error in its measurements that decreased as the spacecraft got closer (settling on one of 12 equally spaced values between 100 and 200 V).

The participants were told that the autonomous system had access to additional information to provide its recommendation, that it was not malicious and always trying to help, and that it was the same system for all trials. In reality, the autonomous system had a 75% reliability and gave the correct recommendation 9 times in the 12 trials. This was motivated by other studies using autonomous decision aid reliability in the 60-90 % range [31–35]. Predicting the trust in this autonomous system is the goal for the first research objective. The analog gauge was provided to give participants environmental context to make their decision on, to avoid blindly trust the autonomous system's recommendation; however, the analog gauge setting was more conclusive for some trials (e.g., settling on 155 V vs 200 V). Participants were given a monetary bonus for selecting the correct voltage setting/not damaging the spacecraft, and for making a quick decision. This reward structure encouraged the participant to optimize for response time and accuracy, which are supported by the autonomous system and analog gauge, respectively. The time it took for participants to make a selection was used as the embedded measure of trust, which is a commonly used metric for understanding trust in autonomous systems [23, 34, 36]; the faster the participant responds the more likely they are to trust the autonomous system.

At the end of each trial participants completed three questionnaires: Modified Bedford, Situation Awareness Rating Technique (SART), and Trust in Automated Systems (TAS) [37–39]. These surveys will be described in more detail in Ch. 3 and 4. After completing the surveys, participants were given a feedback assessment of their trial, consisting of the docking outcome (successful or unsuccessful), and bonuses for secondary rewards (piloting accuracy).

In addition to the 12 trials of the task, participants completed a questionnaire about their demographics, sleep, video game usage, handedness, experience with aerospace-relevant displays, and experience with autonomy. They also completed five trials of an internet version of the Psychomotor vigilance Test (PVT) [40] used to assess alertness. Finally, they completed the Automated

Induced Complacency Potential (AICP) questionnaire [41] to understand their general tendencies towards being compliant.

2.2 Variables

2.2.1 Response Variables

The gold-standard questionnaires provided context for the perceived SA (SART), WL (Modified Bedford) and Trust (Trust in Automated Systems). While the Bedford Scale is ordinal, it was treated as a continuous variable [42]. These are used as the ground truth values for cognitive states in the modeling effort of this thesis and are the outcome variables for the regression.

Table 2.1: Description of Response Variables

Metric	Description	Values
TAS	Trust in Automated System survey score; used as the ground truth for trust	12-84
SART	Situation Awareness Rating Technique survey score; used as the ground truth for SA	-14-46
Bedford Scale	score from the modified Bedford Scale; used as the ground truth for WL	1-10

2.2.2 Predictor variables

The predictor variables for the experiment can be broken down four categories: physiological measures, embedded measures (EM), observable measures, and demographics. A description of the differences in these categories is in Tab. 2.2.

Table 2.2: Description of Predictor Variable Categories

Category	Description
Physiological	Measurement of an individual's body functions, such as heart rate or respiration
Embedded	Actions or behaviors that operators are naturally conducting during task completion and directly associated with a cognitive state
Observable	Actions that are directly observable, such as number of trials, and performance-based measures that aren't designed to be directly associated with a cognitive state and are specific to the experiment
Demographic	Specific to a subject does not depend on experimental outcomes

2.2.2.1 Physiological Measures

ECG, RSP, and EDA were collected with a Biopac MP160 and Bionomadix wireless collection at a 2000 Hz sampling rate. Eye-based measures were collected through Pupil Lab's Pupil Core headset, and the datastreams were synchronized using Lab Streaming Layer. Unfortunately, the EDA data was of poor quality and not used in this analysis. The ECG and RSP data were processed in Python using the Neurokit2 toolbox [43]. Eye-based measures were calculated from Pupil Lab's processing software and blink counts and gaze metrics were manually counted for each subject. A description of the included physiological measures is in Tab. 2.3. Raw data, as opposed to standardized data or data compared to a baseline, was used as future work of these models will be applied to real-time, ongoing data collection. While most Heart Rate Variability (HRV) metrics are typically calculated over 5 minutes or 24 hours, there have been studies suggesting that these metrics can be calculated over a 10-60 second duration (depending on the metric) and achieve comparable results [44–47]. The key factor is that when comparing HRV metrics they are calculated over the same duration epoch [48, 49]. The 50 second piloting period yields the HRV metrics listed in tab. 2.3 for SA and WL. For the trust model, the physiological data is split into pre-lock-in and post-lock-in windows during the 20 second trusting task time frame. The difference and ratio of these metrics are also included. Due to the limiting factors of the epoch duration, the only ECG based measures that can be included in the trust models are Heart Rate and Heart Rate Variability Root Mean Square of Successive Differences(HRV RMSSD). Unlike the measures obtained from

the ECG signal, the respiration and eye-based measures did not have a time dependency, so were retained across models.

Physiological measures will be colored purple for the duration of this thesis.

Table 2.3: Description of physiological measures [48, 50]

Metric	Description	Value
Heart Rate	Number of heart beats per minute	min = 65, max = 112
HRV RMSSD	Square root of the mean squared difference of successive NN Intervals	min =7 , max = 75
HRV Mean NN	Mean NN interval length	min 538= , max =913
HRV SDNN	Standard deviation of the NN interval	min = 15, max = 82
HRV SDSD	Standard deviation of differences between adjacent NN intervals	min =7 , max =76
HRV CVNN	The standard deviation of the NN interval length divided by the mean of the NN interval length	min = 0.02, max = 0.13
HRV CVSD	The root mean square of the sum of successive differences dividend by the mean of the NN interval length	min =0.02 , max =0.14
HRV MedianNN	Median NN interval length	min = 531, max = 913
HRV MadNN	Median absolute deviation of NN interval	min = 13, max = 82
HRV MCVNN	Median absolute deviation of NN interval divided by the median NN interval length	min = 0.01, max = 0.12
HRV IQRNN	Interquartile range of the NN interval	min = 19, max = 108
HRV pNN50	Proportion of successive NN intervals differing by more than 50 ms	min = 0, max = 38
HRV pNN20	Proportion of successive NN intervals differing by more than 20 ms	min = 1, max = 82
RSP Rate	Number of breaths per minute	min = 4.6 , max = 34
RSP Amplitude	Average amplitude of each breath	min = 0.4, max = 11
Blink Count	Number of blinks per trial	min = 0 , max =62
Mean Pupil Diameter	Average pupil diameter, including only diameters in nominal range	min =3 , max =8

2.2.2.2 Embedded Measures

Embedded measures are actions or behaviors taken that are naturally occurring during task completion, and directly related to a cognitive state. There are two embedded measures for trust. The first is the time to lock into the desired voltage setting. The second is a group of gaze-based metrics. These include how long the participant spent on a certain area, the number of times a participant looked at an area, how often the participant switched between two areas, and the total number of switches. The areas of interest are seen in Fig. 2.4. The trust embedded measures are described in more detail in Ch. 3

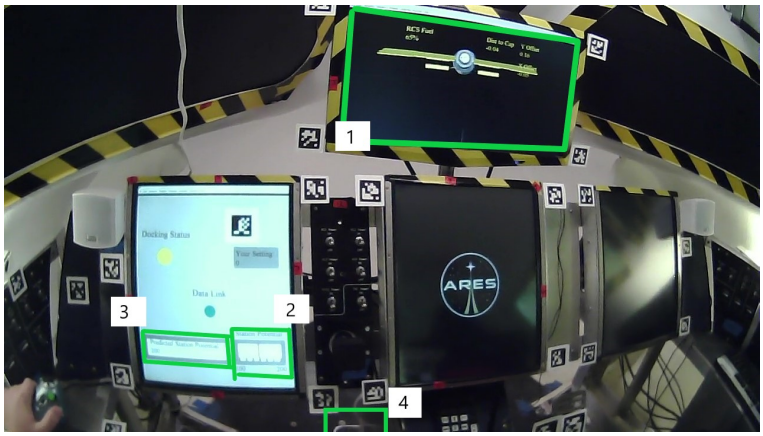


Figure 2.4: The zones of interest for the gaze-based metrics for trust. Zone 1 is the primary flight display, zone 2 is the analog gauge, zone 3 is the autonomous system recommendation. Zone 4 is the controller where the subject inputs their selected voltage; this is off screen of the photo.

Likewise, the details of the SA and WL embedded measures are described in Ch. 4. For workload, the embedded measure is the percent of time that the outer ring was lit, out of the total possible time. The lower the lighting percentage is, the quicker the subject responded and turned off the ring. For SA, the embedded measure is the percentage of successful callouts made of the total number of possible callouts. The total number of callouts depended on the magnitude of the control inputs made, since more frequent inputs caused more frequent "RCS Fuel" level callouts. At a minimum subjects made 6 verbal callouts on each trial for the distance callouts.

Embedded measures will be color coded as yellow for the duration of this thesis.

Table 2.4: Description of Embedded Measures

Metric	Description	Value
Time to lock in	Time to lock in the voltage setting; metric to measure Trust	0-20
Gaze-based metrics	Time spent on various displays, and how often the subject switch between them; metric to measure Trust	NA
% of Callouts	Percent of verbal callouts made; metric to measure SA	0-100
Lighting %	Percent of time the secondary light was on; metric to measure WL	0-100

2.2.2.3 Observable Measures

Observable measures include items that are directly observable, as well as performance based observable measures.

The observable info for trust includes the number of trials previously completed and feedback on correct or incorrect decisions in previous trials (the number of times the subject previously caused an arc). An expectation metric is used as the performance-based observable metric for the trust model. The expectation metric is how the autonomous system's recommendation compared to the observed information from the analog gauge and is calculated by integrating the analog gauge needle's position with respect to time until the participants locked in their decision. The center point of 150 V was treated as the 0 value for integrating. This results in a value that describes how strongly the needle moved to one side or the other. The absolute value of this value was multiplied by +1 when the system recommendation was correct and -1 when the system was incorrect.

For the SA and WL models the observable information also included the number of trials previously completed. In addition, the task load setting is included. Task load was manipulated as the independent variables in the SA/WL part of the experiment, by increasing the magnitude of random perturbations during each trial. Changing the task load was done to force changes in workload and situation awareness. The performance-based observable measures are based on the summed magnitude of the joystick input over the docking time, and the root mean squared of the tracking error (how far off the spacecraft was from the guidance cues).

While some of the observable measures, especially those associated with performance, may

seem like they could be embedded measures, they are distinct as they are not designed with measuring a specific cognitive state in mind. In addition, these performance measures also require knowledge of the task goals and performance criteria.

The observable measures will be colored red for the duration of this thesis.

Table 2.5: Description of observable measures

Metric	Description	Value
Number of Previous Arcs	Number of incorrect decisions in previous trials; metric for trust	0-11
Number of Trials	Previous number of trials completed; metric for trust, SA, and WL to capture some time dependency	0-11
Expectation	How autonomous system recommendation compared to analog gauge, with negative meaning dissonance and positive being agreement; metric for trust	min = -0.18, max = 0.18
Task Load	Low, medium, or high based on the magnitude of perturbations during docking; metric for SA and WL	-1,0,1
Joystick Input	Summed magnitude of joystick control inputs; metric for SA and WL	min = 99, max = 1125
RMS Tracking Error	Root mean square of the tracking error; metric for SA and WL	min = .148, max = 1.68

2.2.2.4 Demographics

Demographic information was collected for each subject; this information does not depend on experimental outcomes. For trust, the demographics include age, sex, hours of sleep the previous night, and quality of sleep . In addition, it includes usage of autonomous systems, navigational aid, and their score on the Monitoring questions of the AICP questionnaire [41]. The questionnaires used are in Appendix D

For the SA and WL model, the demographics collected includes age, sex, video game usage, previous aerospace relevant display usage, hours of sleep the night before, quality of sleep, performance on the PVT (average reaction time in milliseconds over 5 prompts), and what is their dominant hand.

Demographics will be color coded blue for the duration of this thesis.

Table 2.6: Description of demographics measures

Metric	Description	Value
Age	How old the subject is	min = 19, max = 32
Sex	Male or Female	0,1
Hours of Sleep	Hours of sleep the previous night, rounded to nearest half hour	min = 5, max = 9
Sleep Rating	Self-rated quality of sleep	-2, -1, 0, 1, 2
Robot/ Autonomous System User	Previous use with autonomous systems; metric for trust	0,1
Navigational Aid user	Use of navigational aid like google maps once a week; metric for trust	0,1
AICP Monitoring Score	Automated Induced Complacency Potential Score for general tendency to be complacent; metric for trust	1-25
Video Game Usage	Previous experience playing video games; metric for SA and WL	0, 1, 2, 3
Aerospace Display experience	Experience with aerospace like displays; metric for SA and WL	0, 1, 2, 3
PVT Score	Psychomotor vigilance task score for reaction time; metric for SA and WL	min = 211, max = 333
Handedness	Left of Right handed; metric for SA and WL	0, 1

2.3 Statistical Methods

The objectives of this research involve developing models that use embedded measures, physiological measures, observable information, and demographics to predict trust by itself, or SA and WL simultaneously. The models were created by fitting a selection of the dependent and independent variables to the gold-standard scores survey scores. Lease Absolute Shrinkage and Selector Operator (LASSO) regression was used to downselect the predictor variables to those that were most important [51, 52]. These downselected variables were then fit using ordinary least squares regression (OLS) to the subjective survey scores, which are a stand in for the ground truth values.

Three different model types, described in Tab. 2.7, were considered to determine what variables were available to the LASSO algorithm. The models are all universal and can be used

without prior observation of the operator. Model type 1 is to be used as validation of embedded measures and physiological measures on their own, as well as a comparison of the relative use of these metrics. Model 2 also includes observable measures, thus requiring no knowledge of the operator. Model type 3 includes demographic information and allows the models to be customized to a particular operator with information gathered before the task.

Table 2.7: Description of model types and available predictors

Type	Embedded Measures	Physiological Measures	Observable Info (ie. Task load)	Demographic Info	Purpose
1	(included)		(excluded)		Validation of embedded and physiological measures
2					Requires no information about the operator
3					Can be tailored before the task

In total, there are 15 participants over 12 trials, leading to 180 observations. While there may be time dependencies, the trials are treated as independent. For the trust models, two participants were removed because their responses to the TAS questionnaire was less than 10% of the range of the possible survey results, indicating that the subject had a misunderstanding of the questionnaire, or their ratings were immune to the manipulations [26]. As such there are 156 observations. For trust, there are up to 66 predictor variables (2211 including interactions). In the second set of models for SA and WL, all subjects are accounted for, leading to 180 observations and up to 31 predictor variables (496 including interactions).

For each model type, two different regression models were fit: one without interactions, one with interaction terms. Both types were considered to see if adding interactions benefited the model's performance.

2.3.0.1 Model Fitting

Models were fit using the LASSO shrinkage method because it can be used for simultaneous fitting and when the number of predictors is greater than the number of observations, which is the case for some of the models. Previous modeling studies used stepwise to select the predictors, however, due to the desire for simultaneous fitting and constraint of more predictors than observations LASSO was used for this research [26, 28].

The predictors were selected by running a 10-fold cross validation relaxed LASSO in R. The simultaneous SA and WL model used the 'mgaussian' family to enable multivariate selection [52–54]. The LASSO estimator is defined in 2.1 [55].

$$\hat{\beta}^{\lambda} = \min_{\beta_j} \left\{ \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.1}$$

where y is the normalized outcome variables, x are the normalized predictor variables, and β are the coefficients. The coefficients from LASSO are different than the coefficients that would be found through stepwise or least squares regression. λ denotes the amount of shrinkage, $\lambda = 0$ implies all features are considered, and $\lambda = \infty$ implies that no features are considered. λ is decided through cross validation (in this case 10-fold is used), where a λ value can be selected based on mean square error (MSE). The two commonly selected values of λ are λ min which is at the smallest MSE achieved, and λ 1SE which is the largest λ at which the MSE is within one standard error of the lowest MSE. The 1SE location is also chosen as it leads to a more parsimonious model whose accuracy is comparable with the best model [52, 56].

With large numbers of predictor variables, LASSO sometimes selects noisy variables that do not actually contribute to the model. Relaxed LASSO incorporates another coefficient, γ, to represents the degree of shrinkage from the original model, which helps eliminate variables only chosen due to noise [52, 57]. Essentially, relaxed LASSO runs LASSO to identify a set of non-zero coefficients, then apply LASSO to these selected non-zero predictors. Since the variables in the second step have less competition from the noisy variables, cross validation tends to pick more

parsimonious models [52].

The relaxed LASSO estimator is defined for $\lambda \in [0, \infty)$ and $\gamma \in (0, 1]$ in Eq. 2.2:

$$\hat{\beta}^{\lambda,\gamma} = \min_{\beta_j} \left\{ \sum_{i=1}^{N} (y_i - x_i^T \{\beta \cdot \mathbf{1}_{\mathcal{M}_\lambda}\})^2 + \gamma\lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.2}$$

where λ is still the amount of shrinkage, γ represents the degree of relaxation, $\mathcal{M}_\lambda$ is the set of variables selected from the original LASSO algorithm, and $\hat{\beta}$ is the vector of coefficients. $\mathbf{1}_{\mathcal{M}_\lambda}$ is the indicator function on $\mathcal{M}_\lambda \subseteq \{1, ..., p\}$ such that for all $k \in \{1, ..., p\}$

$$\{\beta \cdot \mathbf{1}_{\mathcal{M}_\lambda}\}_k = \begin{cases} 0 & k \notin \mathcal{M}_\lambda \\ \beta_k & k \in \mathcal{M}_\lambda \end{cases}$$

Note that only the predictors in $\mathcal{M}_\lambda$ are considered for the relaxed LASSO estimator, and a γ of 1 is equivalent to the typical LASSO algorithm.

Like before, cross validation can be used to select the optimal λ and γ values as seen in Fig. 2.5. The two vertical lines represent λ min and λ 1SE. The different colored lines represent the degree of relaxation, and the top axis shows the number of predictors. As can be seen 1SE reduces the number of predictors compared to the minimum value, but still has comparable performance (based on MSE). In order to have more parsimonious models with comparable performance, higher degrees of relaxation are needed.

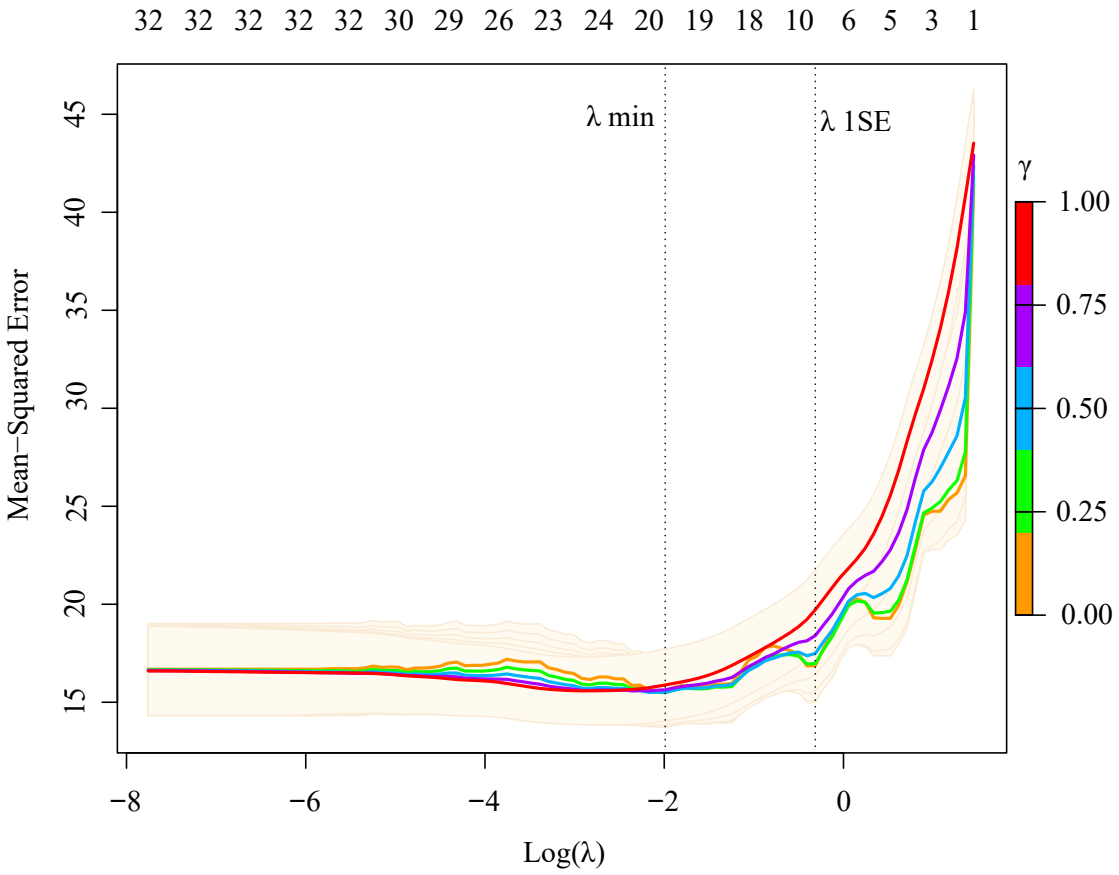


Figure 2.5: Cross Validation for LASSO. γ represents the degree of relaxation, and the top axis is the number of predictors. The vertical lines represent the λ values for minimum MSE and at the 1SE location, respectively. The shaded orange is the confidence bands.

