Explainable predictions for real-time employee workload management in railway control rooms

Léon Sobrie*[1], Marijn Verschelde[2], and Bart Roets[3]

[1]Ph.D. Candidate, Ghent University, Belgium
[2]Full Professor, IÉSEG School of Management, France
[3]Railway Scientist & Ph.D., Infrabel & Ghent University, Belgium

Short summary

Industry 5.0 targets a resilient, sustainable and human-centric European industry. A key initiative to reach this target is adopting a human-centric approach to digital technologies, which places the well-being of the worker at the center. As workload peaks/lows contribute to lower employee well-being, predictive employee workload analytics can empower management to undertake proactive prevention. For this purpose, we develop a real-time machine learning framework to predict and explain future workload. Our feature importance analysis demonstrates the value of human-machine interactions and partner workload exposure. The proposed 2-stage framework, inspired by deep Tobit models, is developed and implemented in an environment with a variable and imbalanced workload: the digital control rooms for railway traffic management of Infrabel, Belgium's railway infrastructure company. The related application is tailored towards the managers, for whom it provides real-time and explainable insights.

Keywords: Analytics, explainable, operators, railways, real-time, workload

1 Introduction

The European Commission (EC) launched Industry 5.0 to complement Industry 4.0 by focusing on research and innovation that serves the transition to a resilient, sustainable and human-centric European industry (European Commission, 2022). A key initiative proposed by the EC is to adopt a human-centric approach toward digital technologies. We aim to contribute to this initiative by focusing on human workload in a highly digitized environment: the railway control rooms of Infrabel, Belgium's railway infrastructure company. In this setting, millions of actions are taken on a monthly basis by traffic operators to control railway traffic in real-time.

Operations literature acknowledges monitoring of adequate employee workload as one of the objectives of the control room management (see e.g. Valls et al. (2009)). Adequate workload has to be considered in comparison to both overload and underload. High levels of workload (i) are connected to lower daily well-being (Ilies et al., 2010), (ii) lead to task preference for easier tasks hurting performance (Kc et al., 2020), (iii) are a causal factor for human fatigue (Li et al., 2020), (iv) induce quality degradation due to cognitive multitasking (Xu et al., 2022) and (v) are likely to trigger the health-impairing mental/physical conditions of individuals, which are often related to safety performance (Derdowski & Mathisen, 2023). Low levels of workload (i) result in extra performance-seeking risks (Xu et al., 2022) and (ii) lead to boredom and lack of attention (Young, 2021). Therefore, creating an environment with a balanced workload amongst and within employees contributes to their well-being and satisfaction (Inegbedion et al., 2020). Moreover, workload is a *multi-attribute concept* (Comstock Jr & Arnegard, 1992) which entails communication, resource management, automation, scheduling, monitoring and tracking. Hence, analytics for workload management should incorporate different attributes to provide granular and explainable insights.

Our contribution is fourfold. First, we contribute to the operations research literature, by showing the applicability of machine learning to provide real-time employee-centric predictive analytics. We answer the calls from literature to connect operations research with human resource management (Roels & Staats, 2021) and to bridge the gap between practice and research by constructing a model that is embedded in a problem observed in practice (Ranyard et al., 2015)). Related literature on employee-centric predictions for French air traffic controllers (TC) postulates a neural network for workload prediction combined with a tree-based search model for optimal airspace partitions Gianazza (2010). We distinguish ourselves from the literature by our focus on granular explainability and our multi-attribute consideration of workload. Second, we contribute to the explainable artificial intelligence literature (Coussement & Benoit, 2021) by illustrating the usefulness for adequate explainability of sample selection as based on the unobserved predicted values, via our advocated two-stage methodology, inspired by Heckman (1979) and Zhang et al. (2021). Third, we contribute to the management science (MS) literature on human-machine interaction (see e.g., Brynjolfsson & McAfee (2014)) and team exposure (see e.g., Akşin et al. (2021)). Our Shapley analysis pinpoints the importance of including automation and team interplay in workload predictions. Last, we contribute to the literature on applications of smart data analytics (see e.g. Baesens et al. (2016)) through our real-time implementation. Not many systems that utilize machine and/or deep learning have been adopted in a real-world setting (Kraus et al., 2020). One of the exceptions utilizes regression trees to build real-time analytics on passenger flows in the control room of Heathrow (Guo et al., 2020).

