Psychophysiological Models of Cognitive States Can Be Operator-Agnostic

Extended Abstract

Erin E. Richardson
University of Colorado Boulder
Boulder, United States
erin.richardson@colorado.edu

Savannah L. Buchner
University of Colorado Boulder
Boulder, United States
savannah.buchner@colorado.edu

Jacob R. Kintz
University of Colorado Boulder
Boulder, United States
jacob.kintz@colorado.edu

Torin K. Clark
University of Colorado Boulder
Boulder, United States
torin.clark@colorado.edu

Allison P. Anderson
University of Colorado Boulder
Boulder, United States
allison.p.anderson@colorado.edu

ABSTRACT

Real-time prediction of a person's trust (T), mental workload (W), and situation awareness (SA) can improve safety and performance in operational environments. We develop psychophysiological models of TWSA both with and without operator-specific demographic information available to them and we assess the impact of this constraint on the models' performance. We demonstrate functional model fit (Adjusted R^2: T = 0.67, W = 0.65, and SA = 0.88) and predictive ability with operator-agnostic models, assessed via leave-one-participant-out cross validation (Q^2: T = 0.58, W = 0.45, and SA = 0.79). Our findings help establish the viability of operator-agnostic psychophysiological models of TWSA which could be used to inform an autonomous agent or manage multi-agent teams.

KEYWORDS

Psychophysiology; Predictive Modeling; Human-Agent Teaming

ACM Reference Format:
Erin E. Richardson, Savannah L. Buchner, Jacob R. Kintz, Torin K. Clark, and Allison P. Anderson. 2024. Psychophysiological Models of Cognitive States Can Be Operator-Agnostic: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Operational environments, such as spaceflight, often require humans to team with multiple agents, which may include other humans and autonomous systems. In future complex operational environments, the ability to model and predict an operator's trust (T), mental workload (W), and situation awareness (SA), collectively "TWSA", can facilitate improved safety and performance [11].

Current methods of measuring or predicting TWSA are insufficient for operational environments. Administering questionnaires requires interrupting an operator's task and behavioral and task-based measures are often specific to a given task, and thus lose applicability with changes to the tasks or protocols. Psychophysiological measures, however, do not require interrupting an operator's work and do not explicitly rely on task-specific elements. As a result, predictive models of TWSA based on psychophysiological measures may show robustness to moderate changes in tasks or protocol as well as utility across different tasks. Predictive models would provide additional utility if they were accurate across users. Collecting demographic information may be impractical in time-constrained situations [15, 18]. Furthermore, necessitating collection of demographic data may decrease a tool's acceptability, especially where an operator's performance affects their career [3, 4]. Since identifiable information facilitates mapping of data to individuals, avoiding collection of operator-specific information may help to decrease privacy risk and improve tool acceptability.

We explore the implications of excluding operator-specific information on modeling TWSA. We build psychophysiological models of operator TWSA in a supervisory task where participants work alongside a simulated autonomous system. We improve on prior work that relies on proxy measures by using validated questionnaires as the target variables in predicting TWSA. Our predictive modeling approach improves upon classification techniques by modeling TWSA as continuous constructs via multiple regression. We perform feature shrinkage to reduce our feature set and stability selection to reduce variability in feature selection. Finally, we use internal cross-validation to assess model predictive accuracy. We hypothesized that model performance would decrease with less information available, but that operator-agnostic models would still demonstrate viability for use in predicting operator TWSA.

2 METHODS

This study was approved by the University of Colorado's Institutional Review Board. Twelve people participated in the study but data from two of them were excluded due to technical issues during data collection. Data from the remaining ten participants (6M/4F, age 25 ±7 [19-42]) were used in our analysis.

Participants worked in a supervisory role alongside a simulated autonomous system to maintain a modeled deep space habitat environmental control and life support system (ECLSS) described in detail in [8]. Each participant was trained on the task and completed practice trials. Before starting the experiment, participants completed a demographic questionnaire, the Automation-Induced

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Table 1: Performance of TWSA models

Model Description		$N_{predictors}$	*Adjusted* R^2	*LOPO* Q^2	*LOTO* Q^2
T	Specific	29	0.71	0.54	0.63
	Agnostic	25	0.67	0.58	0.61
W	Specific	25	0.66	0.48	0.56
	Agnostic	24	0.65	0.45	0.54
SA	Specific	29	0.90	0.79	0.87
	Agnostic	29	0.88	0.79	0.81

Complacency Potential (AICP) rating scale, and the 3-minute Psychomotor Vigilance Test (PVT) [1, 10]. Participants wore a 3-lead electrocardiogram montage, a BIOPAC respiratory chest belt, electrodermal activity electrodes on two fingers, and Pupil Labs' Pupil Core eye-tracking headset [6]. Baseline physiological signals were recorded for 20 seconds before each trial while participants sat still and visually fixated on a crosshair. Each participant completed 15 trials that were 50-95 seconds long. Throughout the trials, scripted events presented anomalies in the ECLSS and the autonomous system suggested actions in response. The autonomous system randomly transitioned between different modes of operation across trials by altering transparency and decision authority [8]. Participants self-reported their TWSA via subjective questionnaires after each trial; Jian et al.'s scale, a modified version of the Bedford Workload scale, and the "10D" Situational Awareness Rating Technique were used to measure TWSA, respectively [5, 13, 14, 16].