LASSO can be used to select predictors and their coefficients, but the coefficients are not consistent. Consistency refers to the idea that as the sample size grows, the estimates converge to their true values; Ordinary least squares (OLS) is consistent. A LASSO-OLS hybrid uses relaxed LASSO to determine the predictor variables in the model but finds the coefficient values using OLS [52]. In this case, the OLS coefficients were fit from the raw data so that the models can be applied on future subjects in real-time without having to normalize the data.

2.3.1 Model Selection

The LASSO process as described above can result in different selected predictor variables. A criterion to select from the host of models generated was created; the final model was selected

based on its quality of fit and predictive capabilities. The quality of fit was measured through adjusted R^2 and RMSE. Adjusted R^2 was selected to evaluate the quality of fit as it accounts for the number of predictors and penalizes for overfitting [58]. This was important, as having a more parsimonious model is desirable. Once fit, the predictive capabilities were measured based on the metric Q^2, which is a measure of predictive performance, on both a per subject and per trial basis. The Q^2 by subject evaluates how effective the model is at predicting the score of a new participant, whereas the Q^2 by trial is an exhaustive leave one out cross validation. Q^2 is a metric analogous to R^2 that assess the predictive power as a proportional reduction of error and can be seen in Eq. 2.3 [59]. However, Q^2 can be negative or have a magnitude greater than 1; a negative Q^2 means that the model has less predictive power than guessing the mean of the training data set values.

$$\begin{gathered} Q^2 = 1 - \frac{\Sigma(y-\hat{\mathbf{y}})^2}{\Sigma(y-\bar{y})^2} \\ y = \text{left out data} \\ \hat{\mathbf{y}} = \text{predicted y for the left-out data} \\ \bar{y} = \text{mean of the training data} \end{gathered} \tag{2.3}$$

Models were judged based on adjusted R^2, RMSE, Q^2 by subject, and Q^2 by trial to determine which was best. Models that had poor performance in any of the categories were eliminated. The final model was selected by prioritized Q^2 by subject as the models should be predictive to new subjects. For future human-autonomy teaming applications, being robust to a new operator and new data is important. In the event that multiple models had the same performance, the more parsimonious model was selected to avoid overfitting.

The simultaneous models for SA and WL tried to achieve a balance between SA and WL performance. A model that may be able to fit SA extremely well might do poorly on WL, and thus should not be considered, as the goal is to fit both SA and WL well.

2.3.2 Algorithms

The process of using LASSO and the final downselection is describe using flowcharts for both the baseline models without interactions and the models with interactions below.

2.3.2.1 Baseline Models Without Interactions

The process for fitting baseline models can be seen in the flowchart (Fig. 2.6). The available measures for a given model type, as was previously described, were fed into the algorithm. The predictor variables were selected using a relaxed LASSO with the λ and γ values at the 1SE location. With a large number of predictor variables, LASSO models to not always converge to the same set of predictors. As such, the LASSO model was ran fifty times, and each unique solution was retained. These models were fit using OLS to determine the coefficient values. From these, the model with the best fit performance and cross validation predictive performance was selected.

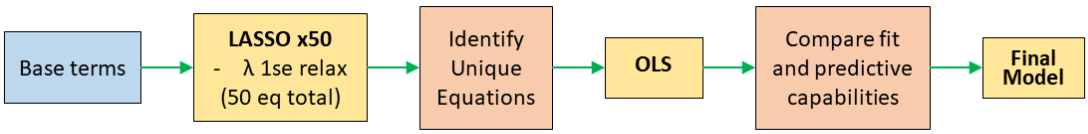


Figure 2.6: Flowchart showing the LASSO based methods used for the without interaction models

2.3.2.2 Models With Interactions

For the models with interactions, the method described above led to results that were over-fitting the data. Due to the large number of predictor variables compared to the observation (2211 predictors vs 156 observations for type 3 trust and 528 vs 180 in the case of SA and WL model), predictors were being selected more due to noise than their quality. In order to solve this issue, the idea of relaxed LASSO was taken further as can be seen in the flowchart in Fig. 2.7.

For each model type, all possible interactions were generated between the available measures. These interactions and the base measures were fed into the algorithm. Like before, LASSO was ran fifty times, and each time the set of predictor variables were selected using relaxed lasso with λ

and γ values at the minimum MSE and 1SE locations. These models were used to downselect the measures. Any term that showed up in one of the resulting equations was fed back into LASSO. Once again LASSO was ran fifty times and the λ and γ values at the 1SE were selected, and each unique solution was retained. These models were fit using OLS, and the final model was selected based on the best fit accuracy and predictive performance.

In order to reach this final algorithm, other algorithms were explored with a summary in Appendix A.

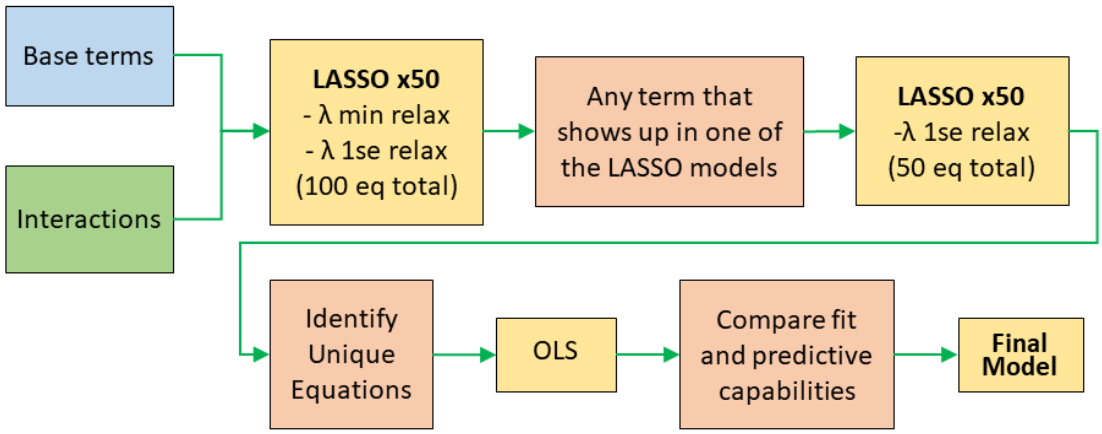


Figure 2.7: Flowchart showing the LASSO based methods used to avoid the issue of overfitting when dealing with interactions

In the case of model type 2 for SA and WL (as presented in Ch. 4), the best model was selected from the initial run through. All other final selected models were results in the second iteration of LASSO. In some cases, they were also results of the initial iteration of LASSO that were re-selected in the second iteration.

The final models are presented in Ch. 3 and 4.

2.3.3 Comparison Models

In order to test the hypotheses that having both physiological and embedded measures allows for a stronger model performance than models fit with only one of them, and that simultaneously fitting SA and WL allows for more parsimonious models than two independently fit, these alter-

native models need to be generated. In addition to the multimodal models fit for trust (MM), unimodal models that only contain physiological or embedded measures will also be fit (UM-EM, UM-phys). For the SA and WL model, the originally fit model is both multimodal and simultaneously fit (MMSF). This can be compared to models that are unimodal, but still simultaneously fit (UMSF-EM, UMSF-phys), as well as to models that are multimodal but independently fit to SA or WL (MMIF-SA, MMIF-WL). The same algorithms as described above are used to generate base models without interactions and models with interactions for the different comparison models. Additionally, comparisons can be made to test this LASSO-based algorithm to other selection methods, like stepwise by using the same predictor variables but different statistical methods.

Chapter 3: Trust

This chapter focuses on building models to predict trust from unobtrusive physiological measures and embedded measures.

One commonly used gold standard trust survey is the subjective Trust in Autonomous Systems (TAS) developed by Jian et al. [39]. This trust scale has been used extensively to capture trust during human autonomy interaction [60–64]. Operators answer twelve questions on a seven-point scale relating to their feeling of trust towards or impression of the system.

Previous work has often been limited to either 'blind trust' or co-location of human and autonomous system. Blind trusting tasks are where the operator has limited information and only has a binary choice to trust or not trust the system [18]. On the other hand, having human and autonomy co-located results in a higher degree environmental context to be able to make a decision [21, 62, 65]. This experiment allows operators to have limited addition information (in the form of the analog gauge) to mirror an operational context more closely.

Prior efforts to model trust based physiological signals has focused on classifying trust in a binary state of either high or low trust, but not estimating it as a continuous variable [21, 22]. Since trust dynamics change with time, only being able to differentiate between two trust states (high and low) has limited utility in future human-autonomy teaming. These models will allow for trust to be estimated on a continuous scale to be able to better understand trust changes.

Modeling trust with embedded measures has previously involved time-dependent measures and gaze-based measures. Autonomous vehicle studies have used metrics like the time the participant waited to take over control as a metric of trust [23, 34, 36]. In addition, gaze-based metrics, such as percentage of fixations in an area of interest, and percentage of time spent viewing an area of interest have been shown to be correlated with trust [36, 66, 67]. Using both physiological and embedded measures can help determine which of the metrics are most useful for predicting trust in future operational use, and the results are presented below.

3.1 Results and Discussions

Table 3.1 lists the predictors that were available to the LASSO algorithm for each of the 3 model types. Cells that are light blue means that predictor is available to the algorithm in that model type; grayed out cells mean the features were not available to the algorithm for that specific model type. The physiological measures include the values from the pre lock-in and post lock-in time frames as the difference and ratio of these values.

Table 3.1: List of possible predictors for trust regression models for each model type grouped by category. If the cell is blue the predictor is available for that model type; grayed out cells are unavailable.

Predictor	Model Type		
	1	2	3
ECG based measures			
Respiration based measures			
Eye based measures			
Time to lock in voltage settings			
Gaze based measures			
Expectation			
Number of previous arcs			
Number of trials previously completed			
Age			
Sex			
Hours of sleep			
Sleep rating			
Robot/autonomous system user			
Navigational aid used			
AICP Monitoring Score			

Note: Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

3.1.1 Without Interactions Models

Table. 3.2 contains the coefficients and performance for the type 1, 2 and 3 models. The coefficients are not normalized. Table 3.3 contains what percentage of the total number of predictors in a model are from the different categories (e.g., physiological, embedded, etc.) to easily see what

metric types are most important. In addition, it also contains the total number of predictor variables and the total number of sensors (out of ECG, RSP, and eye-tracking) that are needed to generate that model type; the fewer sensors required the more operationally useable the model is.

Table 3.2: Coefficients and Performance for Trust without Interactions

	Predictor	Model Type		
		1	2	3
Coefficients	(Intercept)	104.38 *	88.11 *	1.94
	HR Ratio	-40.01 *	-44.43 *	
	HRV RMSSD after		0.10 *	-0.89 *
	RSP Amp Diff.	-2.02 *	-0.82	
	Pupil Diameter After		2.79 *	
	Time to Lock In	-2.72 *	-2.46 *	-1.44 *
	% of Looks on Rec. Screen	7.51	9.98	
	% of Time on Rec. Screen	0.22 *	0.15	
	% of Looks on Throtle Input	-14.30	-9.11	
	% of Looks between Analog Gauge and Throttle Input	-10.84	-7.92	0.12
	Expectation		30.16 *	45.11 *
	Previous Number of Arcs		9.85 *	5.74 *
	Age			0.73 *
	Sleep Rating			3.93 *
	AICP			2.56 *
Performance	Adjusted R^2	0.35	0.49	0.67
	RMSE	11.90	10.49	8.50
	Q^2 by subject	0.21	0.32	0.63
	RMSE by subject	13.70	12.78	9.44
	Q^2 by trial	0.32	0.45	0.64
	RMSE by trial	12.24	10.98	8.83

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment.
*: $p<.05$

Table 3.3: Percentage of predictors in a certain category for the trust without interactions models

Metric	Model Type		
	1	2	3
% ECG based	14.3	18.2	12.5
% RSP based	14.3	9.1	0.0
% Eye based (phys)	0.0	9.1	0.0
% EM	14.3	9.1	12.5
% Gaze (EM)	57.1	36.4	12.5
% Observable		18.2	25.0
% Demographics			37.5
Number of Predictors	7	11	8
Number of Sensors	3	3	2

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

The model breakdown shows that all categories of metrics are useful in predicting trust, and as more metric types become available the accuracy is improved. The *time to lock in* embedded measure is significant in all three model types in understanding trust; the faster someone locks in the more trust they have in the system. In addition, gaze-based metrics are consistently chosen and make up a large portion of the metrics selected.

3.1.1.1 Model Comparisons

The model above in Tab. 3.2 (Multimodal model, MM) can be compared to other similar models. The comparison models focus on the utility of embedded measures and physiological features, and have either embedded measure or physiological metrics available, but not both (Unimodal models, UM-EM, UM-phys). These inform the importance of having both embedded measures and physiological features to improve fit quality and predictive capabilities. Figure 3.1 contains comparisons of adjusted R^2 and Q^2 by subject for each of the three model types. For additional reference, comparison figures for Q^2 by trial, number of predictors, and number of sensors, as well as tables with details on the coefficients and performance for these comparison models can be found in Appendix B.

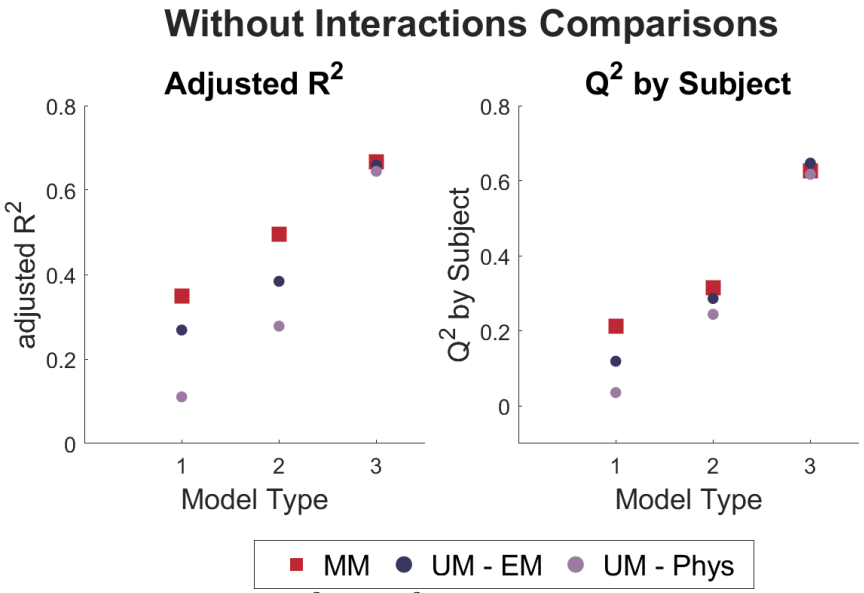


Figure 3.1: Comparison of adjusted R^2 and Q^2 by subject for the trust model without interactions

For all three models (MM, UM-EM, UM-phys), there is a large increase in predictive accuracy with the inclusion of demographic and observable based metrics. For model type 1 and 2, the multimodal models outperform the unimodal models, indicating that having both embedded measures and physiological measures may be useful. In model type 3 the models have comparable performance, and both UM models have 7 predictors vs the 8 in the MM. Additionally the embedded model requires 1 sensor (eye-tracking to gather the gaze information), and both MM and physiological models contain 2. However, all three models rely on the same observable and demographic metrics (expectation, previous number of arcs, age, sleep rating, and AICP score) which also make up the majority of the predictors. This indicates that in building accurate trust models it is important to include measures beyond embedded and physiological.

3.1.2 With Interactions Models

The performance of the trust models with interactions is in Tab. 3.4, and the breakdown of where the predictors come from can be seen in Tab. 3.5. Due to the large number of terms, the table of predictor variables and their coefficients for all three models is in Appendix B (Tab. B.1).

Table 3.4: Performance Metrics for Trust with Interactions

	Metric	Model Type		
		1	2	3
Performance	Adjusted R^2	0.36	0.46	0.72
	RMSE	11.79	10.83	7.77
	Q^2 by subject	0.23	0.39	0.68
	RMSE by subject	13.54	12.06	8.80
	Q^2 by trial	0.34	0.44	0.71
	RMSE by trial	12.02	11.12	8.03

Table 3.5: Percentage of predictors in a certain category for the trust with interactions model

Metric	Model Type		
	1	2	3
% ECG based	12.5	15.0	14.3
% RSP based	37.5	20.0	7.1
% Eye based (phys)	0.0	5.0	0.0
% EM	12.5	10.0	7.1
% Gaze (EM)	37.5	30.0	14.3
% Observable		20.0	17.8
% Demographic			39.3
Number of Predictors	8	10	14
Number of Sensors	3	3	3

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

The with interaction models include both physiological and embedded measures as important in all model types. In addition, there are significant interaction between them indicating that having both embedded and physiological measures available may be useful. Like before, as more metric types become available the model accuracy and performance increases.

Compared to the without interaction models, interactions are able to improve predictive accuracy and quality of fit. For model type 3, the adjusted R^2 improved from 0.67 to 0.72, Q^2 by subject from 0.63 to 0.68, but the number of predictors increases from 8 to 14 and the number of sensors increases by 1. The with interaction models additionally have a higher percentage of metrics containing physiological terms, and include a wider variety of physiological metrics.

3.1.2.1 Model Comparisons

The interaction model presented above (Multimodal - MM) can be compared to the unimodal models with interactions (UM-EM and UM-phys) as previously described. The comparison of adjusted R^2 and Q^2 by subject for each of the three model types is in Fig. 3.2. The details of the comparison model coefficients and performance, as well as comparison plots for Q^2 by trial, number of predictors, and number of sensors are in Appendix B.