The paper is structured as follows. In section 2, we elaborate on the 2-stage methodology. In section 3, we dive into the results of the accuracy and explainability of the proposed model, and show its implementation for real-time management in traffic control centers (TCCs). The last section formulates the main takeaways.

2 Methodology

Our methodological framework is developed to provide explainable and accurate predictive analytics on employee workload aggregated in 15-minute intervals. The idea is to dissect workload into different operational categories (see section 3). Methodologically, we are confronted with a classical sample selection issue (see Heckman (1979)). This is because we have no a priori reason to believe that the mechanisms affecting the presence of workload are the same as the mechanisms that affect the workload magnitude, when present. Recently, Zhang et al. (2021) have introduced the sample selection issue into the literature on deep learning via a two-stage '*deep Tobit model*'. Our proposed two-stage approach comprises an LSTM encoder-decoder model in stage 1 to select the workload categories and an XGBoost model in stage 2 to predict the amount of workload within the selected categories.

Stage 1: LSTM encoder-decoder for binary classification

We utilize a Long Short-Term Memory (LSTM) encoder-decoder model (Hochreiter & Schmidhuber, 1997; Cho et al., 2014) to predict the presence of workload. This approach leverages an input-to-output sequence data structure that incorporates information from previous intervals to predict the occurrence of different workload categories for future intervals. Moreover, the prediction model has a memory cell to store the past. This aligns with the call from Corman & Quaglietta (2015) to close the loop in real-time railway traffic control. The training set contains N (input sequence, workload sequence) pairs with K features per input sequence. Each pair combines T input feature sequences $x = (x_1, ..., x_t, ..., x_T)$, with $x_{t \in T} \in R^K$, together with M workload sequences $y = (y_1 ..., y_m, ..., y_M)$, with $y_{m \in M}$ a binary variable, representing the presence of workload in a category. Via the LSTM-based encoder-decoder model, we learn the conditional expectation $E(y_m | y_{m-1}, ..., y_1, x_{T-1}, ..., x_1)$. by estimating $\widehat{y}$ via minimizing squared error over the N training pairs:

$$\min_{\theta} \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \left(y_{m,n} - \widehat{y}_{m,n}(\theta) \right)^2 , \tag{1}$$

with θ the model parameters. Operationalization implies the use of an encoder to model $(x_T, ..., x_1)$ into c. The latter serves as an input for the decoder model to learn the conditional expectation $E(y_m | y_{m-1}, ..., y_1, c)$. By uncovering this conditional expectation, the encoder-decoder model learns to predict y_{m+1}, based on $y_{\leq m}$ and c.

Stage 2: XGBoost for quantifying selected workload categories

The extreme gradient boosting utilizes the gradient boosting algorithm which grows trees sequentially by minimizing a regularized objective function (Chen et al., 2015):

$$\sum_{i} l(\widehat{}_{i}, y_i) + \sum_{j} \Omega(f_j), \text{ where } \Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2, \tag{2}$$

with l a differentiable convex loss function related to the difference between the prediction $\widehat{y}$ and the actual output y, Ω a regularization function, i=1,...N representing the input-output pair, $j = 1, ..., J$ representing the tree, T the number of leaves in a tree, w the leaf weight, γ a user-defined pruning parameter, and λ a shrinkage parameter.
We leverage Shapley values to provide explainable insights for the proposed XGBoost model. Shapley values (Shapley, 1953) incorporate potential synergies between the features, by averaging the marginal contributions over all possible subsets and by taking the order of the marginal contribution into account (Lundberg & Lee, 2017).

3 Results and discussion

Empirical setting

All railway traffic in Belgium is managed in the digital control rooms of Infrabel, with every zone being monitored by one traffic and safety controller at any given time. The basis for the predictive model is a real-time data structure containing all actions taken in the control rooms, including over 5,000,000 tasks executed per month. In this paper, we focus on traffic controllers for the month of February 2022, aggregated per control room, workstation and 15-minute time interval.