Predictive models of TWSA were built using the subjective questionnaire scores as the ground truth. Features were extracted from the physiological data and the demographic questionnaires. First-order interaction terms were also generated. We divided the potential predictors into two subsets: operator-specific and operator-agnostic. Operator-specific predictors include the demographic questionnaire responses, AICP scores, and PVT scores while the operator-agnostic features do not. Sets of predictors were down-selected from the subset of potential predictors for each model. 10-fold cross validation relaxed LASSO was used to identify two sets of predictors by: 1) setting the shrinkage coefficient, λ, at the one standard error (1-SE) location and 2) setting λ at the minimum mean squared error location [2, 17]. This was repeated 50 times, resulting in 100 sets of predictors. Next, any terms that appeared in any of the 100 sets were used in a subsequent run of LASSO with λ at the 1-SE location. This was repeated 50 times, resulting in 50 more sets of predictors. This stability selection aims to address the instability in predictor sets resulting from cross-validation embedded in the LASSO method [9]. Ordinary least squares was used to fit coefficients to each unique set of predictors output by LASSO, generating a set of model options. Two forms of exhaustive cross-validation were performed to assess each model option: leave-one-participant-out (LOPO) and leave-one-trial-out (LOTO), generating LOPO and LOTO Q^2s [12]. It should be noted that in both cross-validation measures, the left-out participant/trial was not left out of the feature selection process, such that the final model features remained the same while the fitted coefficients differed with each left-out prediction. All of the trials were used to fit the final coefficients and calculate the adjusted R^2 values. To further protect against overfitting, we constrained models to contain no more predictors than 1/5 of the number of observations. We also constrained models to have Q^2s within 0.2 of the adjusted R^2, as disparity in these values is indicative of overfitting. The final model was the one which achieved the highest adjusted R^2 while satisfying these constraints. This process was conducted for both feature subsets for each of TWSA.

3 RESULTS

The model performances are summarized in Table 1. Overall, limiting TWSA models to operator-agnostic features did not substantially decrease performance. The largest drop in adjusted R^2 was 0.04 for the trust model. Critically, the operator-agnostic models still performed well when predicting the TWSA of participants whose data were not used to fit their coefficients, as indicated by their LOPO Q^2s of 0.58, 0.45, and 0.79. This result is important as we aim to build models that can generalize to new operators.

4 DISCUSSION

Across all three cognitive states, the operator-agnostic models demonstrated predictive relevance. These models' basis in psychophysiological signals rather than task-specific information and operator-specific information further affirm their utility for predicting TWSA in operational environments. Previous work found trust and workload model performance to greatly decrease without demographic predictors [7]. These models, however, did not have physiological measures available to them and only used task-embedded measures. In our study, the exclusion of operator-specific predictors resulted in only small drops in performance.

Going forward, data should be collected from larger sample sizes and from real field operators. Our cross-validation is limited in that the test data sets were not left out of the predictor selection process, giving rise to the potential for internal validation bias. More subjects would allow for separate train and test datasets, mitigating internal validation bias. Our ongoing work further addresses this limitation by including feature selection in the cross-validation process.

5 CONCLUSION

This work demonstrates a viable pathway for operator-agnostic prediction of TWSA using psychophysiological data. We fit models derived from psychophysiological data to gold-standard questionnaire scores of TWSA. TWSA models that were not privy to operator-specific demographic information achieved good fit (adjusted R^2s of 0.67, 0.65, and 0.88) and predictive ability (LOPO Q^2s of 0.58, 0.45, and 0.79). Furthermore, the models do not interrupt users' actions and do not rely on operationally-cumbersome demographic questionnaires. They could be used in a variety of settings, such as to inform the behavior of adaptive autonomous systems or to allocate resources in a team of operators, thereby improving performance and safety of human-autonomy teams and of humans working in operational environments.

ACKNOWLEDGMENTS

This work was supported by NASA STRI 80NSSC19K1052.