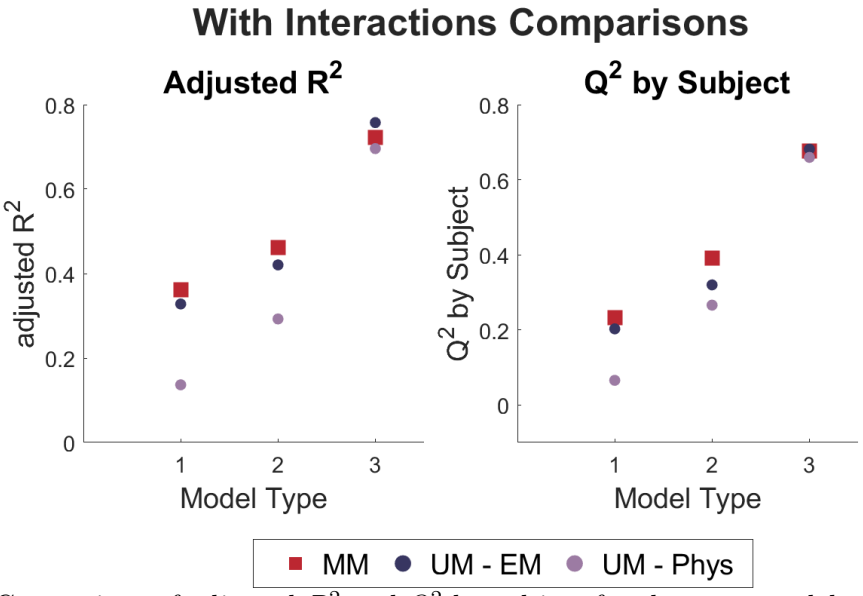


Figure 3.2: Comparison of adjusted R^2 and Q^2 by subject for the trust model with interactions

The comparisons for the with interaction models also show the utility of the combination of both embedded and physiological sensors, especially for model type 1 and 2. Like before, the inclusion of demographics (model type 3) makes it so that all the models converge to a similar performance that is better than model types 1 and 2. Comparing the required number of predictors and sensors for model type 3, the physiological model is able to provide this same fit with 2 sensors (as opposed to 3 from MM) and less predictors (8 vs 14); Fig. B.3 contains the comparisons for number of predictors and sensors. While the embedded measures are able to provide the fit with 1 sensor (for the gaze metrics), it also requires 18 predictors variables.

3.1.3 LASSO compared to Stepwise

The trust models were also used as a way to understand the utility of our LASSO algorithm as opposed to stepwise regression as a method for selecting the predictor variables. This was done by comparing models selected through LASSO to the previous stepwise models created by Kintz [26]. This previous research only considered non-gaze-based embedded measures (e.g., only *Time to lock in*), the same performance measures, and a reduced number of demographics; the LASSO algorithm was rerun with this same set of predictor variables and interactions. This use case is limited as the total number of potential predictors terms (28) is less than the number of observations (156).

Figure 3.3 compares adjusted R^2, Q^2 by subject and number of predictors for the LASSO models and stepwise models. The coefficients and performance for these models are both in Appendix B.

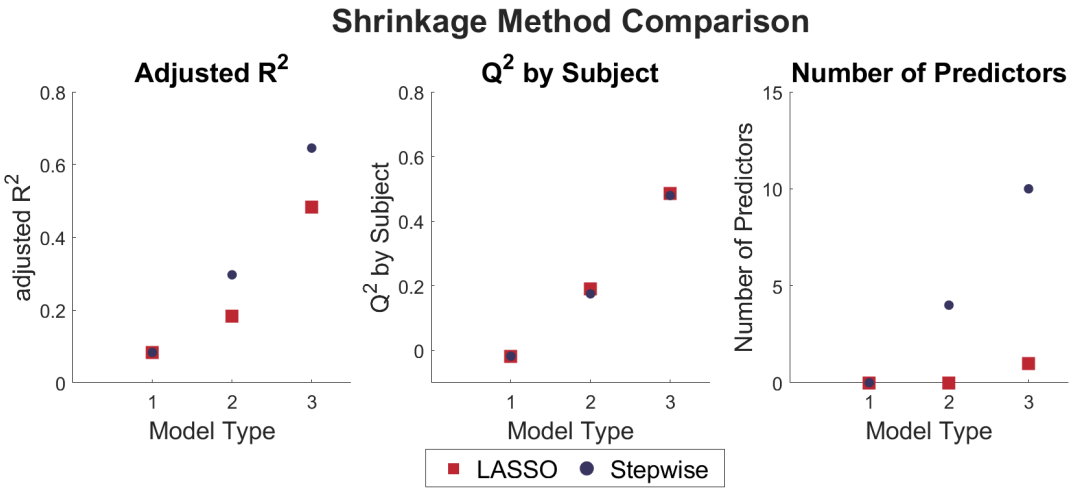


Figure 3.3: Comparison of adjusted R^2, Q^2 by subject, and number of predictors for the comparing LASSO and Stepwise capabilities

In model type 1, there was only one predictor variable available and both models selected it. For model type 2 and 3, the LASSO algorithm is able to select models with slight better predictive capabilities and a smaller number of predictor terms. However, the adjusted R^2 is reduced, which is likely a side effect of the decrease in the number of predictors.

While this test case does not show the full benefits of LASSO, it indicates that LASSO can

generate more parsimonious models with comparable predictive accuracy as stepwise regression. LASSO was chosen for this research due to its ability to fit simultaneous models, which is needed for the SA and WL combined models presented in chapter 4, as well for its efficiency when the number of predictors is greater than the number of observations (p>>n) [52]. When considering both physiological and embedded measures with interactions, this becomes the case. In addition, relaxed LASSO tends to outperform stepwise both in computational time and model quality as p gets larger [68]. There have been concerns from statisticians about stepwise selecting nuisance variables and having a low predictive accuracy; these problems become more serious as the number of predictors increases [69] [70].

3.2 Conclusion

As more multimodal features are available to the algorithm the model fit and predictive capabilities improve without overfitting, supporting hypothesis one. For model types 1 and 2, having both embedded measures and physiological measures (MM) improves performance over having only one of them (UM-EM, UM-phys). For model type 3, the performance of all models is comparable, and the majority of predictors contain observable and demographic based measures. In some circumstances, demographic information about the operator may not be available. As such, maximizing performance for model types 1 and 2 is critical. Further, some operational environments may require minimizing the number of physiological sensors the operator is required to wear. Therefore, achieving parsimony while maintaining performance is a high priority.

Including interaction terms can improve quality of fit and predictive accuracy. However, the number of predictors and sensors required also increases. The slight improvements may not be worth the additional overhead and operator discomfort.

It should be noted that there could be errors in the gaze metrics, including time spent in a zone and number of gazes to a zone, as these had to be manually counted. This was done through watching videos captured by the eye tracker. Many of these gaze metrics were seen as significant

and included in the models, however errors in counting could affect which predictors were chosen. Since this was such an important measure identified, it is important to continue to pursue eye tracking to estimate trust and work is ongoing to automate the process. In addition, while the TAS scale is the most popular trust scale and widely used, the ordering of the questions may cause a positive bias in the survey results [60].

Limitations that apply to both this Trust model and the SA and WL models will further be discussed in Chapter 5.

While this trust model fulfilled research objective one, confirmed our first hypothesis, and showed the usefulness of the LASSO algorithm, there are still open questions about the importance of simultaneous fitting. This will be addressed in the next chapter.

Chapter 4: Situation Awareness and Workload

This chapter focuses on the simultaneous fitting of situation awareness and workload. While working towards adaptive autonomy, it is undesirable to rely on signals that are ambiguously tied to a state or are not specific to that state. Creating models that fit both SA and WL together and use multiple sensors can help reduce ambiguity and determine if a change in one metric is due to a change in SA, WL, or both, which can enable the appropriate response.

To quantify SA the Situation Awareness Rating Technique (SART) [38] was used. SART is a popular subjective post trial questionnaire, where operators answer 10 questions each on a seven-point scale focused on different aspects of SA including understanding, demand, and supply (e.g., spare mental capacity). SART was selected because it was not desirable to stop the simulation during the 50 second time frame. Due to this short time length, the difference between overall operator SA and SA at the end of the trial is likely to be small.

The modified Bedford scale was used to quantify WL in this research and was chosen due to the speed at which it can be completed [37]. The Bedford Scale is a unidimensional 10-point rating scale to identify an operator's spare mental capacity while completing a task. A hierarchical design tree is used to select the rating, with each point accompanied by a description. This measures if it was possible to complete the task, if workload was tolerable, and if workload was satisfactory without reduction [71].

Most of the previous studies using physiological signals have focused on SA or WL separately, and the number of physiological sensors used has often been limited to only a few measures at a time [9,14–20]. Having access to multiple physiological data streams can allow for the determination of which metrics are important in correlating states. In addition, SA and WL are sometimes related to one another, and thus in individual studies it is not always clear if the physiological change was altered due to only a change in SA or WL.

Embedded measures may similarly have difficulty in being distinctly tied to either SA or WL. While embedded measures have typically been used only to evaluate a specific cognitive state, there has been limited research into the crossover of these embedded measures in terms of other cognitive

states. The effect of tertiary callouts for measuring SA and the cognitive state of workload were found to be correlated but they did not always reliably track each other [27]. By building models that predict both SA and WL together, the crossover effect of the embedded metrics can be further understood.

The selected models for simultaneously fitting SA and WL with embedded and physiological models are presented below.

Secondary workload metrics and tertiary verbal callouts have been used as embedded measures for WL and SA respectively. Secondary workload metrics have been used extensively to measure "how much additional work the operator can undertake while still performing the primary task to meet system criteria" [30]. While there are many examples of secondary workload metrics, the use of a pushbutton lighting/response task as described by Knowles, as used in this experiment, has been commonly used to evaluate surplus mental capacity [27, 29, 30].

Tertiary verbal callouts have been used as an embedded measure for situation awareness [27]. The operator is tasked to verbally report when certain states have changed (e.g., in a lunar landing task calling out the vehicle altitude every 100 ft). In many aerospace tasks, verbal callouts of relevant information are part of standard operation and thus this is naturally part of the task [27,29].

4.1 Results and Discussion

Table 4.1 contains the predictors that were available to the LASSO algorithm for each of the 3 model types. Cells that are blue mean that predictor was included for that model type, whereas grayed out cells means that predictor was not available to the algorithm for that model type.

Table 4.1: List of possible predictors for SA and WL regression models for each model type grouped by category. If the cell is blue the predictor is available for that model type; grayed out cells are unavailable.

Predictor	Model Type		
	1	2	3
ECG based measures			
Respiration based measures			
Eye tracking based measures			
% of callouts made successfully			
Secondary workload lighting %			
Summed magnitude of joystick control inputs			
Root mean square of tracking error			
Task load settings			
Number of trials previously completed			
Age			
Sex			
Video games rating			
Display skill rating			
Sleep rating			
Hours of sleep			
PVT score			
Handedness			

Note: Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

4.1.1 Without Interactions Models

The without interaction model predictors, coefficients, and performance for type 1, 2, and 3 are in Tab. 4.2 below. A breakdown of the percentage of each metric type (e.g., whether the metrics are physiological, embedded, etc.), like that which was presented in Chapter 3, is in Tab. 4.3.

Table 4.2: Coefficients and Performance for SA and WL models without interactions

	Predictor	Model Type 1		Model Type 2		Model Type 3	
		SA	WL	SA	WL	SA	WL
Coefficients	(Intercept)	4.30	7.86 *	17.18 *	3.65 *	5.42	1.73
	RSP Rate	0.69 *	-0.14 *	0.37 *	-4.33e-2	6.72e-2	5.97e-3
	% Callouts made	9.18e-2 *	-1.71e-2 *				
	Lighting %	-6.80e-2 *	3.28e-2 *	-2.77e-2 *	2.08e-2 *		
	Joystick Input			-2.19e-3	1.79e-3 *	-6.74e-3 *	2.04e-3 *
	RMS Tracking Error			-6.30 *	2.77 *	1.20	2.47 *
	Task load			-3.60 *	0.27	-3.53 *	0.44 *
	Number of Trials			0.30 *	-4.28e-2	0.32 *	-6.30e-2 *
	Age					0.26 *	0.16 *
	Display Skill					2.14 *	-0.57 *
	PVT					3.14e-2 *	-5.80e-3
Performance	Adjusted R^2	0.24	0.26	0.55	0.48	0.64	0.56
	RMSE	5.48	1.67	4.22	1.40	3.79	1.28
	Q^2 by subject	0.16	0.22	0.43	0.39	0.53	0.45
	RMSE by subject	5.84	1.75	4.84	1.54	4.37	1.47
	Q^2 by trial	0.23	0.25	0.53	0.46	0.62	0.54
	RMSE by trial	5.54	1.69	4.31	1.42	3.87	1.31

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment
*: $p<.05$

Table 4.3: Percentage of predictors in a certain category for the SA and WL without Interactions models

Metric	Model Type		
	1	2	3
% ECG based	0.0	0.0	0.0
% RSP based	33.3	16.7	12.5
% Eye based (phys)	0.0	0.0	0.0
% EM	66.7	16.7	0.0
% Observable		66.7	50.0
% Demographic			37.5
Number of Predictors	3	6	8
Number of Sensors	1	1	1

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment.

The model breakdown shows that all categories of variables are useful in predicting SA and WL simultaneously, and as more variable types become available, accuracy is improved. *Respiration rate* is the only physiologically feature selected, but it is selected consistently across all model types. Previous studies suggest that respiration rate may be better indicative of workload than heart-based metrics, which these models agree with [17]. While previous literature has suggested a negative correlation with SA and a positive with WL [72, 73], the models in Tab. 4.2 generally suggest a positive correlation with SA, and a negative with WL. This could be due differences in the scenario used to elicit different cognitive states in this research compared to those in the literature.

For embedded measures, the SA based metric (*% Callouts made*) is only selected in model type 1. This suggests that a better embedded measure for SA may be needed, as it was not selected over the other metrics available to the later model types. In addition, when selected both embedded measures are significant for both cognitive states. The more callouts made, the higher the SA, and the lower the workload. Likewise, the WL based embedded measure (*Lighting %*) being on more corresponds to higher workload and lower SA. In Hanley et. al. there was evidence of the correlation between the callout percentage and workload, although they did not track one-to-one [27]. However, a literature review does not seem to indicate prior work on the crossover effect between the secondary workload lighting and SA.

4.1.1.1 Model Comparisons

The model above in Tab. 4.2 (Multimodal Simultaneously Fit - MMSF) can be compared to other similar models. Models that have the same multimodal predictor variables available to them but are independently fit to either SA or WL (Multimodal Independently Fit, MMIF-SA, MMIF-WL) were created for comparison. These models were selected based on their ability to fit only one cognitive state and can allow for the characterization of the usefulness of simultaneously fitting SA and WL in reducing the number of required predictors and sensors. Models that are simultaneously fit to SA and WL but are unimodal and have either embedded measure or physiological metrics available, but not both (Unimodal Simultaneously Fit, UMSF-EM, UMSF-phys) were also fit for

comparison. These inform the importance of having multimodal models to potentially increase quality of fit and predictive accuracy. Figure 4.1 contains comparisons of adjusted R^2 and Q^2 by subject. These plots show adjusted R^2 and Q^2 by subject for all three model types with SA on the x-axis and WL on the y-axis. An excellent model fit, or prediction would be in the upper right corner of the plot. Data points that fall along the diagonal are equally good at both SA and WL. Models that fall far above or below the diagonal would be good at one cognitive state but not the other. For additional reference, comparison figures for Q^2 by trial, number of predictors, and number of sensors, as well as a table of the comparison models' coefficients and performance are in Appendix C.

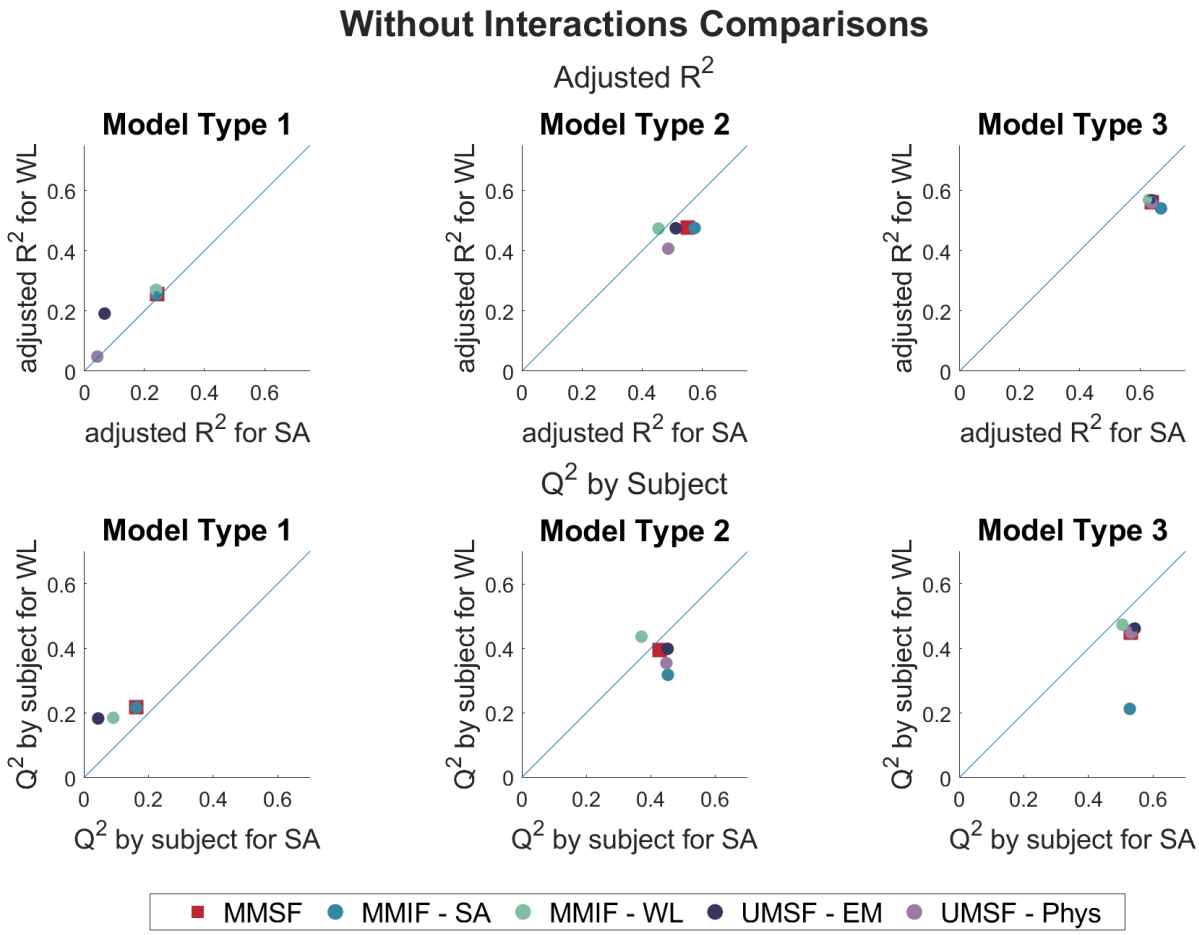


Figure 4.1: The comparison of SA and WL models without interactions to other similar models. The top row of plots compares the adjusted R^2, the bottom row compares the predictive Q^2 by subject. The x-axis represents the value for SA, and the y-axis is for WL. The blue line represents models that are equally good at SA and WL; an excellent model would fall in the top right corner.

For model type 1, the multimodal models outperform the unimodal models. This advocates for the use of including both physiological and embedded measures, versus just one of them. For model type 3 the performance of the MMSF and UMSF-EM and UMSF-phys models converge, as all the models mostly rely on performance and demographic information. Looking more closely at the comparison of just embedded measures and physiological measures, embedded measures outperform physiological measures, especially in WL. This is likely an artifact of the fact that the lighting % is a strong embedded measure for WL, but the % of callouts made is a weaker embedded measure.

While the models independently fit to SA and WL may be able to outperform the simultaneously fit model in SA and WL respectively, the use of two independently fit models would require and additional ECG sensor, as HR and HRV metrics are also now considered (C.2 and C.6). For model type 3, using the two independently fit models would slightly improve performance, however, it comes with the cost of extra sensors as in Tab. 4.4. For future operational real-time use, the reduction of number of predictors and sensors in the simultaneously fit models for comparable performance and predictive capabilities highlights the benefit of using this over the two independently fit models.