We build on previous research (Topcu et al., 2019) that divides workload into 6 operational categories as presented in Table 1. Further, Table 1 contains the occurrence of zeros for each category. We notice that there is no workload in more than 40% of the considered 15-minute intervals, except for MOVE and AUT. This strengthens our empirical choice to propose a 2-stage methodology to first filter for which categories the operator will have workload and, thereafter, predict the workload magnitude in an explainable way. MOVE is excluded in the first stage as the occurrence of zero workload for this category is below 1%, making a binary filter redundant.

Table 1: Operational workload categories

Operational workload categories		**Zero occurrence**
MOVE	Monitoring of railway traffic by opening signals	0.7%
ADAPT	Reducing train delays by changing tracks and station platforms	40.7%
AUT	Changing the automation - Automatic Route Setting	16.3%
SAFETY	Safety interventions	78.2%
PHONE	Phone calls between the control room operator and train drivers	60.6%
JUSTIF	Justification of train delays	48.7%

Similarly to the operational categories, the considered features for our predictive model are grouped by the control room, workstation and 15-minute interval. The features for near-future workload prediction for the different categories consist of (i) the experience and training level of the operator, (ii) automation usage, (iii) trains monitored, (iv) delays, (v) current workload for each category, (vi) partner controller features, and (vii) control room and temporal fixed effects. The features are tested for potential multicollinearity to ensure meaningful explainability.

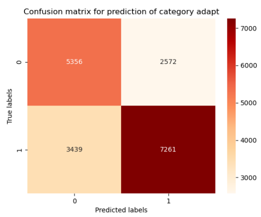
Accuracy of LSTM encoder-decoder for binary classification

The binary LSTM encoder-decoder uses 4 input sequences to predict 4 output sequences. We reach the lowest classification error for dropout=0.2, optimizer=Adam, learning_rate=0.0001, batch_size=25, and number_of_nodes=50. The Area Under the Receiver Operator Curve (AUC) values, provided in Table 2, shows the classification capability for the different workload categories.

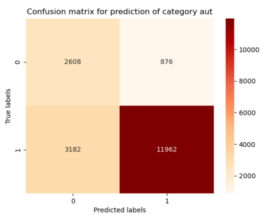
Table 2: AUC values

	SEQ1	SEQ2	SEQ3	SEQ4
ADAPT	0.73	0.74	0.74	0.74
AUT	0.84	0.86	0.86	0.86
SAFETY	0.82	0.82	0.82	0.82
PHONE	0.78	0.78	0.78	0.77
JUSTIF	0.71	0.72	0.72	0.71

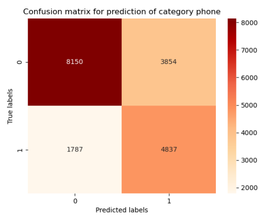
The AUC is above 0.7 for all categories, with over 0.84 for AUT and 0.82 for SAFETY. The stability of the AUC over the four output sequences demonstrates the classification power of stage 1. Figure 1 presents the confusion matrices for the different categories with the tuned thresholds.



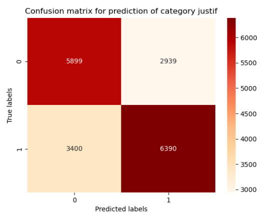
(a) ADAPT, threshold = 0.7



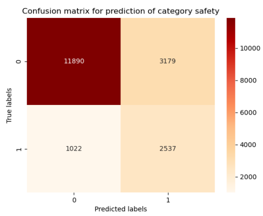
(b) AUT, threshold=0.9



(c) PHONE, threshold=0.4



(d) JUSTIF, threshold=0.6



(e) SAFETY, threshold=0.1

Figure 1: Confusion matrices

Explainability of extreme gradient boosting

The explainability analysis focuses on the first 15-minute interval in the future, as this is the most relevant horizon for real-time management. The XGBoost model reaches the lowest error for overall workload prediction with learning_rate=0.1, max_depth=3, number_of_estimators=100, percentage_sampled_columns_per_tree=0.4, and subsample_percentage_training_data=1. Table 3 presents the root mean square error (RMSE) and Spearman correlation.