REFERENCES

[1] Mathias Basner, Adam Savitt, Tyler Moore, Allison Port, Sarah McGuire, Adrian Ecker, Jad Nasrini, Daniel Mollicone, Christopher Mott, Thom McCann, David Dinges, and Ruben Gur. 2015. Development and Validation of the Cognition Test Battery for Spaceflight. *Aerospace Medicine and Human Performance* 86 (Nov. 2015), 942–952. https://doi.org/10.3357/AMHP.4343.2015

[2] Savannah L. Buchner. 2022. *Multimodal Feature Selection to Unobtrusively Model Trust, Workload, and Situation Awareness.* Master's thesis. University of Colorado at Boulder, United States – Colorado. https://www.proquest.com/docview/2681067437/abstract/FB66870A43AA40A0PQ/1

[3] Cherrylyn Buenaflor, Hee-Cheol Kim, and S. Korea. 2013. Six Human Factors to Acceptability of Wearable Computers. https://www.semanticscholar.org/paper/Six-Human-Factors-to-Acceptability-of-Wearable-Buenaflor-Kim/3ae7835e1270d4d60bf011532c07ad979437a671

[4] Byungjoo Choi, Sungjoo Hwang, and SangHyun Lee. 2017. What drives construction workers' acceptance of wearable technologies in the workplace?: Indoor localization and wearable health devices for occupational safety and health. *Automation in Construction* 84 (Dec. 2017), 31–41. https://doi.org/10.1016/j.autcon.2017.08.005

[5] Jiun-Yin Jian, Ann Bisantz, and Colin Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4 (March 2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[6] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1151–1160. https://doi.org/10.1145/2638728.2641695

[7] Jacob R. Kintz, Neil T. Banerjee, Johnny Y. Zhang, Allison P. Anderson, and Torin K. Clark. 2023. Estimation of Subjectively Reported Trust, Mental Workload, and Situation Awareness Using Unobtrusive Measures. *Human Factors* 65, 6 (Sept. 2023), 1142–1160. https://doi.org/10.1177/00187208221129371

[8] Jacob R. Kintz, Young-Young Shen, Savannah L. Buchner, Allison P. Anderson, and Torin K. Clark. 2023. A Simulated Air Revitalization Task to Investigate Remote Operator Human-Autonomy Teaming With Communication Latency. In *52nd International Conference on Environmental Systems.* https://ttu-ir.tdl.org/handle/2346/94539

[9] Nicolai Meinshausen and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, 4 (2010), 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

[10] Stephanie M. Merritt, Alicia Ako-Brew, William J. Bryant, Amy Staley, Michael McKenna, Austin Leone, and Lei Shirase. 2019. Automation-Induced Complacency Potential: Development and Validation of a New Scale. *Frontiers in Psychology* 10 (2019). https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00225

[11] Raja Parasuraman, Thomas Sheridan, and Christopher Wickens. 2008. Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making* 2 (July 2008), 140–160. https://doi.org/10.1518/155534308X284417

[12] Nguyen T. Quan. 1988. The Prediction Sum of Squares as a General Measure for Regression Diagnostics. *Journal of Business & Economic Statistics* 6 (1988), 501–504. https://doi.org/10.2307/1391469

[13] A. H. Roscoe. 1984. *Assessing Pilot Workload in Flight.* Technical Report. ROYAL AEROSPACE ESTABLISHMENT BEDFORD. https://apps.dtic.mil/sti/citations/ADP004109

[14] A. H. Roscoe and G. A. Ellis. 1990. *A Subjective Rating Scale for Assessing Pilot Workload in Flight: A decade of Practical Use.* Technical Report. ROYAL AEROSPACE ESTABLISHMENT FARNBOROUGH. https://www.semanticscholar.org/paper/A-Subjective-Rating-Scale-for-Assessing-Pilot-in-A-Roscoe-Ellis/fcb2e3627e7ca07101ac1d1ad6e0f79e1c23f5c2

[15] Mark R. Rosekind, R. Curtis Graeber, David F. Dinges, Linda J. Connell, Michael S. Rountree, Cheryl L. Spinweber, and Kelly A. Gillen. 1994. *Crew factors in flight operations 9: Effects of planned cockpit rest on crew performance and alertness in long-haul operations.* Technical Report DOT/FAA/92/24. https://ntrs.nasa.gov/citations/19950006379

[16] R. M. Taylor. 1990. Situational Awareness Rating Technique (SART): The Development of a Tool for Aircrew Systems Design. In *Situational Awareness in Aerospace Operations (AGARD-CP-478).* NATO - AGARD, Neuilly Sur Seine, France.

[17] Robert Tibshirani. 1996. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[18] Danni Tu, Mathias Basner, Michael G. Smith, E. Spencer Williams, Valerie E. Ryder, Amelia A. Romoser, Adrian Ecker, Daniel Aeschbach, Alexander C. Stahn, Christopher W. Jones, Kia Howard, Marc Kaizi-Lutu, David F. Dinges, and Haochang Shou. 2022. Dynamic ensemble prediction of cognitive performance in spaceflight. *Scientific Reports* 12 (June 2022). https://doi.org/10.1038/s41598-022-14456-8