Table 4.4: Independently fit vs Simultaneously fit performance without Interactions for Model Type 3

	Simultaneously Fit		Independently Fit	
	SA	WL	SA	WL
Adjusted R^2	0.64	0.56	0.67	0.57
RMSE	3.79	1.28	3.62	1.27
Q^2 by subject	0.53	0.45	0.53	0.47
RMSE by subject	4.37	1.47	4.39	1.43
Q^2 by trial	0.62	0.54	0.65	0.55
RMSE by trial	3.87	1.31	3.73	1.30
Number of Predictors	8	8	11	8
Number of Sensors	1	1	2	0

4.1.2 With Interactions Models

The performance of the multimodal simultaneously fit model with interaction terms is in Tab: 4.5. Table 4.6 contains the percentage of the predictors that contain that metric category. Details on the predictor variables and their associated coefficients are found in Tab. C.1 in Appendix C.

Table 4.5: Performance for SA and WL models with interactions

		Model Type 1		Model Type 2		Model Type 3	
		SA	WL	SA	WL	SA	WL
Performance	Adjusted R^2	0.25	0.23	0.50	0.39	0.68	0.59
	RMSE	5.45	1.70	4.46	1.51	3.58	1.24
	Q^2 by subject	0.17	0.15	0.46	0.31	0.56	0.44
	RMSE by subject	5.81	1.82	4.68	1.64	4.24	1.48
	Q^2 by trial	0.24	0.21	0.49	0.37	0.65	0.55
	RMSE by trial	5.50	1.72	4.53	1.54	3.71	1.30

Table 4.6: Percentage of predictors in a certain category for the SA and WL with Interactions models

Metric	Model Type		
	1	2	3
% ECG based	28.6	44.4	10.0
% RSP based	28.6	0.0	10.0
% Eye based (phys)	0.0	0.0	3.3
% EM	42.9	0.0	6.7
% Observable		55.5	30
% Demographic			40.0
Number of Predictors	4	5	15
Number of Sensors	2	1	3

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment.

Compared to the without interaction models (Tab. 4.6) the with interaction models include more physiological sensors. Preciously only respiration rate was chosen, with the inclusion of interactions more emphasis is placed on HRV metrics. In addition, for model types 2 and 3 a greater percentage of predictors include some form of physiological measure. This indicates that physiological measures may provide more utility in an interaction with embedded measures, observable, or demographic information than on their own.

Embedded measures and physiological measures were found to significantly interact, reinforcing the importance of having both sets of metrics available to choose from. For example, *RSP Rate* and the SA metric *% Callouts* interaction was significant in the type 1 model for both SA

and WL, suggesting that the interactions of both physiological and embedded metric interaction can provide additional context and value.

4.1.2.1 Model Comparisons

The interaction model above in Tab. C.1 (MMSF) can be compared other similar models (MMIF-SA, MMIF-WL, UMSF-EM, UMSF-phys) generated with interactions. Figure 4.1 contains comparisons of adjusted R^2 and Q^2 by subject. For additional reference, comparison figures for Q^2 by trial, number of predictors, and number of sensors, as well and the coefficients and performance for the comparison models are in Appendix C.

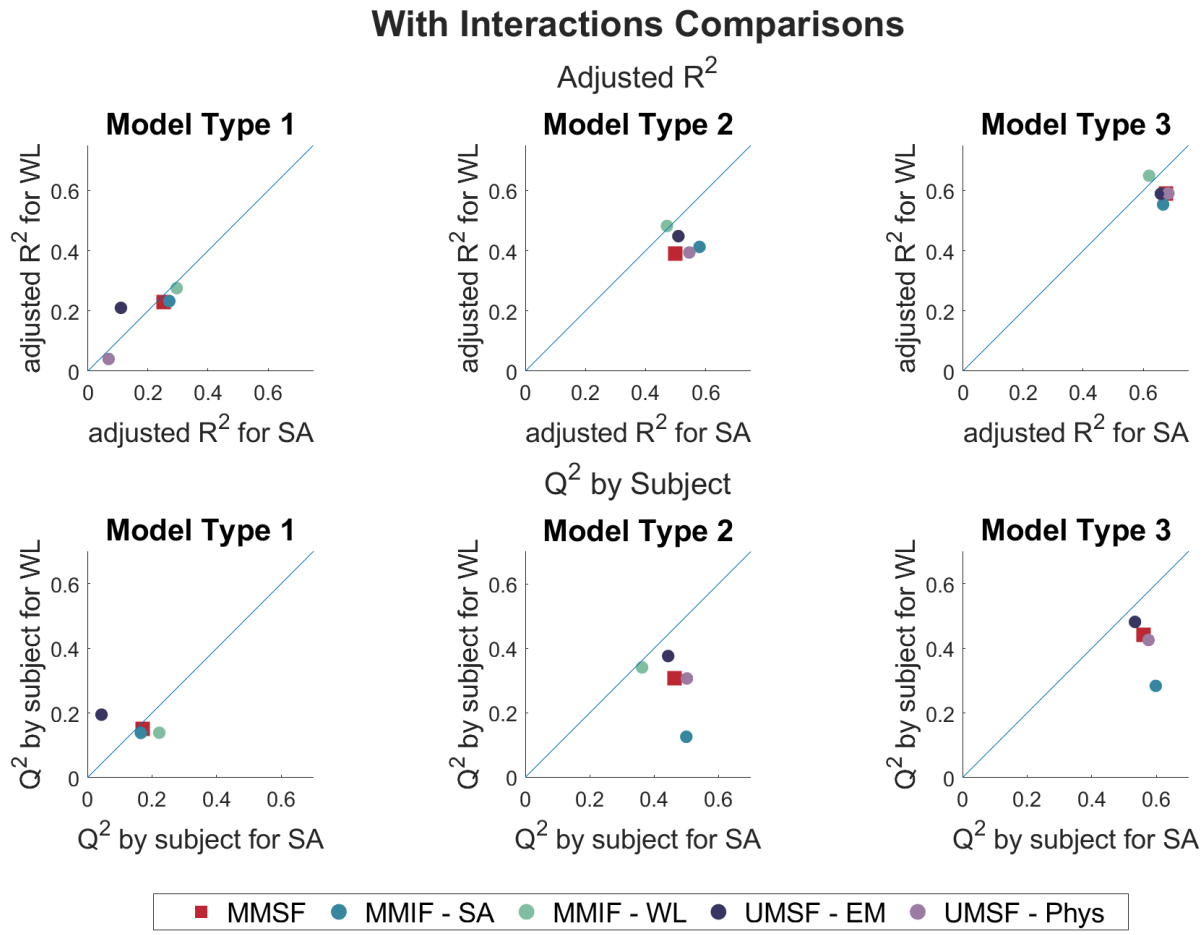


Figure 4.2: The comparison of SA and WL models with interactions to other similar models. The top row of plots compares the adjusted R^2, the bottom row compares the predictive Q^2 by subject. The x-axis represents the value for SA, and the y-axis is for WL. The blue line represents models that are equally good at SA and WL.

Unlike the without interaction models, the with interaction models do not generally outperform the comparison models. This could be due to the large number of predictors to choose from and many correlated interactions. Using two independently fit models versus one simultaneous fit model would result in the same number of sensors for model type 3. The only benefit is a reduction in the total number of predictors to be calculated. For the simultaneously fit model only 15 predictors are needed, whereas 25 predictors are needed to calculate both independently fit models.

4.2 Conclusion

In general, having the use of both embedded and physiological measures (MMSF) improves performance over the unimodal models (UMSF-EM, UMSF-phys), supporting hypothesis one. This can be especially seen in the without interaction models. In the with interaction models, it additionally provides a chance to interactions between physiological and embedded measures, which are seen to be significant in some model types. with the addition of observable and demographic based features, this difference is not as apparent, as the models rely more on these feature sets, and in some cases the embedded measure only model outperforms the multimodal model.

Overall, as the model type increases (e.g., 1 to 3) the model improves, and the comparison models converge to similar performances. This is enabled by the demographics and performance-based measures as they make up the bulk of the predictor terms, further reinforcing the importance of having a variety of predictors available.

While two independently fit models can be used together to provide a slightly better performance (MMIF-SA, MMIF-WL), it also comes with the need for increased number of sensors and predictors. For without interactions, an addition eye sensor is needed to measure pupil diameter, for with interactions the number of sensors does not change, but more predictors are needed to be computed. This supports the second hypothesis that simultaneous fitting will result in more parsimonious models.

Additionally, in general the no interaction models lead to better workload performance, while the interaction models provide better SA performance. The interaction models tend to have more predictors in the models and require more sensors. The no interaction models only require respiration sensors, while the interaction model for type 3 requires ECG, RSP, and eye sensors. The slight improvement in performances with interaction models cost parsimony and explainability; for future operational settings, the without interaction models are more beneficial due to operator comfort and usability.

An issue for these models comes from the issues with using subjective surveys as ground

truths for the cognitive states, however there was no better alternative for this task. The SART survey used is not the best available to objectively quantify SA. SART is subjective, which can result in biases as operators may not be aware of having low SA and thus cannot rate their SA accurately. It is also possible that participants failed to recall when they had poor SA during the trial and instead rated based on the end of the trial. However, since the trial lengths are short, the difference between overall SA and at the end of the trial is likely small. Additionally, SART has been shown to correlate with performance and confidence and be confounded with workload [10]. However, SART was selected as it was the best for the experimental design.

Likewise, for the main limitation for WL comes from the Bedford scale being treated as a continuous variable when it is actually ordinal. Prior research has treated the Bedford scale as continuous and ordinal variables are often used as continuous in statistical models; however, this can potentially cause improper WL classification.

While simultaneously fitting was used to help reduce ambiguity between the cognitive states, the subject responses from the task did invoke a correlated SA and WL response ($r = -.59$, $t(178) = -9.81$, $p < 0.005$) where a high SA corresponded to low WL. Not having the full range of SA and WL responses (for example low WL and low SA) makes it harder to ensure predictive accuracy in boundary cases where WL and SA do not follow each other. This limits the effectiveness of these models working towards human autonomy teaming as the autonomous system may not react properly with low SA and WL or high SA and WL.

Finally, task load was assumed to be equally spaced (such that the change in difficulty from low to medium was the same as the change in difficulty from medium to high). While efforts were made to ensure this, it is possible that there is uneven spacing in the actual task load, which would change the effect that task load has on model performance.

Limitations that apply to both the Trust and these SA and WL models will further be discussed in Chapter 5.

This chapter fulfills research objective two of creating a model that. Additionally, it further supports hypothesis one and supports hypothesis two.

Chapter 5: Conclusion

This research developed models for trust, as well as situation awareness and workload simultaneously, from a variety of predictor variables. As more categories of predictor variables were available, the performance and predictive capabilities of the model improved, supporting the first hypothesis that multimodal models will improve performance. Without the use of demographics, models containing both embedded and physiological measures outperformed the unimodal models; when demographics were included performances converged. Additionally, simultaneously fitting allowed for more parsimonious models, supporting the second hypothesis.

Models with interactions were found to slightly improve fit, but cost model parsimony and required more sensors which may not be worth the operational encumbrance and overhead. In addition, the interaction models were not as stable compared to the without interaction models when using the LASSO algorithm, suggesting that in future application these models may not be as robust as the non-interaction algorithms

One aspect that remains unknown is the degree to which these models transfer to other tasks. Many embedded and observable metrics were specific to this task. However, corollary metrics from new tasks could be identified. The exact coefficients and included metrics may be different for other experimental designs. Additionally, there is a limited number of embedded measures used for this task, other measures may provide better correlation with cognitive states (especially for SA) and can be included in future models. While model transferability is an issue across the literature [26], the methods identified here inform the importance of multi-modal data across a broad range of operational outcomes.

In addition, the dependence of HRV metrics on timescale means that these exact models could not be used for future experiments due to differing timescales; but these are able provide background on what metrics are important to take and use for future human-autonomy teaming systems. Not all HRV metrics were used due to the shorter timescales [48], but one benefit of this short timescale is that changes can be identified quicker allowing for a better estimation of state dynamics.

5.1 Limitations

There are some limitations that are the same for all the models created, beyond what was discussed in Chapters 3 and 4.

The models treated the data as 180 independent trials, even though there was a limited sample size of 15 subjects. The data is not likely completely independent, as there may be dependencies from learning effects. While we attempted to mitigate this through pre-experiment training, it cannot be certain these effects were not present. Additionally, this experiment occurred over a short duration of time so some of the results may be influenced by habituation versus actual response. It was noted that heart rate decreased throughout the experiment, even though the task load order was randomized. The physiological measures are also not compared to a subject's baseline value. While this was done to enable future work for real-time modeling for an unknown operator, the differences in subject's natural values could influence what variables were selected as being important. The models also do not account for temporal dynamics completely. While efforts were made to include aspects (such as number of trials or previous number of arcs), these may not fully capture the complete time dynamics.

5.2 Future Work

As human-autonomy teaming becomes more prevalent, there becomes a need to be able to unobtrusively measure the human operator's cognitive states. This research has developed models for predicting situation awareness and workload simultaneously using embedded measures, physiological based measures, as well as observable information. While these models are not generated in real-time, they provide insight into what metrics are most important to collect for future experiments. Additionally, these future experiments can inform on the generalizability of these models if similar metrics and parameters are chosen. They also highlight the limitations in physiological and embedded measures alone for modeling cognitive states.

In addition, future work should consider including trust into the same model as situation

awareness and workload. This will enable similar analysis to be done about the possible interaction between the embedded measures of different cognitive states, as well as enable greater understanding of how the cognitive states may be influenced by each other.

This work has shown that multimodal fitting is important, especially when demographic information may not be available. Other physiological signals such as EDA, respiration variability, and neurological measures have been shown to have high correlation with cognitive states [15, 18, 74–79], but no cross analysis including these multimodal streams have been done. Future work can consider these additional measures in generating models.

5.3 Summary

Per the research objects described previously the following aims were met:

- Developed a model to predict operator's trust based on unobtrusive physiological and embedded measures
- Developed a model to simultaneously predict an operator's situation awareness and mental workload based on unobtrusive physiological and embedded measures

These were done using the developed LASSO-based algorithm. The results supported hypothesis one that the inclusion of both physiological and embedded measures improves model performance and hypothesis two that simultaneously fitting allows for more parsimonious models. Both embedded measures and physiological measures were found to be important, however, alone they could not accurately predict cognitive states. When combined with observable information and demographics, models were able to describe and predict operator states during human-autonomy teaming. Additionally, simultaneous fitting of SA and WL results in more parsimonious models as compared to two independently fit models. These can be applied to future human-autonomy teaming problems.

Bibliography

[1] Schwarz, J. and Fuchs, S., "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)**, Oct. 2018, pp. 704–709, ISSN: 2577-1655.

[2] Lee, J. D. and See, K. A., "Trust in Automation: Designing for Appropriate Reliance," **Human Factors**, Vol. 46, No. 1, March 2004, pp. 50–80, Publisher: SAGE Publications Inc.

[3] Endsley, M. R., "Design and Evaluation for Situation Awareness Enhancement," **Proceedings of the Human Factors Society Annual Meeting**, Vol. 32, No. 2, Oct. 1988, pp. 97–101, Publisher: SAGE Publications.

[4] Hart, S. G. and Staveland, L. E., "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," **Advances in Psychology**, edited by P. A. Hancock and N. Meshkati, Vol. 52 of **Human Mental Workload**, North-Holland, Jan. 1988, pp. 139–183.

[5] Hooey, B. L., Kaber, D. B., Adams, J. A., Fong, T. W., and Gore, B. F., "The Underpinnings of Workload in Unmanned Vehicle Systems," **IEEE Transactions on Human-Machine Systems**, Vol. 48, No. 5, Oct. 2018, pp. 452–467, Conference Name: IEEE Transactions on Human-Machine Systems.

[6] Lee, J. D. and Moray, N., "Trust, self-confidence, and operators' adaptation to automation," **International Journal of Human-Computer Studies**, Vol. 40, No. 1, Jan. 1994, pp. 153–184.

[7] Stanton, N. A., Chambers, P. R. G., and Piggott, J., "Situational awareness and safety," **Safety Science**, Vol. 39, No. 3, Dec. 2001, pp. 189–204.

[8] Young, M. S. and Stanton, N. A., "Attention and automation: New perspectives on mental underload and performance," **Theoretical Issues in Ergonomics Science**, Vol. 3, No. 2, Jan. 2002, pp. 178–194.

[9] Lohani, M., Payne, B. R., and Strayer, D. L., "A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving," **Frontiers in Human Neuroscience**, Vol. 13, 2019, pp. 57.

[10] Endsley, M. R., Selcon, S. J., Hardiman, T. D., and Croft, D. G., "A Comparative Analysis of Sagat and Sart for Evaluations of Situation Awareness," **Proceedings of the Human Factors and Ergonomics Society Annual Meeting**, Vol. 42, No. 1, Oct. 1998, pp. 82–86.

[11] Schmidt, E. A., Schrauf, M., Simon, M., Fritzsche, M., Buchner, A., and Kincses, W. E., "Drivers' misjudgement of vigilance state during prolonged monotonous daytime driving," **Accident; Analysis and Prevention**, Vol. 41, No. 5, Sept. 2009, pp. 1087–1093.

[12] Dirican, A. C. and Göktürk, M., "Psychophysiological measures of human cognitive states applied in human computer interaction," **Procedia Computer Science**, Vol. 3, 2011, pp. 1361–1367.

[13] Parasuraman, R., Sheridan, T. B., and Wickens, C. D., "Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs," **Journal of Cognitive Engineering and Decision Making**, Vol. 2, No. 2, June 2008, pp. 140–160.

[14] Grassmann, M., Vlemincx, E., von Leupoldt, A., Mittelstädt, J. M., and Van den Bergh, O., "Respiratory Changes in Response to Cognitive Load: A Systematic Review," **Neural Plasticity**, Vol. 2016, 2016, pp. 8146809.

[15] Wilson, G. F. and Russell, C. A., "Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks," **Human Factors**, Vol. 45, No. 4, Dec. 2003, pp. 635–644, Publisher: SAGE Publications Inc.

[16] Zhang, T., Yang, J., Liang, N., Pitts, B. J., Prakah-Asante, K. O., Curry, R., Duerstock, B. S., Wachs, J. P., and Yu, D., "Physiological Measurements of Situation Awareness: A Systematic Review," **Human Factors**, Nov. 2020, pp. 0018720820969071, Publisher: SAGE Publications Inc.

[17] Roscoe, A. H., "Assessing pilot workload. Why measure heart rate, HRV and respiration?" **Biological Psychology**, Vol. 34, No. 2, Nov. 1992, pp. 259–287.

[18] Akash, K., Hu, W.-L., Jain, N., and Reid, T., "A Classification Model for Sensing Human Trust in Machines Using EEG and GSR," **ACM Transactions on Interactive Intelligent Systems**, Vol. 8, No. 4, Nov. 2018, pp. 27:1–27:20.

[19] Fairclough, S. H. and Venables, L., "Prediction of subjective states from psychophysiology: A multivariate approach," **Biological Psychology**, Vol. 71, No. 1, Jan. 2006, pp. 100–110.

[20] Khalid, H., Shiung, L., Nooralishahi, P., Rasool, Z., Helander, M., Chu Kiong, L., and Chin, A.-V., "Exploring Psycho-Physiological Correlates to Trust: Implications for Human-Robot-Human Interaction," **Proceedings of the Human Factors and Ergonomics Society Annual Meeting**, Vol. 60, Sept. 2016, pp. 697–701.

[21] Ajenaghughrure, I. B., Sousa, S. D. C., and Lamas, D., "Measuring Trust with Psychophysiological Signals: A Systematic Mapping Study of Approaches Used," **Multimodal Technologies and Interaction**, Vol. 4, No. 3, Sept. 2020, pp. 63, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[22] Hu, W.-L., Akash, K., Jain, N., and Reid, T., "Real-Time Sensing of Trust in Human-Machine Interactions," **IFAC-PapersOnLine**, Vol. 49, No. 32, Jan. 2016, pp. 48–53.

[23] Kunze, A., Summerskill, S. J., Marshall, R., and Filtness, A. J., "Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces," **Ergonomics**, Vol. 62, No. 3, March 2019, pp. 345–360, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00140139.2018.1547842.