Table 3: RMSE and Spearman correlation

	RMSE (in seconds)	**Spearman correlation**
MOVE	17.54	0.73
ADAPT	43.54	0.42
AUT	19.09	0.56
SAFETY	9.08	0.41
PHONE	66.00	0.42
JUSTIF	55.55	0.28

Figure 2 presents the most important features for predicting the overall workload. Automation usage is the top feature demonstrating the value of including human-machine interaction in explainable employee-centric analytics. In particular, TCs automate route setting in non-complex, low-workload situations to further reduce their workload (Balfe et al., 2015). Further, features on current workload have predictive power, next to experience and workstation fixed effects.

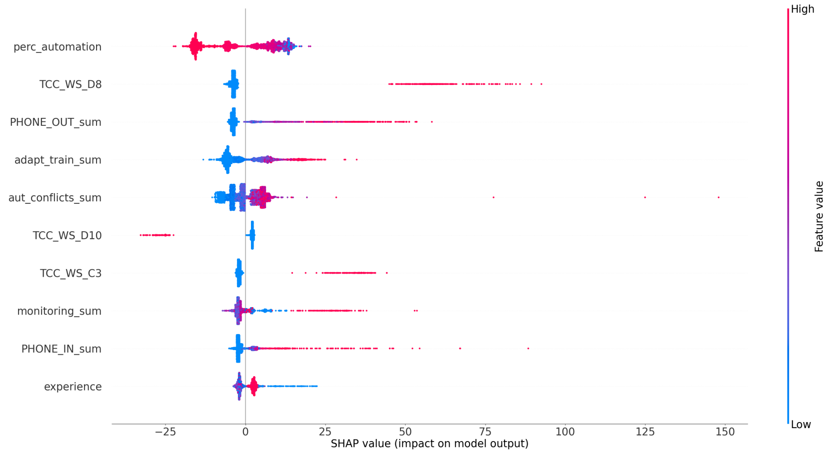
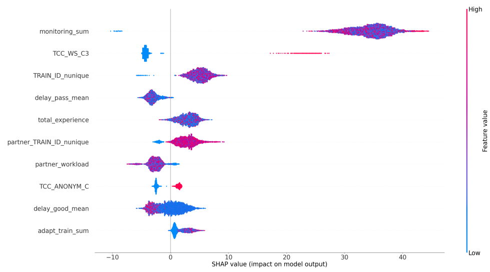
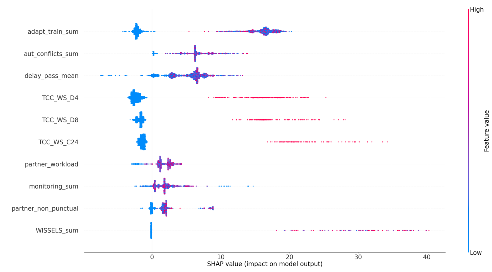


Figure 2: Global Shapley plot

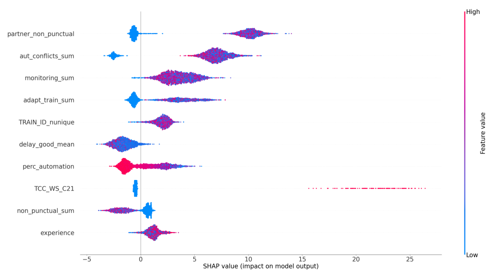
Figure 3 demonstrates the importance of including the partner features for predicting the workload in the operational categories. This finding is in line with the previous MS research (Tan & Netessine, 2019) on the importance of the relationship between observed workload and partner characteristics.