[24] Mehta, R. K., Peres, S. C., Shortz, A. E., Hoyle, W., Lee, M., Saini, G., Chan, H.-C., and Pryor, M. W., "Operator situation awareness and physiological states during offshore well control scenarios," **Journal of Loss Prevention in the Process Industries**, Vol. 55, Sept. 2018, pp. 332–337.

[25] Wei, H., Zhuang, D., Wanyan, X., and Wang, Q., "An experimental analysis of situation awareness for cockpit display interface evaluation based on flight simulation," **Chinese Journal of Aeronautics**, Vol. 26, No. 4, Aug. 2013, pp. 884–889.

[26] Kintz, J. R., **Estimating Operator Trust, Mental Workload, and Situation Awareness Through Embedded Measures for Human-Autonomy Teaming**, Master's thesis, University of Colorado Boulder, 2021.

[27] Hainley, C. J., Duda, K. R., Oman, C. M., and Natapoff, A., "Pilot Performance, Workload, and Situation Awareness During Lunar Landing Mode Transitions," **Journal of Spacecraft and Rockets**, Vol. 50, No. 4, July 2013, pp. 793–801.

[28] Zhang, J., **Methods for Assessing and Managing Operator Workload and Situation Awareness in Human Spaceflight Operations**, Master's thesis, University of Colorado Boulder, 2021.

[29] Karasinski, J. A., Robinson, S. K., Duda, K. R., and Prasov, Z., "Development of real-time performance metrics for manually-guided spacecraft operations," **2016 IEEE Aerospace Conference**, March 2016, pp. 1–9.

[30] Knowles, W. B., "Operator Loading Tasks," **Human Factors**, Vol. 5, No. 2, April 1963, pp. 155–161, Publisher: SAGE Publications Inc.

[31] Nunnally, J. C., "Psychometric Theory 2nd ed." 1978, Publisher: Mcgraw hill book company.

[32] Wickens, C. D., Fitzgerald, N. J., Clegg, B. A., Smith, C., Orth, D., and Kincaid, K., "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**, Vol. 64, No. 1, Dec. 2020, pp. 1950–1954, Publisher: SAGE Publications Inc.

[33] Wickens, C. D. and Dixon, S. R., "The benefits of imperfect diagnostic automation: a synthesis of the literature," **Theoretical Issues in Ergonomics Science**, Vol. 8, No. 3, May 2007, pp. 201–212, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14639220500370105.

[34] Petersen, L., Robert, L., Yang, X. J., and Tilbury, D. M., "Situational Awareness, Driver's Trust in Automated Driving Systems and Secondary Task Performance," 2019, pp. 26.

[35] Akash, K., McMahon, G., Reid, T., and Jain, N., "Human Trust-Based Feedback Control: Dynamically Varying Automation Transparency to Optimize Human-Machine Interactions," **IEEE Control Systems Magazine**, Vol. 40, No. 6, Dec. 2020, pp. 98–116, Conference Name: IEEE Control Systems Magazine.

[36] Hergeth, S., Lorenz, L., Vilimek, R., and Krems, J. F., "Keep Your Scanners Peeled: Gaze Behavior as a Measure of Automation Trust During Highly Automated Driving," **Human Factors**, Vol. 58, No. 3, May 2016, pp. 509–519, Publisher: SAGE Publications Inc.

[37] Roscoe, A. and Ellis, G. A., "A Subjective Rating Scale for Assessing Pilot Workload in Flight: A decade of Practical Use," Tech. Rep. ADA227864, ROYAL AEROSPACE ESTABLISHMENT FARNBOROUGH, March 1990.

[38] Taylor, R. M., "Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design. Situational Awareness in Aerospace Operations (AGARD-CP-478)," **Neuilly Sur Seine, France: NATO-AGARD**, 1990.

[39] Jian, J.-Y., Bisantz, A. M., and Drury, C. G., "Foundations for an Empirically Determined Scale of Trust in Automated Systems," **International Journal of Cognitive Ergonomics**, Vol. 4, No. 1, March 2000, pp. 53–71, Publisher: Routledge _eprint: https://doi.org/10.1207/S15327566IJCE0401_04.

[40] Dinges, D. F., Pack, F., Williams, K., Gillen, K. A., Powell, J. W., Ott, G. E., Aptowicz, C., and Pack, A. I., "Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4-5 hours per night," **Sleep**, Vol. 20, No. 4, April 1997, pp. 267–277.

[41] Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., and Shirase, L., "Automation-Induced Complacency Potential: Development and Validation of a New Scale," **Frontiers in Psychology**, Vol. 10, 2019, pp. 225.

[42] Bachelder, E. and Godfroy-Cooper, M., "Pilot Workload Estimation: Synthesis of Spectral Requirements Analysis and Weber's Law," Jan. 2019.

[43] Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., and Chen, S. H. A., "NeuroKit2: A Python toolbox for neurophysiological signal processing," **Behavior Research Methods**, Vol. 53, No. 4, Aug. 2021, pp. 1689–1696.

[44] Esco, M. R. and Flatt, A. A., "Ultra-Short-Term Heart Rate Variability Indexes at Rest and Post-Exercise in Athletes: Evaluating the Agreement with Accepted Recommendations," **Journal of Sports Science & Medicine**, Vol. 13, No. 3, Sept. 2014, pp. 535–541.

[45] Salahuddin, L., Cho, J., Jeong, M. G., and Kim, D., "Ultra Short Term Analysis of Heart Rate Variability for Monitoring Mental Stress in Mobile Settings," **2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society**, Aug. 2007, pp. 4656–4659, ISSN: 1558-4615.

[46] Shaffer, F., Shearman, S., and Meehan, Z. M., "The Promise of Ultra-Short-Term (UST) Heart Rate Variability Measurements," **Biofeedback**, Vol. 44, No. 4, Dec. 2016, pp. 229–233.

[47] Nussinovitch, U., Elishkevitz, K. P., Katz, K., Nussinovitch, M., Segev, S., Volovitz, B., and Nussinovitch, N., "Reliability of Ultra-Short ECG Indices for Heart Rate Variability," **Annals of Noninvasive Electrocardiology**, Vol. 16, No. 2, 2011, pp. 117–122, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1542-474X.2011.00417.x.

[48] Shaffer, F. and Ginsberg, J. P., "An Overview of Heart Rate Variability Metrics and Norms," **Frontiers in Public Health**, Vol. 5, Sept. 2017, pp. 258.

[49] Electrophysiology, T. F. o. t. E. S. o. C. t. N. A. S. o. P., "Heart Rate Variability," **Circulation**, Vol. 93, No. 5, March 1996, pp. 1043–1065, Publisher: American Heart Association.

[50] "Heart rate variability," **European Heart Joural**, Vol. 17, 1996, pp. 354–381.

[51] Bühlmann, P. and van de Geer, S., "Lasso for linear models," **Statistics for High-Dimensional Data**, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 7–43, Series Title: Springer Series in Statistics.

[52] Hastie, T., Tibshirani, R., and Friedman, J., **The Elements of Statistical Learning**, Springer Series in Statistics, Springer New York, New York, NY, 2009.

[53] Friedman, J. H., Hastie, T., and Tibshirani, R., "Regularization Paths for Generalized Linear Models via Coordinate Descent," **Journal of Statistical Software**, Vol. 33, Feb. 2010, pp. 1–22.

[54] Simon, N., Friedman, J. H., Hastie, T., and Tibshirani, R., "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent," **Journal of Statistical Software**, Vol. 39, March 2011, pp. 1–13.

[55] Tibshirani, R., "Regression Shrinkage and Selection Via the Lasso," **Journal of the Royal Statistical Society: Series B (Methodological)**, Vol. 58, No. 1, Jan. 1996, pp. 267–288.

[56] Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S., "Cross-validation pitfalls when selecting and assessing regression and classification models," **Journal of Cheminformatics**, Vol. 6, March 2014, pp. 10.

[57] Meinshausen, N., "Relaxed Lasso," **Computational Statistics & Data Analysis**, Vol. 52, No. 1, Sept. 2007, pp. 374–393.

[58] Ezekiel, M., **Methods of correlation analysis.**, Methods of correlation analysis., Wiley, Oxford, England, 1930, Pages: xiv, 427.

[59] Quan, N. T., "The Prediction Sum of Squares as a General Measure for Regression Diagnostics," **Journal of Business & Economic Statistics**, Vol. 6, No. 4, 1988, pp. 501–504, Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[60] Gutzwiller, R. S., Chiou, E. K., Craig, S. D., Lewis, C. M., Lematta, G. J., and Hsiung, C.-P., "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**, Vol. 63, No. 1, Nov. 2019, pp. 217–221.

[61] Hartwich, F., Witzlack, C., Beggiato, M., and Krems, J. F., "The first impression counts – A combined driving simulator and test track study on the development of trust and acceptance of highly automated driving," **Transportation Research Part F: Traffic Psychology and Behaviour**, Vol. 65, Aug. 2019, pp. 522–535.

[62] Morris, D. M., Erno, J. M., and Pilcher, J. J., "Electrodermal Response and Automation Trust during Simulated Self-Driving Car Use," **Proceedings of the Human Factors and Ergonomics Society Annual Meeting**, Vol. 61, No. 1, Sept. 2017, pp. 1759–1762, Publisher: SAGE Publications Inc.

[63] Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., and Yanco, H., "Effects of changing reliability on trust of robot systems," **2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)**, March 2012, pp. 73–80, ISSN: 2167-2148.

[64] Kaniarasu, P., Steinfeld, A., Desai, M., and Yanco, H., "Potential measures for detecting trust changes," **2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)**, March 2012, pp. 241–242, ISSN: 2167-2148.

[65] Nikolaidis, S., Zhu, Y. X., Hsu, D., and Srinivasa, S., "Human-Robot Mutual Adaptation in Shared Autonomy," **Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction**, HRI '17, Association for Computing Machinery, New York, NY, USA, March 2017, pp. 294–302.

[66] Walker, F., Wang, J., Martens, M. H., and Verwey, W. B., "Gaze behaviour and electrodermal activity: Objective measures of drivers' trust in automated vehicles," **Transportation Research Part F: Traffic Psychology and Behaviour**, Vol. 64, July 2019, pp. 401–412.

[67] Körber, M., Baseler, E., and Bengler, K., "Introduction matters: Manipulating trust in automation and reliance in automated driving," **Applied Ergonomics**, Vol. 66, Jan. 2018, pp. 18–31.

[68] Hastie, T., Tibshirani, R., and Tibshirani, R., "Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons," **Statistical Science**, Vol. 35, No. 4, Nov. 2020.

[69] Smith, G., "Step away from stepwise," **Journal of Big Data**, Vol. 5, No. 1, Dec. 2018, pp. 32.

[70] Harrell, F. E., **Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis**, Springer Series in Statistics, Springer International Publishing, Cham, 2015.

[71] Rehmann, A. J., "Handbook of Human Performance Measures and Crew Requirements for Flightdeck Research: (664922007-001)," Tech. rep., American Psychological Association, 1995, Type: dataset.

[72] Guoqiang, S., Wanyan, X., Wu, X., and Zhuang, D., "The Influence of HUD Information Visual Coding on Pilot's Situational Awareness," Aug. 2017, pp. 139–143.

[73] Opmeer, C. H. and Krol, J. P., "Towards an objective assessment of cockpit workload. I. Physiological variables during different flight phases." **Aerospace medicine**, Vol. 44, No. 5, May 1973, pp. 527–32.

[74] So, W. K. Y., Wong, S. W. H., Mak, J. N., and Chan, R. H. M., "An evaluation of mental workload with frontal EEG," **PLOS ONE**, Vol. 12, No. 4, April 2017, pp. e0174949.

[75] Hogervorst, M. A., Brouwer, A.-M., and van Erp, J. B. F., "Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload," **Frontiers in Neuroscience**, Vol. 8, Oct. 2014.

[76] Berka, C., Levendowski, D. J., Davis, G., Whitmoyer, M., Hale, K., and Fuchs, S., "Objective Measures of Situational Awareness Using Neurophysiology Technology," , pp. 11.

[77] Aghajani, H., Garbey, M., and Omurtag, A., "Measuring Mental Workload with EEG+fNIRS," **Frontiers in Human Neuroscience**, Vol. 11, July 2017, pp. 359.

[78] Hirshfield, L., Costa, M., Bandara, D., and Bratt, S., "Measuring Situational Awareness Aptitude Using Functional Near-Infrared Spectroscopy," **Foundations of Augmented Cognition**, edited by D. D. Schmorrow and C. M. Fidopiastis, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2015, pp. 244–255.

[79] Causse, M., Chua, Z., Peysakhovich, V., Del Campo, N., and Matton, N., "Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS," **Scientific Reports**, Vol. 7, No. 1, Dec. 2017, pp. 5222.

Appendix A: Model Generation Exploration

Before settling upon the LASSO procedure for the with interaction models that was describe in Ch. 2 (Fig. 2.7), we explored different pathways of running relaxed LASSO to reduce the overfitting issue.

The general flow for the model exploration done is in Fig. A.1. For each model type, all possible interactions were generated between the available measures. These interactions and the base terms are fed into LASSO. For each of the runs, four models were selected corresponding to: minimum lambda, 1SE lambda, minimum lambda for relaxed LASSO, and 1SE lambda for relaxed LASSO. This was repeated 50 times resulting in 200 models. From these 200 different models, the data was further downselected into three different options: one that contained any variable that showed up at least once in the 200 models, one that contained variables that showed up in one-third of the 200 models, and finally one that contained variables that showed up in one-third of the unique models. The unique models considered that LASSO often selects the same models on different runs; by only considering these it eliminates the bias of variables that in a commonly selected model having a higher chance of being selected again. From each of the three downselected options of predictor variables, LASSO was once again run 50 times, and the same 4 equations were selected.

All together this procedure resulted in 800 models. From these, the unique equations were identified and an OLS was performed to get the coefficients. Finally, their accuracy and predictive capabilities were assessed, and the best model was chosen.

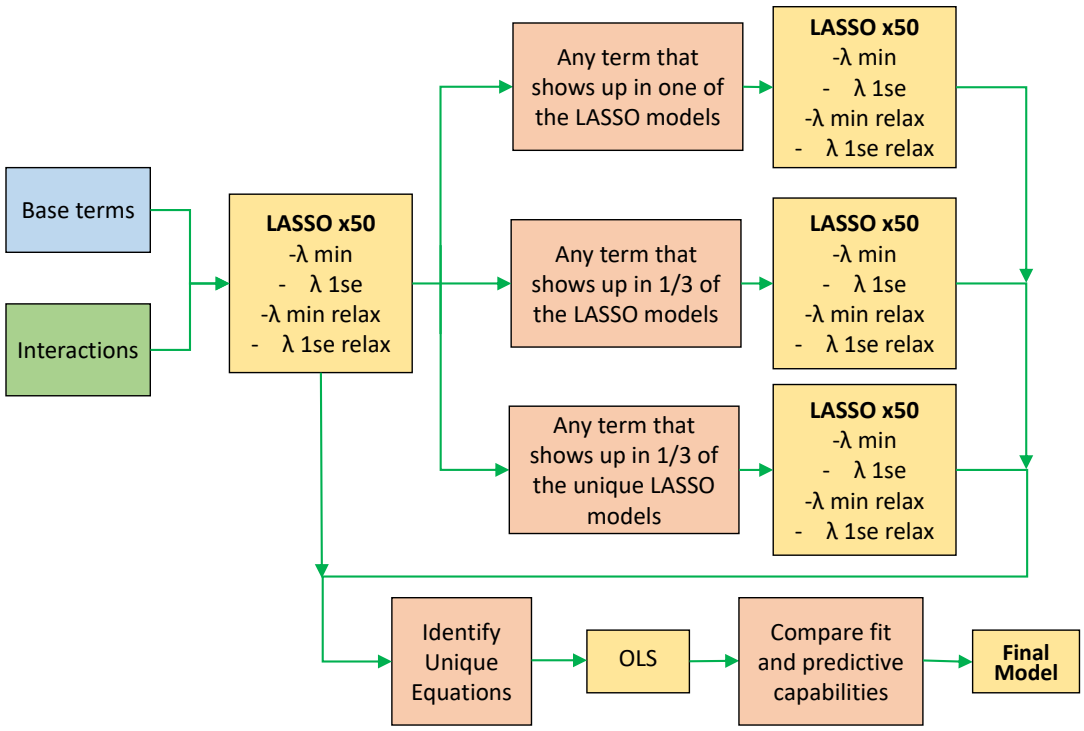


Figure A.1: Flowchart showing the exploratory LASSO based methods used to avoid the issue of overfitting when dealing with interactions

While running this on test data, it was discovered that the best equations were constantly being selected from the results of relaxed LASSO, eliminating the need to run a standard LASSO. In addition, during the second iteration of LASSO, the same equations often showed up in all three model types, and the best equations were all included in the run that included any term. As this was already a small subset of variables, LASSO was able to choose the ones that contributed most to the model and did not need the further downselection done by using the 1/3 criteria. This led to the elimination of these two blocks. Finally, from the second LASSO run through, the best models were always at the 1SE location (as opposed to min)

Appendix B: Additional Trust Models

This appendix includes the coefficients and performance for the trust model with interactions. It also contains the coefficients, performance, and percentage of terms from a certain category for the comparison models. The comparison models include the unimodal embedded measures (UM-EM), unimodal physiological measures (UM-phys), and the models generated to compare the LASSO to stepwise algorithm.

B.1 With Interactions Models

This is the final model for the multimodal trust with interactions.

Table B.1: Coefficients and Performance for Trust with Interactions

	Predictor	Model Type 1	Model Type 2	Model Type 3
Coefficients	(Intercept)	64.65 *	56.03 *	19.00 *
	RSP Rate Before: RSP Amp diff.	-0.16 *	-0.15 *	
	HR before: % of time on Rec. Screen	-1.14e-3	-2.50e-3	
	HR Ratio: Time to Lock In	-4.03 *	-3.02 *	-0.72
	RSP Rate Before: % of Looks on Rec. Screen	-0.13		
	RSP Rate Before: % of Time on Rec. Screen	6.73e-3	6.52e-3	
	RSP Amp Ratio: Time to Lock In	1.59 *	1.21 *	-0.49
	RSP Amp Ratio: % of Looks on Throttle Input	-17.47 *		
	HR RMSSD Before: Prev. Num. of Arcs			6.30e-2
	HR RMSSD After:Prev. Num. of Arcs		0.29 *	3.59e-2
	Pupil Diameter Ratio: Prev. Num. of Arcs		-1.87	
	HR Before: AICP			9.23e-3
	RSP Amp Before: Sleep Rating			0.14
	% of Looks on Rec. Screen: % of Time on Rec. Screen	0.38	0.63 *	
	Num. of Looks on Rec Screen: Expectation		20.48 *	15.15
	% of Time on Rec. Screen: Expectation			0.48
	% of Time on Rec. Screen: Prev. Num. of Arcs		4.82e-2	
	% of Time on Rec. Screen: Sleep Rating			9.47e-2
	% of Time on Rec. Screen: AICP			-1.80e-3
	Prev. Num. of Arcs: Sex			23.44 *
	Age: AICP			4.35e-2 *
	Hours of Sleep: AICP			5.94e-2
	PVT:AICP			2.85e-3
Performance	Adjusted R^2	0.36	0.46	0.72
	RMSE	11.79	10.83	7.77
	Q^2 by subject	0.23	0.39	0.68
	RMSE by subject	13.54	12.06	8.80
	Q^2 by trial	0.34	0.44	0.71
	RMSE by trial	12.02	11.12	8.03

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment
*: $p<.05$

B.2 Comparison Models

This section contains all the comparison models, including unimodal with embedded measures, and unimodal using physiological measures. This includes both the interaction and without interaction models.

B.2.1 Unimodal Embedded Measures

The unimodal embedded measure models did not have any physiological metrics available to them. However, the other metrics (like observable, or demographics) were available during the appropriate model type.

B.2.1.1 Without Interactions Models

Table B.2: Coefficients and Performance for Trust using only Embedded Metrics without Interactions

	Predictor	Model Type		
		1	2	3
Coefficients	(Intercept)	54.12 *	46.14 *	-1.22
	Time to Lock In	-2.64 *	-1.71 *	-1.34 *
	% of Looks on Rec. Screen	5.49	22.99 *	
	% of Time on Rec. Screen	0.35 *	0.24 *	0.15 *
	% of Looks on Throttle Input	-9.64		
	% of Time on Analog Gauge	0.15		
	% of Looks btw Analog Gauge and Throttle Input	-16.87		
	Expectation		41.58 *	46.34 *
	Previous Number of Arcs		9.64 *	6.52 *
	Age			0.74 *
	Sleep Rating			3.88 *
	AICP			2.67 *
Performance	Adjusted R^2	0.27	0.38	0.66
	RMSE	12.62	11.58	8.61
	Q^2 by subject	0.12	0.29	0.65
	RMSE by subject	14.50	13.05	9.18
	Q^2 by trial	0.24	0.36	0.63
	RMSE by trial	12.94	11.84	8.94

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment
*: $p<.05$

Table B.3: Breakdown of Metrics for Trust using Embedded Metrics without Interactions

Metric	Model Type		
	1	2	3
% ECG based			
% RSP based			
% Eye based (phys)			
% EM	16.7	20.0	14.3
% Gaze (EM)	83.3	40.0	14.3
% Observable		40.0	28.6
% Demographic			42.9
Number of Predictors	6	5	7
Number of Sensors	1	1	1

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

B.2.1.2 With Interactions Models

Table B.4: Coefficients and Performance for Trust using only Embedded Metrics with Interactions

	Predictor	Model Type		
		1	2	3
Coefficients	(Intercept)	60.02 *	57.88 *	23.12 *
	Time to Lock In	-0.26	-1.71 *	
	Previous Number of Arcs		8.66	
	Time to Lock In : Num. of Looks on Analog Gauge			1.24e-2
	Time to Lock In % of Looks on Analog Gauge	-5.09		-2.34
	Time to Lock In :% of Time on Analog Gauge	-6.37e-2	0.49 *	-8.34e-3
	% of Looks on Rec. Screen: % of Time on Rec. Screen			
	% of Time on Rec. Screen: % of Time on Analog Gauge	5.39e-3		
	% of Looks on Analog Gauge: % of Time on Rec. Screen	0.52		
	Num. of Looks on Throttle Input: % of Looks on Throttle Input	-16.00		
	% of Looks on Throttle Input: % of Time on Throttle Input	-0.26		
	% of Looks on Throttle Input: % of Looks btw. Analog Gauge and Throttle Input	-38.70	-70.01 *	
	Num. of Looks on PFD: Num of Looks btw. Analog Gauge and Throttle Input	-4.93		
	Num of Looks from Analog Gauge to PFD: Num. of Looks from Throttle Input to Analog Gauge	6.01		

Continuation of Tab B.4				
	Num of Looks from Analog Gauge to PFD: Num. of Looks btw. Analog Gauge and Throttle Input	2.60		
	% of Looks from Throttle Input to Analog Gauge: % of Looks btw. Analog Gauge and Throttle Input	-122.86 *		
	Num. of Looks on Rec Screen: Expectation		24.75 *	-30.27
	% of Time on Rec. Screen: Expectation			1.27
	% of Looks on PFD: Expectation			283.60 *
	Num of Looks between PFD and Analog Gauge: Expectation			19.39
	% of Time on Rec. Screen: Prev. Num. of Arcs		0.03	
	Time to Lock In : Sex			0.13
	% of Time on Rec. Screen: Sleep Rating			8.40e-2 *
	% of Looks from Throttle Input to Analog Gauge: Sex			-4.66
	% of Looks from Throttle Input to Analog Gauge: Robot User			-19.95
	% of Looks btw. PFD and Analog Gauge: AICP			0.22
	Prev. Num of Arcs:e Number of Trials			0.49 *
	Prev. Num. of Arcs: Sex			25.73 *
	Number of Trials : Sex			-0.60 *
	Age: AICP			5.59e-2 *
	Hours of Sleep: AICP			6.78e-2
	PVT: AICP			3.14e-3 *
Performance	Adjusted R^2	0.33	0.42	0.76
	RMSE	12.10	11.23	7.27
	Q^2 by subject	0.20	0.32	0.68
	RMSE by subject	13.79	12.74	8.72
	Q^2 by trial	0.29	0.40	0.72
	RMSE by trial	12.43	11.48	7.81

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

*: $p<.05$

Table B.5: Percentage of predictors in a certain category for Trust using Embedded Metrics with Interactions

Metric	Model Type		
	1	2	3
% ECG based			
% RSP based			
% Eye based (phys)			
% EM	8.7	10.0	11.1
% Gaze (EM)	91.3	60.0	30.6
% Observable		30.0	22.2
% Demographic			36.1
Number of Predictors	12	6	18
Number of Sensors	1	1	1

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

B.2.2 Unimodal Physiological Measures

The unimodal physiological measure models did not have any embedded metrics available to them. However, the other metrics (like observable, or demographics) were available during the appropriate model type.

B.2.2.1 Without Interactions Models

Table B.6: Coefficients and Performance for Trust using only Physiological Sensors without Interactions

	Predictor	Model Type		
		1	2	3
Coefficients	(Intercept)	56.59 *	53.55 *	-18.39 *
	HR Before			0.18 *
	RSP Amp Diff.	-2.69 *	-1.84 *	-0.73
	Expectation		50.26 *	54.66 *
	Previous Number of Arcs		9.78 *	4.81 *
	Age			0.55 *
	Sleep Rating			5.41 *
	AICP			2.89 *
Performance	Adjusted R^2	0.11	0.28	0.64
	RMSE	13.92	12.54	8.80
	Q^2 by subject	0.04	0.24	0.62
	RMSE by subject	15.17	13.43	9.56
	Q^2 by trial	0.11	0.26	0.62
	RMSE by trial	13.99	12.74	9.07

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment
*: $p<.05$

Table B.7: Percentage of predictors in a certain category for Trust using Physiological Sensors without Interactions

Metric	Model Type		
	1	2	3
% ECG based	0.0	0.0	14.3
% RSP based	100.0	33.3	14.3
% Eye based (phys)	0.0	0.0	0.0
% EM			
% Gaze (EM)			
% Observable		66.7	28.6
% Demographic			42.9
Number of Predictors	1	3	7
Number of Sensors	1	1	2

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

B.2.2.2 With Interactions Models

Table B.8: Coefficients and Performance for Trust using only Physiological Sensors with Interactions

	Predictor	Model Type		
		1	2	3
Coefficients	(Intercept)	56.20 *	53.35 *	8.69 *
	Number of Arcs		-0.43	
	RSP Rate Before:RSP Amp diff.	-0.14 *	-0.11 *	
	HRV RMSSD Before:Expectation		1.21 *	
	HR Ratio:Prev. Num. of Arcs		4.82	
	HR RMSSD After:Prev. Num. of Arcs		0.18	9.42e-2
	Pupil Diameter Ratio:Prev. Num. of Arcs		-0.61	
	HR Before:AICP			5.50e-3
	RSP Amp Before:Sleep Rating			0.47 *
	Expectation:PVT			0.22 *
	Prev. Num. of Arcs:Sex			24.68 *
	Age:AICP			3.42e-2 *
	Hours of Sleep:AICP			0.18 *
	PVT:AICP			3.22e-3 *
Performance	Adjusted R^2	0.14	0.29	0.70
	RMSE	13.71	12.41	8.14
	Q^2 by subject	0.07	0.27	0.66
	RMSE by subject	14.93	13.23	9.01
	Q^2 by trial	0.14	0.28	0.68
	RMSE by trial	13.76	12.55	8.34

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

*: $p<.05$

Table B.9: Percentage of predictors in a certain category for Trust using Physiological Sensors with Interactions

Metric	Model Type		
	1	2	3
% ECG based	0.0	27.3	12.5
% RSP based	100.0	18.2	6.3
% Eye based (phys)	0.0	9.1	0.0
% EM			
% Gaze (EM)			
% Observable		45.5	18.8
% Demographic			62.5
Number of Predictors	1	6	8
Number of Sensors	1	3	2

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

B.2.3 LASSO and Stepwise Comparisons

These models were created given the same initial predictor variables. The stepwise models are from Kintz [26].

B.2.3.1 LASSO

Table B.10: Coefficients and Performance for Trust using LASSO for a LASSO Stepwise comparison

	Predictor	Model Type		
		1	2	3
Coef.	(Intercept)	67.23 *	53.99 *	16.24 *
	Time to Lock In	-2.06 *		
	Previous Number of Arcs		12.21 *	10.31 *
	AICP			2.92 *
Performance	Adjusted R^2	0.08	0.18	0.48
	RMSE	14.13	13.34	10.61
	Q^2 by subject	-0.02	0.19	0.49
	RMSE by subject	15.59	13.89	11.08
	Q^2 by trial	0.07	0.18	0.48
	RMSE by trial	14.26	13.37	10.72

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment
*: $p<.05$

Table B.11: Percentage of predictors in a certain category for Trust using LASSO for a LASSO Stepwise comparison

Metric	Model Type		
	1	2	3
% ECG based			
% RSP based			
% Eye based (phys)			
% EM	100.0	0.0	0.0
% Gaze (EM)			
% Observable		100.0	50.0
% Demographic			50.0
Number of Predictors	1	1	1
Number of Sensors	0	0	0

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

B.2.3.2 Stepwise

Table B.12: Coefficients and Performance for Trust using the Stepwise Model [26]

	Predictor	Model Type		
		1	2	3
Coefficients	(Intercept)	67.23 *	72.89 *	30.01 *
	Time to Lock In	-2.06 *	-3.21 *	-1.61 *
	Expectation		38.24 *	294.37 *
	Previous Number of Arcs		12.60 *	41.06 *
	Number of Trials		-1.98 *	-2.56 *
	AICP			2.64 *
	Time to Lock In :Expectation			-9.64 *
	Time to Lock In :Num. of Trials		0.27	
	Time to Lock In :Prev. Num. of Arcs			3.03 *
	Expectation :Prev. Num. of Arcs			-54.94
	Expectation :AICP			-13.55 *
	Prev. Num. of Arcs:AICP			-3.12 *
	Num. of Trials:AICP			0.15
Performance	Adjusted R^2	0.08	0.30	0.65
	RMSE	14.13	12.37	8.79
	Q^2 by subject	-0.02	0.17	0.48
	RMSE by subject	15.59	14.03	11.15
	Q^2 by trial	0.07	0.26	0.55
	RMSE by trial	14.26	12.75	9.97

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

*: $p<.05$

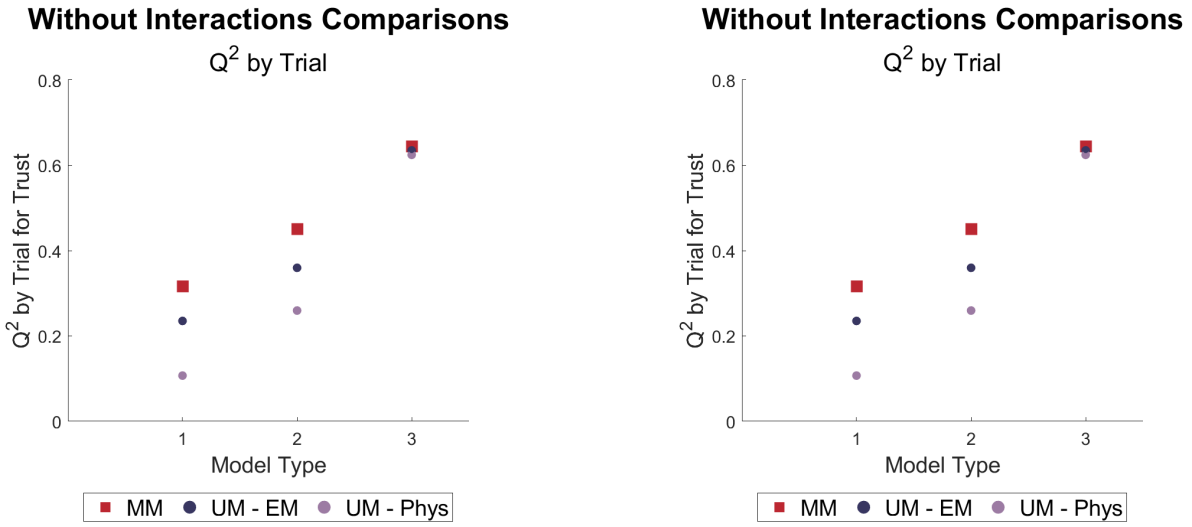
Table B.13: Percentage of predictors in a certain category for Trust using the Stepwise Model [26]

Metric	Model Type		
	1	2	3
% ECG based			
% RSP based			
% Eye based (phys)			
% EM	100.0	28.6	17.6
% Gaze (EM)			
% Observable		71.4	58.8
% Demographic			23.5
Number of Predictors	1	5	11
Number of Sensors	0	0	0

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

B.3 Additional Comparison Plots

B.3.1 Q^2 by Trial



(a) Comparison of trust models without interaction on the basis of Q^2 by trial

(b) Comparison of trust models with interaction on the basis of Q^2 by trial

Figure B.1: Q^2 by trial comparisons

B.3.2 Number of Predictors and Sensors

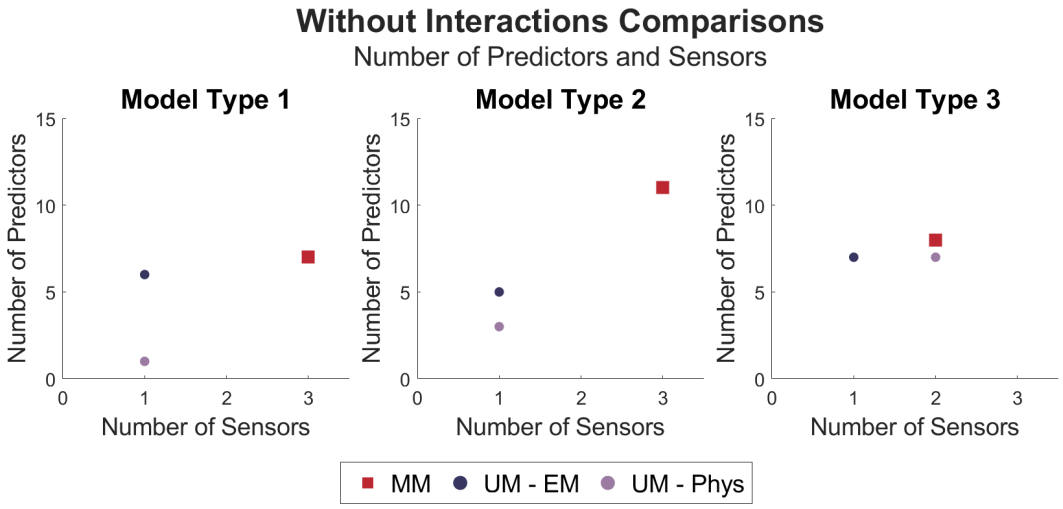


Figure B.2: The comparison of trust models without interactions to other similar models. The x-axis is the number of sensors required (of ECG, RSP, EYE) and the y axis is the number of predictors in the models.

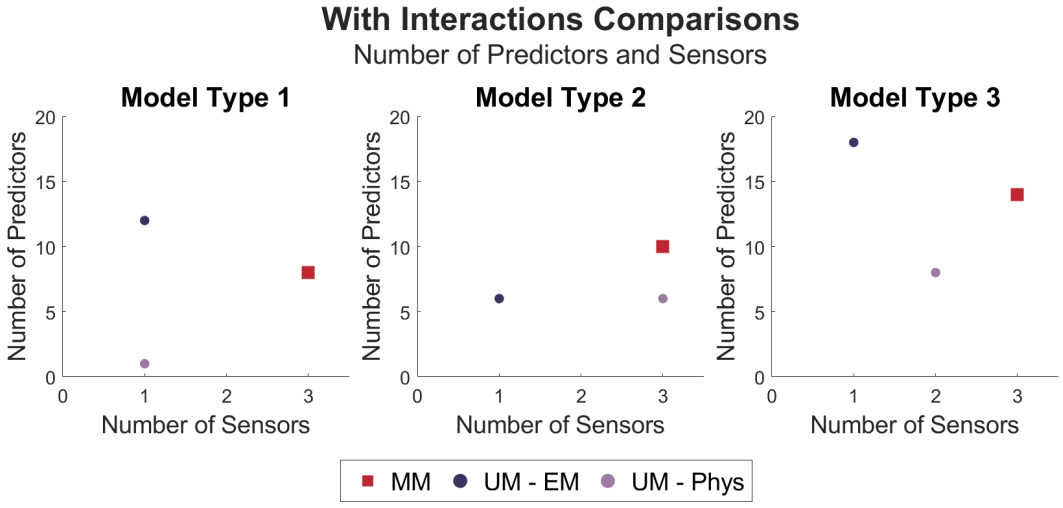


Figure B.3: The comparison of trust models with interactions to other similar models. The x-axis is the number of sensors required (of ECG, RSP, EYE) and the y axis is the number of predictors in the models.

Appendix C: Additional Situation Awareness and Workload Models

This appendix includes the coefficients and performance for the SA and WL model with interactions. It also contains the coefficients, performance, and percentage of terms from a certain category for the comparison models. The comparison models include the unimodal simultaneously fit with embedded measures (UMSF-EM), unimodal simultaneously fit physiological measures (UMSF-phys), multimodal independently fit to SA (MMIF-SA), and multimodal independently fit to WL (MMIF-WL)

C.1 With Interactions Models

These are the models for multimodal simultaneously fit SA and WL with interactions.

Table C.1: Coefficients and Performance for SA and WL models with Interactions

	Predictor	Model Type 1		Model Type 2		Model Type 3	
		SA	WL	SA	WL	SA	WL
Coefficients	(Intercept)	10.16 *	6.88 *	28.12 *	3.06 *	12.41 *	2.97 *
	RSP Rate	0.45 *	-7.64e-2 *				
	Joystick Input			-4.77e-3 *	2.27e-3 *		
	RSP Rate: % Callouts Made	3.97e-3 *	-8.96e-4 *			8.03e-4	-6.07e-4 *
	HRV MeanNN: Lighting %	-4.44e-5	2.38e-5				
	HRV SDNN: Lighting %	-8.66e-4	2.54e-4				
	HRV MeanNN: Joystick input					-4.47e-6	2.94e-6 *
	HRV SDNN: Joystick Input					-7.20e-5	1.82e-6
	HRV MeanNN: RMS track. err.			-1.26e-2 *	4.08e-3 *		
	HR: Task load			-3.25e-3	-8.51e-3		
	HRV CVNN: Task load			-10.19	7.36		
	HRV pNN20: Task load			-3.46e-2	1.04e-2	2.13e-3	1.28e-2
	RSP Rate: Number of Trials					1.21e-2 *	-2.20e-3
	RSP Rate: Display Skill					3.13e-2	2.93e-2

Continuation of Tab C.1							
Coefficients	Mean Pupil Dia.: Display Skill					0.20	-0.24 *
	Lighting % : Joystick Input					-1.74e-5	-1.03e-5
	RMS Track. err: Age					4.94e-2	0.12 *
	Task load : Hours of Sleep					-0.39 *	-9.40e-2
	Task load : Sleep Rating					-0.24	0.39 *
	Num. of Trials: Display Skill					4.88e-2	-4.90e-3
	Age: Hour of Sleep					3.67e-3	1.85e-2 *
	Age : PVT					9.47e-4 *	1.83e-4
	Hours of Sleep: PVT					1.04e-3	-1.22e-3 *
Performance	Adjusted R^2	0.25	0.23	0.50	0.39	0.68	0.59
	RMSE	5.45	1.70	4.46	1.51	3.58	1.24
	Q^2 by subject	0.17	0.15	0.46	0.31	0.56	0.44
	RMSE by subject	5.81	1.82	4.68	1.64	4.24	1.48
	Q^2 by trial	0.24	0.21	0.49	0.37	0.65	0.55
	RMSE by trial	5.50	1.72	4.53	1.54	3.71	1.30

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment
*: $p<.05$

C.2 Comparison Models

This section contains all the comparison models, including unimodal simultaneously fit with embedded measures, unimodal simultaneously fit using physiological measures, multimodal independently fit to SA, and multimodal independently fit to WL. This includes both the interaction and non interaction models.