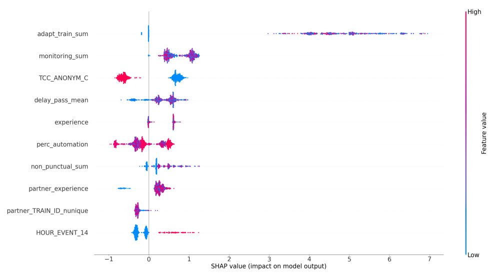
(a) MOVE



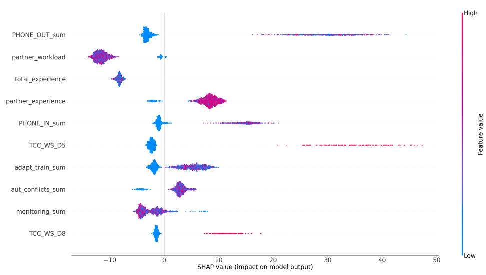
(b) ADAPT



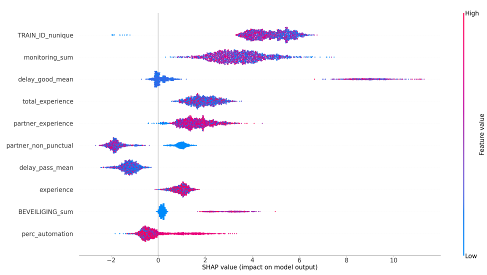
(c) AUT



(d) SAFETY



(e) PHONE

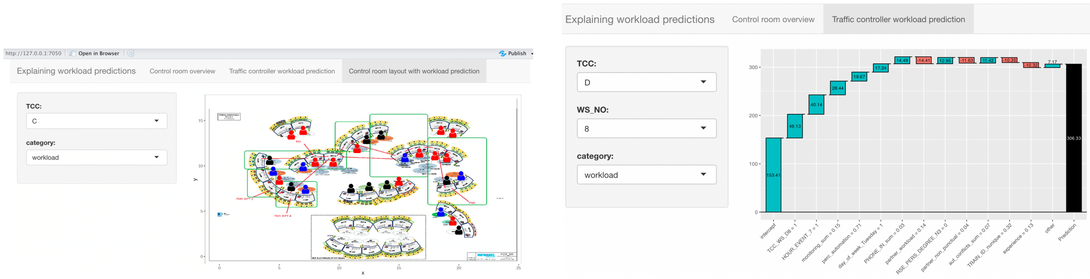


(f) JUSTIF

Figure 3: Global Shapley plots per workload category

Implementation

We implement the proposed prediction framework for control room managers in R Shiny. Figure 4 contains two panels with visualizations. Panel 4(a) presents the control room layout to more easily depict the over-/underloaded operators. Panel 4(b) provides explainable real-time insights into the expected near-future workload by breaking down the contribution of each feature. The real-time implementation showcases the usefulness of human-centric decision support in safety-critical settings.



(a) Control room layout (b) Workstation

Figure 4: R Shiny implementation

4 Conclusions

The impact of workload and its imbalance between and within employees should not be underestimated, as the repercussions on well-being and satisfaction are noticeable. Highly digitized environments, such as railway control rooms, provide an opportunity to introduce data-driven decision support. We leverage this opportunity and propose explainable predictive analytics, customized for employee workload prediction. This is facilitated by our real-time data structure, covering all actions taken in the control rooms. Empirically, we show the binary classification capability of our proposed LSTM encoder-decoder model for the different workload categories. Next, we deploy Shapley values on the output of the XGBoost model, which demonstrates the importance of automation usage for near-future workload prediction. In addition, we unravel the impact of partner workload on the different workload categories, highlighting the value of including team dynamics in predictive analytics. Furthermore, an R Shiny application for the traffic supervisor deploys the proposed framework to provide near-future workload predictions in real time.

Acknowledgements

We would like to thank Infrabel for funding this research, preparing and implementing the datasets, and providing operational feedback. Special thanks to Jasmijn De Clercq, Pierre Gramme, Kristof Van der Strieckt, and the Fellows of the On Track Lab. The views expressed in this paper are those of the authors and do not necessarily reflect the opinions of Infrabel.
We would like to thank the participants of the EURO 2022 conference, the INFORMS 2022 Annual Meeting, and the 2022 On Track Lab workshop, for their comments.
Marijn Verschelde benefited from financial support from the French National Research Agency through the project BALANCE.

References

Akşin, Z., Deo, S., Jónasson, J. O., & Ramdas, K. (2021). Learning from many: Partner exposure and team familiarity in fluid teams. *Management Science*, *67*(2), 854–874.

Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., & Zhao, J. L. (2016). Transformational issues of big data and analytics in networked business. *MIS quarterly*, *40*(4), 807–818.

Balfe, N., Sharples, S., & Wilson, J. R. (2015). Impact of automation: Measurement of performance, workload and behaviour in a complex control environment. *Applied ergonomics*, *47*, 52–64.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies.* WW Norton & Company.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... others (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1–4.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Comstock Jr, J. R., & Arnegard, R. J. (1992). *The multi-attribute task battery for human operator workload and strategic behavior research* (Tech. Rep.).

Corman, F., & Quaglietta, E. (2015). Closing the loop in real-time railway control: Framework design and impacts on operations. *Transportation Research Part C: Emerging Technologies*, *54*, 15–39.

Coussement, K., & Benoit, D. F. (2021). Interpretable data science for decision making. *Decision Support Systems*, *150*, 113664.

Derdowski, L. A., & Mathisen, G. E. (2023). Psychosocial factors and safety in high-risk industries: A systematic literature review. *Safety Science*, *157*, 105948.

European Commission. (2022). *Industry 5.0 roundtable : Brussels 27 april 2022 : meeting report.* Publications Office of the European Union. doi: doi/10.2777/982391

Gianazza, D. (2010). Forecasting workload and airspace configuration with neural networks and tree search methods. *Artificial intelligence*, *174*(7-8), 530–549.

Guo, X., Grushka-Cockayne, Y., & De Reyck, B. (2020). London heathrow airport uses real-time analytics for improving operations. *INFORMS Journal on Applied Analytics*, *50*(5), 325–339.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153–161.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Ilies, R., Dimotakis, N., & De Pater, I. E. (2010). Psychological and physiological reactions to high workloads: implications for well-being. *Personnel Psychology*, *63*(2), 407–436.

Inegbedion, H., Inegbedion, E., Peter, A., & Harry, L. (2020). Perception of workload balance and employee job satisfaction in work organisations. *Heliyon*, *6*(1), e03160.

Kc, D. S., Staats, B. R., Kouchaki, M., & Gino, F. (2020). Task selection and workload: A focus on completing easy tasks hurts performance. *Management Science*, *66*(10), 4397–4416.

Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research*, *281*(3), 628–641.

Li, F., Chen, C.-H., Zheng, P., Feng, S., Xu, G., & Khoo, L. P. (2020). An explorative context-aware machine learning approach to reducing human fatigue risk of traffic control operators. *Safety science*, *125*, 104655.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Ranyard, J. C., Fildes, R., & Hu, T.-I. (2015). Reassessing the scope of or practice: The influences of problem structuring methods and the analytics movement. *European Journal of Operational Research*, *245*(1), 1–13.

Roels, G., & Staats, B. R. (2021). OM Forum-People-Centric Operations: Achievements and future research directions. *Manufacturing & Service Operations Management*, *23*(4), 745–757.

Shapley, L. (1953). A value for n-person games. In *Contributions to theory games (am-28).* Princeton Univ. Press.

Tan, T. F., & Netessine, S. (2019). When you work with a superman, will you also fly? An empirical study of the impact of coworkers on performance. *Management Science*, *65*(8), 3495–3517.

Topcu, T. G., Triantis, K., & Roets, B. (2019). Estimation of the workload boundary in socio-technical infrastructure management systems: The case of Belgian railroads. *European Journal of Operational Research*, *278*(1), 314–329.

Valls, V., Pérez, Á., & Quintanilla, S. (2009). Skilled workforce scheduling in service centres. *European Journal of Operational Research*, *193*(3), 791–804.

Xu, Y., Tan, T. F., & Netessine, S. (2022). The impact of workload on operational risk: Evidence from a commercial bank. *Management Science*, *68*(4), 2668–2693.

Young, M. S. (2021). In search of the redline: Perspectives on mental workload and the 'underload problem'. In *International symposium on human mental workload: Models and applications* (pp. 3–10).

Zhang, J., Li, Z., Song, X., & Ning, H. (2021). Deep tobit networks: A novel machine learning approach to microeconometrics. *Neural Networks*, *144*, 279–296.