C.2.1 Multimodal Independently Fit - Situation Awareness

These models were fit with all the predictor terms available, but only to SA. The WL portion is calculated from the given SA predictors for comparison purposes.

C.2.1.1 Without Interactions Models

Table C.2: Coefficients and Performance for SA and WL models Independently fit to SA without Interactions

	Predictor	Model Type 1		Model Type 2		Model Type 3	
		SA	WL	SA	WL	SA	WL
Coefficients	(Intercept)	4.30	7.86 *	6.36	2.93 *	4.46	0.94
	RSP Rate	0.69 *	-0.14 *	0.45	-4.65e-2	8.17e-2	4.09e-3
	HR			7.23e-2 *	1.19e-2	6.37e-2	2.26e-2
	HRV SDNN					-6.66e-2 *	6.97e-3
	% Callouts made	9.18e-2 *	-1.71e-2 *	4.31e-2 *	-1.12e-3		
	Lighting %	-6.80e-2 *	3.28e-2 *	-2.54e-2	2.04e-2		
	Joystick Input			-2.59e-3	1.71e-3 *	-8.33e-3 *	2.40e-3 *
	RMS Tracking Error			-5.85 *	2.50 *		
	Task load			-3.37 *	0.31	-2.88 *	0.61 *
	Number of Trials			0.32 *	-3.57e-2	0.37 *	-5.30e-2
	Age					0.18	0.15 *
	Sex					-1.69 *	0.23
	Display Skill					2.01 *	-0.71 *
	Hours of Sleep					0.56	6.21e-2
	PVT					2.29e-2	-9.17e-3
Performance	Adjusted R^2	0.24	0.26	0.57	0.48	0.67	0.54
	RMSE	5.48	1.67	4.11	1.40	3.62	1.31
	Q^2 by subject	0.16	0.22	0.45	0.32	0.53	0.21
	RMSE by subject	5.84	1.75	4.72	1.63	4.39	1.75
	Q^2 by trial	0.23	0.25	0.55	0.46	0.65	0.51
	RMSE by trial	5.54	1.69	4.21	1.43	3.73	1.35

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment
*: $p<.05$

Table C.3: Percentage of predictors in a certain category for SA and WL Independently fit to SA without Interactions

Metric	Model Type		
	1	2	3
% ECG based	0.0	12.5	18.2
% RSP based	33.3	12.5	9.1
% Eye based (phys)	0.0	0.0	0.0
% EM	66.7	25.0	0.0
% Observable		50.0	27.3
% Demographic			45.5
Number of Predictors	3	8	11
Number of Sensors	1	2	2

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

C.2.1.2 With Interactions Models

Table C.4: Coefficients and Performance for SA and WL models Independently fit to SA with Interactions

	Predictor	Model Type 1		Model Type 2		Model Type 3	
		SA	WL	SA	WL	SA	WL
Coefficients	(Intercept)	-552.15 *	138.78	16.36 *	3.42 *	13.67 *	2.88 *
	RSP Rate	0.39 *	-6.77e-2	0.24 *	2.36e-2		
	Joystick Input			5.24e-3	6.24e-3 *	6.06e-3	5.19e-3 *
	HR : HRV MeanNN	8.45e-3	-2.12e-3				
	HR: HRV MedianNN	9.31e-4	-8.16e-5				
	RSP Rate: % Callouts Made	4.15e-3 *	-9.41e-4 *	1.90e-3 *	-2.51e-4	1.19e-3	-3.88e-4
	HRV MeanNN: Lighting %	-5.61e-5	2.29e-5	-1.09e-5	-6.02e-6 *		
	HRV SDNN: Lighting %	-5.24e-4	2.39e-4				
	HRV MeanNN: Joystick input			-9.04e-3 *	3.76e-3 *	-1.79e-5 *	-4.39e-6
	HRV MeanNN: RMS track. err.					1.13e-3	2.02e-3
	RSP Rate: Task load			-8.54e-2	-2.45e-2		
	HRV CVNN: Task load			-16.05	-0.10	-14.30	3.84

Continuation of Tab C.4							
Coefficients	HRV pNN20: Task load			-1.08e-2	1.99e-2 *	9.23e-3	2.20e-2 *
	RSP Rate: Number of Trials			1.52e-2 *	-1.71e-3		
	RSP Rate: Display Skill					0.14 *	3.38e-2
	Task load : Hours of Sleep					-0.37	-0.11
	Num. of Trials: Display Skill					0.23 *	-1.71e-2
	Age: Display Skill					-4.04e-2	4.04e-2 *
	Age : PVT					1.31e-3 *	3.35e-4 *
	Display Skill: PVT					-5.01e-3	-8.88e-3 *
Performance	Adjusted R^2	0.11	0.21	0.51	0.45	0.66	0.59
	RMSE	5.94	1.72	4.41	1.44	3.68	1.24
	Q^2 by subject	0.04	0.19	0.44	0.38	0.53	0.48
	RMSE by subject	6.25	1.77	4.76	1.56	4.36	1.42
	Q^2 by trial	0.10	0.20	0.49	0.43	0.64	0.57
	RMSE by trial	5.98	1.73	4.53	1.46	3.77	1.27

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment
*: p<.05

Table C.5: Percentage of predictors in a certain category for SA and WL Independently fit to SA with Interactions

Metric	Model Type		
	1	2	3
% ECG based	54.5	25.0	17.4
% RSP based	18.2	25.0	8.7
% Eye based (phys)	0.0	0.0	0.0
% EM	27.3	6.3	4.3
% Observable		43.8	30.4
% Demographic			39.1
Number of Predictors	2	6	9
Number of Sensors	2	2	2

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

C.2.2 Multimodal Independently Fit - Workload

These models were fit with all the predictor terms available, but only to WL. The SA portion is calculated from the given SA predictors for comparison purposes.

C.2.2.1 Without Interactions Models

Table C.6: Coefficients and Performance for SA and WL models Independently fit to WL without Interactions

	Predictor	Model Type 1		Model Type 2		Model Type 3	
		SA	WL	SA	WL	SA	WL
Coefficients	(Intercept)	1.64	5.51 *	31.60 *	2.12 *	7.41	3.12 *
	RSP Rate	0.70 *	-0.13 *				
	HR	2.93e-2	2.58e-2 *				
	% Callouts made	9.39e-2 *	-1.53e-2 *				
	Lighting %	-6.94e-2 *	3.15e-2 *	-1.71e-2	1.94e-2 *	-2.37e-2	9.77e-3 *
	Joystick Input			-1.02e-2 *	2.41e-3	-9.49e-3 *	1.50e-3 *
	RMS Tracking Error			-9.56 *	3.01	0.33	1.90 *
	Task load					-2.22 *	0.53 *
	Number of Trials						
	Age					0.43 *	0.13 *
	Display Skill					2.15 *	-0.39 *
	PVT					4.55e-2 *	-5.90e-3
	Handedness					-3.86 *	-1.11 *
Performance	Adjusted R^2	0.24	0.27	0.45	0.47	0.63	0.57
	RMSE	5.49	1.65	4.65	1.40	3.83	1.27
	Q^2 by subject	0.09	0.19	0.37	0.44	0.51	0.47
	RMSE by subject	6.09	1.78	5.06	1.48	4.49	1.43
	Q^2 by trial	0.22	0.25	0.43	0.46	0.62	0.55
	RMSE by trial	5.57	1.68	4.75	1.42	3.91	1.30

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment
*: $p<.05$

Table C.7: Percentage of predictors in a certain category for SA and WL Independently fit to WL without Interactions

Metric	Model Type		
	1	2	3
% ECG based	25.0	0.0	0.0
% RSP based	25.0	0.0	0.0
% Eye based (phys)	0.0	0.0	0.0
% EM	50.0	33.3	12.5
% Observable		66.7	37.5
% Demographic			50.0
Number of Predictors	4	3	8
Number of Sensors	2	0	0

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

C.2.2.2 With Interactions Models

Table C.8: Coefficients and Performance for SA and WL models Independently fit to WL with Interactions

	Predictor	Model Type 1		Model Type 2		Model Type 3	
		SA	WL	SA	WL	SA	WL
Coefficients	(Intercept)	8.44 *	7.18 *	26.27 *	2.98 *	19.79 *	4.23 *
	Lighting %	-0.37 *	4.82e-2	-1.49e-2	1.88e-2 *		
	Age					0.19	0.12 *
	RSP Rate: HRV MeanNN	6.54e-4 *	-1.36e-4 *				
	RSP Amp: HRV pNN20					-8.41e-3	1.96e-3
	RSP Amp: HRV pNN50					-2.07e-3	9.14e-3 *
	RSP Rate % Callouts made	5.02e-3 *	-4.90e-4				
	HRV RMSSD: % Callouts made	-0.35 *	7.77e-2				
	HRV SDSD: % Callouts made	0.35 *	-7.74e-2				
	HR: Lighting %	3.83e-3 *	-3.03e-4				
	HRV IQRNN: Lighting %	-2.02e-4	1.44e-4				
	HR: Joystick Input			-2.91e-5	1.57e-5	-8.56e-5 *	6.19e-6

Continuation of Tab C.8

Coefficients	HR: RMS Track. Err.				-5.21e-2	2.19e-2		
	HRV IQRNN: Task load				-8.03e-3	1.92e-3		
	HRV pNN50: Task load				4.18e-2	1.65e-2	1.92e-2	3.05e-3
	HRV pNN20: Task load				-7.13e-2 *	2.27e-3		
	Mean Pupil Dia.: Number of Trials						0.12 *	-2.01e-2
	RSP Amp: Video Games Rating							
	HRV CVSD: PVT							
	HRV HTI: PVT							
	% Callouts made: PVT						2.34e-5	-4.86e-5 *
	Lighting %: Video Games Rating						-1.10e-2	7.59e-3 *
	Lighting %: Age						8.29e-4	-3.16e-4
	Joystick Input : RMS Track. Err.				-1.40e-3	5.98e-4	2.24e-3	2.47e-3
	RMS Track. Err.: Age						-0.28	-1.69e-2
	Task load: Video Games Rating						-0.80 *	0.21 *
	Task load: Sleep Rating						-1.45 *	0.35 *
	Number of Trials: Handedness						-0.34	2.76e-2
	Display Skill: Video Games Rating						-0.51	-0.32 *
	Display Skill : PVT						1.42e-2 *	7.11e-4
	PVT: Handedness						-3.67e-3	-6.58e-3 *

Continuation of Tab C.8							
Performance	Adjusted R^2	0.30	0.28	0.47	0.48	0.62	0.65
	RMSE	5.28	1.65	4.58	1.39	3.88	1.15
	Q^2 by subject	0.22	0.14	0.36	0.34	-0.18	0.52
	RMSE by subject	5.63	1.83	5.10	1.60	6.93	1.36
	Q^2 by trial	0.27	0.25	0.45	0.46	0.59	0.60
	RMSE by trial	5.38	1.68	4.68	1.43	4.06	1.22

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment
*: p<.05

Table C.9: Percentage of predictors in a certain category for SA and WL Independently fit to WL with Interactions

Metric	Model Type		
	1	2	3
% ECG based	38.5	38.5	12.1
% RSP based	15.4	0.0	6.1
% Eye based (phys)	0.0	0.0	3.0
% EM	46.2	7.7	9.1
% Observable		53.1	27.3
% Demographic			42.4
Number of Predictors	7	7	17
Number of Sensors	2	1	3

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

C.2.3 Unimodal Simultaneously Fit - Embedded Measures

The unimodal simultaneously fit embedded measure models did not have any physiological metrics available to them. However, the other metrics (like observable, or demographics) were available during the appropriate model type. It was fit to both SA and WL together.

C.2.3.1 Without Interactions Models

Table C.10: Coefficients and Performance for SA and WL models using only Embedded Measures without Interactions

	Predictor	Model Type 1		Model Type 2		Model Type 3	
		SA	WL	SA	WL	SA	WL
Coefficients	(Intercept)	25.47 *	3.77 *	26.19 *	2.60 *	3.31	0.72
	Lighting %	-6.24e-2 *	3.14e-2 *	-8.40e-3	1.85e-2 *		
	Joystick Input			-4.78e-3 *	2.09e-3 *	-7.15e-3 *	2.06e-3 *
	RMS Tracking Error			-7.38 *	2.89 *	1.58	2.55 *
	Task load			-2.89 *	0.18	-3.46 *	0.42 *
	Number of Trials			0.31 *	-4.40e-2	0.32 *	-6.29e-2 *
	Age					0.30 *	0.17 *
	Display Skill					1.94 *	-0.67 *
	Hours of Sleep					0.41	0.18
	PVT					3.11e-2 *	-6.90e-3 *
Performance	Adjusted R^2	0.07	0.19	0.51	0.47	0.64	0.57
	RMSE	6.08	1.74	4.40	1.40	3.77	1.27
	Q^2 by subject	0.04	0.18	0.45	0.40	0.54	0.46
	RMSE by subject	6.24	1.79	4.73	1.53	4.32	1.45
	Q^2 by trial	0.06	0.19	0.49	0.46	0.63	0.55
	RMSE by trial	6.11	1.75	4.49	1.43	3.85	1.30

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

*: $p<.05$

Table C.11: Percentage of predictors in a certain category for SA and WL using only Embedded Measures without Interactions

Metric	Model Type		
	1	2	3
% ECG based			
% RSP based			
% Eye based (phys)			
% EM	100.0	20.0	0.0
% Observable		80.0	50.0
% Demographic			50.0
Number of Predictors	1	5	8
Number of Sensors	0	0	0

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

C.2.3.2 With Interactions Models

Table C.12: Coefficients and Performance for SA and WL models using only Embedded Measures with Interactions

	Predictor	Model Type 1		Model Type 2		Model Type 3	
		SA	WL	SA	WL	SA	WL
Coefficients	(Intercept)	18.63 *	6.16 *	25.78 *	3.42 *	15.24 *	1.54 *
	% Callouts made	0.10 *	-3.42e-2 *				
	Joystick Input			-4.16e-3	9.98e-4	-6.21e-3 *	1.67e-3 *
	RMS Tracking Error			-7.57 *	3.00 *	1.25	2.55 *
	Task load			-2.16	-0.32		
	Number of Trials			0.31 *	-5.41e-2		
	% Callouts made : Lighting %	-8.56e-4 *	4.54e-4 *				
	Lighting % : Joystick Input			-1.14e-5	2.49e-5 *		
	% Callouts made: Task load			-1.20e-2	8.69e-3		
	Task load : Hours of Sleep					-0.49 *	7.03e-2 *
	Num. of Trials: Display Skill					0.18 *	-2.63e-2
	Age: Display Skill					0.34 *	-0.15 *
	Age: Hour of Sleep					-7.32e-2 *	5.44e-2 *
	Age : PVT					1.60e-3 *	-1.37e-4
	Display Skill: PVT					-2.57e-2 *	1.06e-2 *
	Hours of Sleep: PVT					5.45e-3 *	-3.14e-3 *
Performance	Adjusted R^2	0.11	0.21	0.51	0.45	0.66	0.59
	RMSE	5.94	1.72	4.41	1.44	3.68	1.24
	Q^2 by subject	0.04	0.19	0.44	0.38	0.53	0.48
	RMSE by subject	6.25	1.77	4.76	1.56	4.36	1.42
	Q^2 by trial	0.10	0.20	0.49	0.43	0.64	0.57
	RMSE by trial	5.98	1.73	4.53	1.46	3.77	1.27

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

*: $p<.05$

Table C.13: Percentage of predictors in a certain category for SA and WL using only Embedded Measures with Interactions

Metric	Model Type		
	1	2	3
% ECG based			
% RSP based			
% Eye based (phys)			
% EM	100.0	25.0	0.0
% Observable		75	25.0
% Demographic			75.0
Number of Predictors	2	6	9
Num. of Sensors	0	0	0

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

C.2.4 Unimodal Simultaneously Fit - Physiological Measures

The unimodal simultaneously fit physiological measure models did not have any embedded measures available to them. However, the other metrics (like observable, or demographics) were available during the appropriate model type. It was fit to both SA and WL together.

C.2.4.1 Without Interactions Models

Table C.14: Coefficients and Performance for SA and WL Models using only Physiological Measures without Interactions

	Predictor	Model Type 1		Model Type 2		Model Type 3	
		SA	WL	SA	WL	SA	WL
Coefficients	(Intercept)	13.51 *	6.94 *	27.78 *	2.95 *	5.42	1.73
	RSP Rate	0.34 *	-1.73e-2			6.72e-2	5.97e-3
	HRV pNN20	3.30e-2	-2.80e-2 *				
	Joystick Input			-5.19e-3 *	2.31e-3 *	-6.74e-3 *	2.04e-3 *
	RMS Tracking Error			-7.64 *	3.24 *	1.20	2.47 *
	Task load			-2.63 *	0.22	-3.53 *	0.44 *
	Number of Trials					0.32 *	-6.30e-2 *
	Age					0.26 *	0.16 *
	Display Skill					2.14 *	-0.57 *
	PVT					3.14e-2 *	-5.80e-3
Performance	Adjusted R^2	0.04	0.05	0.49	0.41	0.64	0.56
	RMSE	6.16	1.89	4.51	1.49	3.79	1.28
	Q^2 by subject	0.02	-0.11	0.45	0.35	0.53	0.45
	RMSE by subject	6.32	2.08	4.74	1.59	4.37	1.47
	Q^2 by trial	0.04	0.04	0.47	0.40	0.62	0.54
	RMSE by trial	6.20	1.90	4.58	1.51	3.87	1.31

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

*: $p<.05$

Table C.15: Percentage of predictors in a certain category for SA and WL using only Physiological Measures without Interactions

Metric	Model Type		
	1	2	3
% ECG based	50.0	0.0	0.0
% RSP based	50.0	0.0	12.5
% Eye based (phys)	0.0	0.0	0.0
% EM			
% Observable		100.0	50.0
% Demographic			37.5
Number of Predictors	2	3	8
Number of Sensors	2	0	1

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

C.2.4.2 With Interactions Models

Table C.16: Coefficients and Performance for SA and WL models using only Physiological with Interactions

	Predictor	Model Type 1		Model Type 2		Model Type 3	
		SA	WL	SA	WL	SA	WL
Coefficients	(Intercept)	-556.38 *	147.06 *	25.55 *	3.49 *	12.37 *	1.95 *
	RSP Rate	0.25 *	1.85e-2 *				
	Joystick Input			-3.79e-3 *	2.04e-3 *		
	RSP Rate: HRV pNN20	1.15e-3 *	-1.04e-3 *				
	HR : HRV MeanNN	8.37e-3 *	-2.22e-3 *				
	HR: HRV MedianNN	1.17e-3 *	-1.35e-4 *				
	RSP Amp : Joystick Input					-7.13e-4 *	9.48e-5 *
	HRV MeanNN: Joystick input					-3.93e-6 *	3.07e-6 *
	HRV SDNN: Joystick Input					-8.50e-5 *	-3.81e-6 *
	HRV MeanNN: RMS track. err.			-1.26e-2 *	3.95e-3 *		
	RSP Rate: Task load			-6.28e-2 *	-1.77e-2 *	-0.13 *	-7.71e-3 *
	RSP Rate: Number of Trials			1.73e-2 *	-2.26e-3 *	1.63e-2 *	-7.72e-4 *

Continuation of Tab C.16							
Coefficients	HRV CVNN: Task load			-12.24 *	3.69 *		
	HRV pNN20: Task load			-1.99e-2 *	1.15e-2 *		
	RSP Rate: Display Skill					5.01e-2 *	1.45e-2 *
	Mean Pupil Dia.: Display Skill					0.14 *	-0.22 *
	RMS track.err: Age					8.05e-2 *	0.12 *
	Task load : Hours of Sleep					0.11 *	-1.02e-2 *
	Task load : Sleep Rating					-0.72 *	0.48 *
	Num. of Trials: Display Skill					4.36e-2 *	2.78e-2 *
	Num. of Trials: Sleep Rating					-3.99e-2 *	-6.72e-2 *
	Age: Hour of Sleep					-2.46e-4 *	2.23e-2 *
	Age : PVT					1.16e-3 *	-1.72e-4 *
	Hours of Sleep: PVT					1.33e-3 *	-3.48e-4 *
Performance	Adjusted R^2	0.07	0.04	0.55	0.39	0.68	0.59
	RMSE	6.07	1.90	4.25	1.51	3.54	1.24
	Q^2 by subject	0.00	-0.14	0.50	0.31	0.58	0.43
	RMSE by subject	6.40	2.11	4.51	1.65	4.16	1.50
	Q^2 by trial	0.05	0.01	0.53	0.37	0.66	0.55
	RMSE by trial	6.14	1.93	4.33	1.54	3.67	1.30

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

*: $p<.05$

Table C.17: Percentage of predictors in a certain category for SA and WL using only Physiological Measures with Interactions

Metric	Model Type		
	1	2	3
% ECG based	71.4	27.3	6.7
% RSP based	28.6	18.2	13.3
% Eye based (phys)	0.0	0.0	3.3
% EM			
% Observable		54.6	33.3
% Demographic			43.3
Number of Predictors	4	6	15
Number of Sensors	2	2	3

Note: If a predictor was not available to the LASSO algorithm that cell is darkened. If a predictor was available, but not selected that cell is empty. Purple is physiological data, Yellow is embedded measures, Red is observable information, and blue is participant-based information obtained before the experiment

C.3 Additional Comparison Plots

C.3.1 Q^2 by Trial

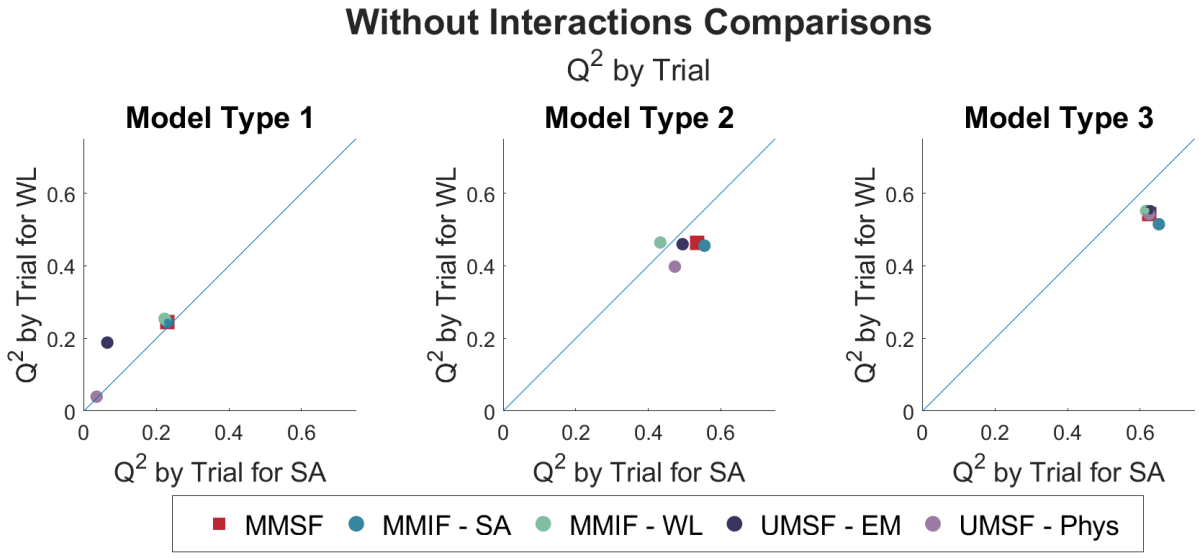


Figure C.1: The comparison of SA and WL models without interactions to other similar models using Q^2 by trial. The x-axis represents the value for SA, and the y-axis is for WL. The blue line represents models that are equally good at SA and WL.

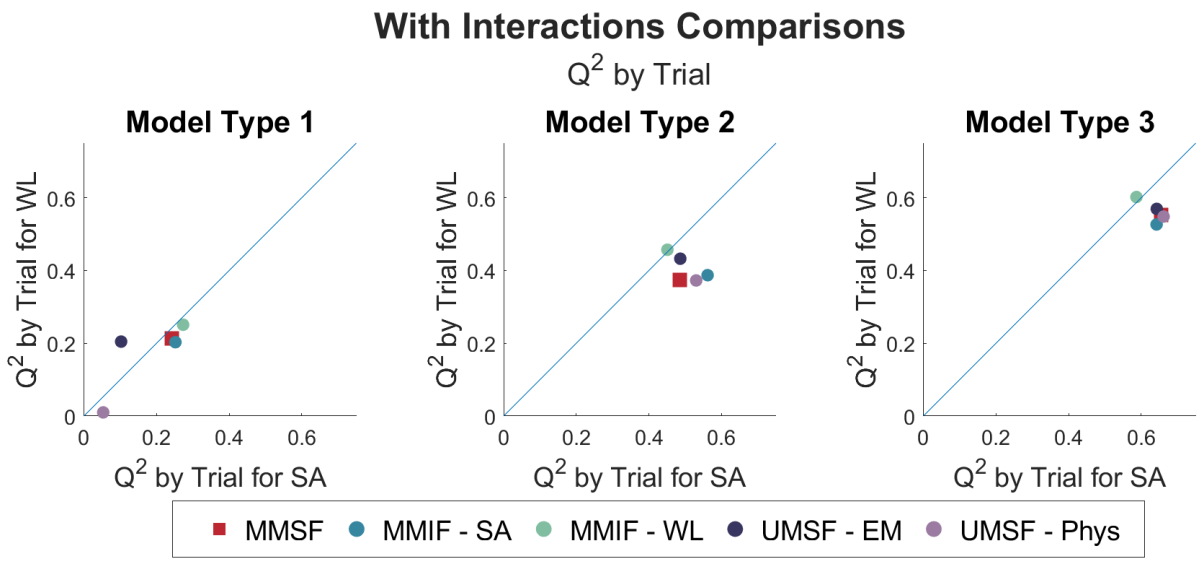


Figure C.2: The comparison of SA and WL models with interactions to other similar models using Q^2 by trial. The x-axis represents the value for SA, and the y-axis is for WL. The blue line represents models that are equally good at SA and WL.

C.3.2 Number of Predictors and Sensors

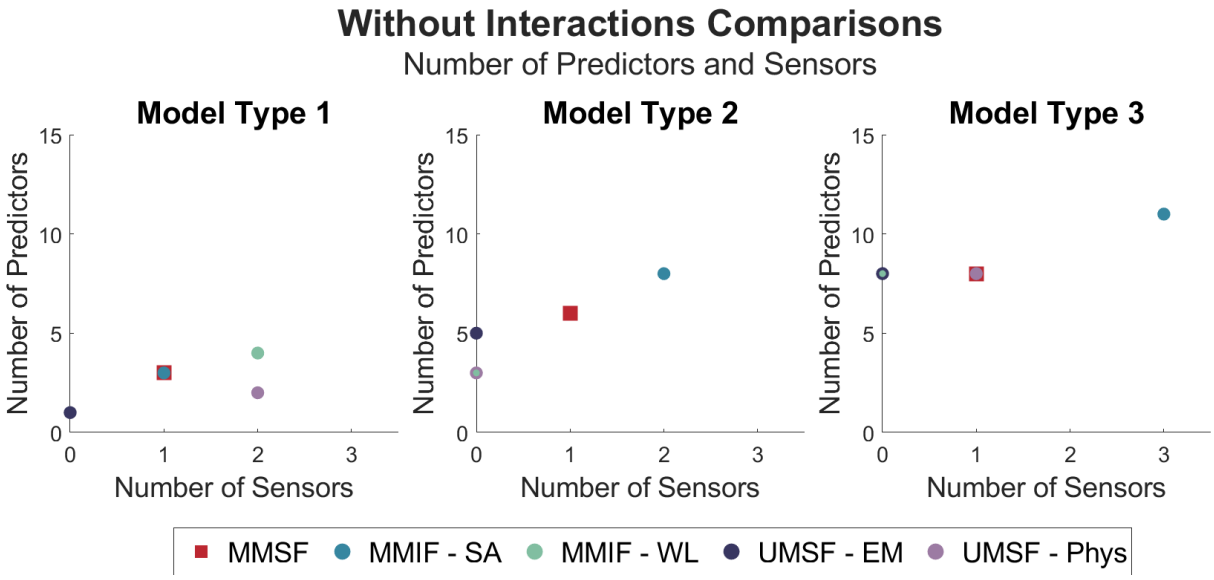


Figure C.3: The comparison of SA and WL models without interactions to other similar models. The x-axis is the number of sensors required (of ECG, RSP, EYE) and the y axis is the number of predictors in the models.

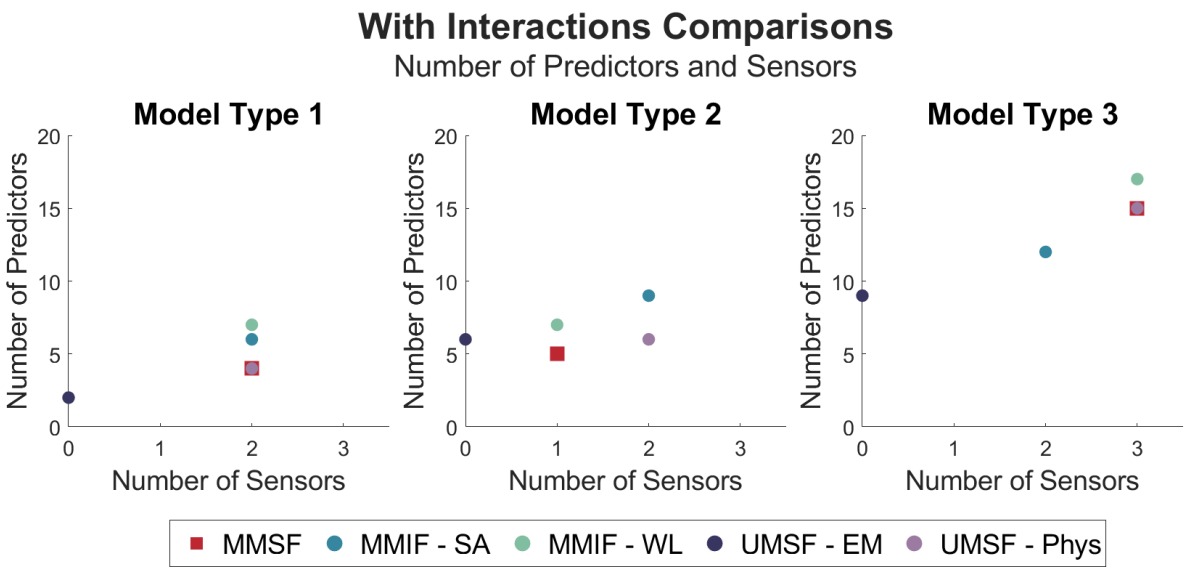


Figure C.4: The comparison of SA and WL models with interactions to other similar models. The x-axis is the number of sensors required (of ECG, RSP, EYE) and the y axis is the number of predictors in the models.

Appendix D: Surveys

D.1 Pre-Experiment Demographics Questionnaire

Pre-Experiment Demographic Questionnaire

Title of research study: Space Habitat Optimized for Mission Exploration

Investigator: Dr. Torin Clark

1. Subject Number: _____
2. Please respond to the following statement: I got adequate sleep last night. (circle one)

   Strongly Agree Agree Neutral Disagree Strongly Disagree
3. How many hours of sleep did you get last night? _____
4. Have you consumed alcohol in the last 6 hours? (please circle one) Yes / No
5. Do you have a known history of seizures? (please circle one) Yes / No
6. What is your handedness/which is your dominant hand? _____
7. How often do you play video games? (please circle one or fill in one blank)

   Never *Monthly _____hrs.* *Weekly_____hrs.* *Daily_____hrs.*
8. Do you use robotic or autonomous systems at least once per week? Yes / No

   Please explain _____
9. Do you use a navigational aid (e.g., Google Maps, Waze, etc.) at least once week? Yes / No
10. What is your approximate level of experience with aerospace or spaceflight relevant information displays? (please circle one)

    No experience *Some experience* *Moderate experience* *Extensive experience*

    Please explain _____

D.2 Post-Experiment Demographics Questionnaire

Post-Experiment Demographic Questionnaire

Title of research study: Space Habitat Optimized for Mission Exploration

Investigator: Dr. Torin Clark

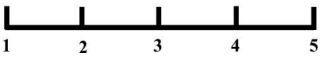
1. Subject Number: _____
2. Sex: ☐ Female ☐ Intersex ☐ Male
3. Age (in years, on today's date): _____
4. Race ("X" ONLY one with which you MOST CLOSELY identify):
   - ☐ American Indian or Alaska Native
   - ☐ Asian
   - ☐ Black or African American
   - ☐ Native Hawaiian or Other Pacific Islander
   - ☐ White
   - ☐ More than one race
   - ☐ Unknown or not reported
5. Ethnicity ("X" ONLY one with which you MOST CLOSELY identify):
   - ☐ Hispanic or Latino
   - ☐ Not Hispanic or Latino
   - ☐ Unknown or not reported

D.3 Automation Induced Complacency Potential questionnaire

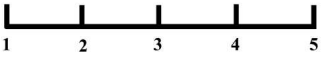
Monitoring score = (6 – Question 5) + Question 7 + (6 – Question 8) + Question 9 + Question 10

Please mark on each line at the point which best describes your feeling or impression (1 = disagree/never, 5 = agree/constantly).

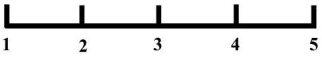
1. When I have a lot to do, it makes sense to delegate a task to automation.



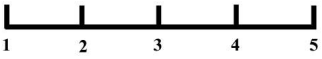
2. If life were busy, I would let an automated system handle some tasks for me.



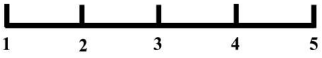
3. Automation should be used to ease people's workload.



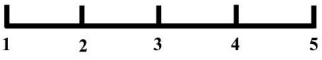
4. If automation is available to help me with something, it makes sense for me to pay more attention to my other tasks.



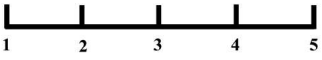
5. Even if an automated aid can help me with a task, I should pay attention to its performance.



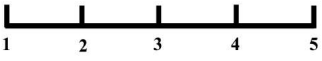
6. Distractions and interruptions are less of a problem for me when I have an automated system to cover some of the work.



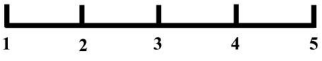
7. Constantly monitoring an automated system's performance is a waste of time.



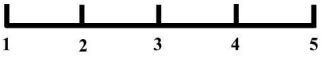
8. Even when I have a lot to do, I am likely to watch automation carefully for errors.



9. It's not usually necessary to pay much attention to automation when it is running.



10. Carefully watching automation takes time away from more important or interesting things.

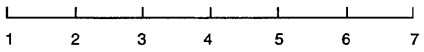


D.4 Trust in Autonomous System Survey

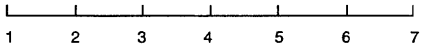
Score = (8 – Question 1) + (8 – Question 2) + (8 – Question 3) + (8 – Question 4) + (8 – Question 5) + (Question 6 + Question 7 + Question 8 + Question 9 + Question 10 + Question 11 + Question 12)

(Note: not at all=1; extremely=7)

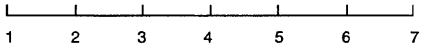
1. The system is deceptive



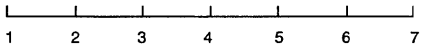
2. The system behaves in an underhanded manner



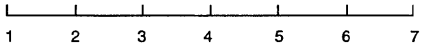
3. I am suspicious of the system's intent, action, or outputs



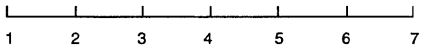
4. I am wary of the system



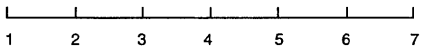
5. The system's actions will have a harmful or injurious outcome



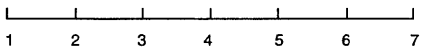
6. I am confident in the system



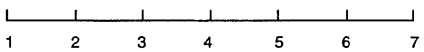
7. The system provides security



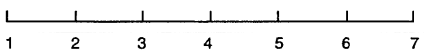
8. The system has integrity



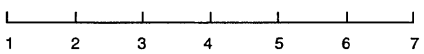
9. The system is dependable



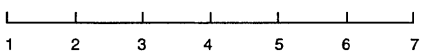
10. The system is reliable



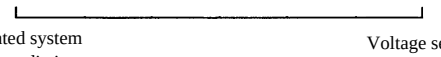
11. I can trust the system



12. I am familiar with the system



13. Place a single mark indicating how much you relied upon the sensor and the prediction to make your setting during the previous trial. A mark in the middle indicates you relied on both equally.

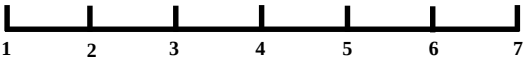


Automated system voltage prediction — Voltage sensor

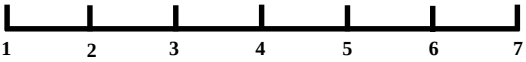
D.5 Situation Awareness Rating Technique

"10D SART" Score = (Question 8 + Question 9 + Question 10) – ((Question 1 + Question 2 + Question 3)) – (Question 4 + Question 5 + Question 6 + Question 7))

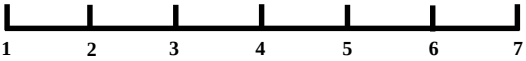
Instability of situation (likeliness to change suddenly)



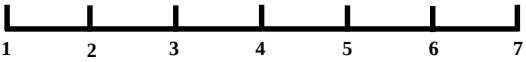
Variability of situation (number of variables/factors changing)



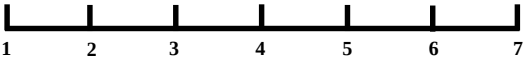
Complexity of situation (degree of complication**)**



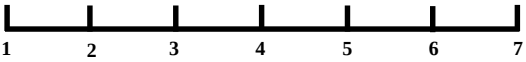
Arousal (degree of alertness; readiness for activity)



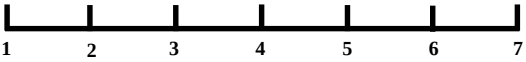
Spare mental capacity (mental ability available for new variables)



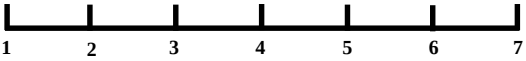
Concentration (degree to which thoughts are brought to bear)



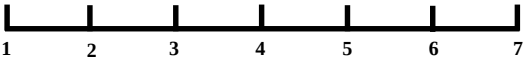
Division of attention (distribution/spread of focus of attention)



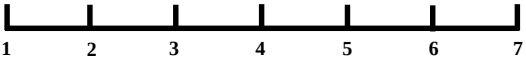
Information quantity (amount of knowledge received and understood)



Information quality (goodness or value of knowledge communicated)



Familiarity (degree of prior experience/knowledge)



D.6 Modified Bedford Scale

Please start at the bottom left and follow the yes/no questions to yield an estimate of your mental workload on the previous trial. Verbally report the number and circle the number on the scale.