Individualizing a Biomathematical Fatigue Model with Attention Data

Bella Z. Veksler (b.veksler@tier1performance.com)
Tier1 Performance Solutions
Covington, KY 41011 USA

Megan B. Morris (megan.morris.3@us.af.mil) and Glenn Gunzelmann (glenn.gunzelmann@us.af.mil)
Air Force Research Laboratory
WPAFB, OH 45433 USA

Abstract

Fatigue is a problematic factor in many workplace environments, resulting in safety and health risks that require monitoring and management. One means to monitor and manage fatigue is through the use of tools implementing biomathematical fatigue models to create assessment and predictions of operator fatigue based on sleep habits. Unfortunately, these models tend to provide assessments and predictions for an "average" operator given work schedules, lacking individualization. One way in which these models can be individualized is through the use of at-the-moment performance data that can modulate the model estimates. In the current effort, we describe an initial attempt at developing an algorithm to individualize fatigue assessments and predictions from a widely-used biomathematical fatigue model with performance data from a common attention task. We discuss the sleep datasets used for the effort, scaling procedure, and model fitting using a genetic algorithm. We then discuss future directions we will take to further increase the effectiveness and efficiency of the individualization capability and its implications.

Keywords: Psychomotor Vigilance Test; Genetic Algorithm

Fatigue is a problematic factor in several workplace environments such as aviation (Caldwell & Caldwell, 2016), commercial motor vehicle (National Academies of Sciences, Engineering, and Medicine, 2016), railroad (Gertler, DiFiore, & Raslear, 2013), and medical (Kancherla et al., 2020) operations. Given the resulting safety and health risks associated with fatigue, it is crucial that organizations implement fatigue risk management (FRM) programs, policies, and other mitigation efforts to combat fatigue. Traditionally, organizations have commonly implemented policy limits regarding work/duty hour limits and rest breaks to allay fatigue. Increasingly, organizations have implemented various types of FRM programs that provide resources and tools to help mitigate fatigue, document fatigue, and examine incidents involving fatigue (Gander et al., 2011). One tool found within some high-risk operational setting programs is the use of biomathematical fatigue models to create assessments and predictions of operator fatigue. Biomathematical fatigue models include homeostatic regulation and circadian rhythm processes, among other factors, to create predictions of fatigue for operators (Mallis et al., 2004).

One particular model that is used by organizations such as the United States Air Force (USAF) Air Mobility Command (AMC), the U.S. Federal Railroad Administration, among others, is the Sleep, Activity, Fatigue, and Task Effectiveness biomathematical fatigue model (SAFTE; Hursh, Redmond, et al., 2004). This model is typically used as the basis of the Fatigue Avoidance Scheduling Tool (FAST; Hursh, Balkin, et al., 2004), a tool that provides fatigue predictions based on prescriptive sleep schedules given work and rest times. One issue with the SAFTE model and other similar biomathematical models is that the model provides predictions for an "average" operator, lacking individualization. Some researchers have had success individualizing predictions of biomathematical fatigue models. Recently, Liu et al. (2017) had success in individualizing the Unified Model of Performance (UPM; Rajdev et al., 2013) with Psychomotor Vigilance Test (PVT; Dinges & Powell, 1985) reaction times. Since the SAFTE model is the basis of several FRM programs and research has provided support of its effectiveness (Hursh, Redmond, et al., 2004; Van Dongen, 2004), we believe it is advantageous to implement a similar technique as Liu et al. (2017). In the current effort, we develop an algorithm to modulate the SAFTE model fatigue estimates with PVT data. This will provide more valid fatigue assessments from the biomathematical model through individualization gained from use of PVT data.

SAFTE Model

The SAFTE model is a three process model that includes homeostatic regulation, circadian rhythm, and sleep inertia processes to calculate general performance effectiveness

(fatigue) predictions. The model also includes a process to account for chronic sleep deprivation. The SAFTE model embedded in FAST is proprietary and includes additional features to take time zone changes and light into account. In the current effort we utilize the non-proprietary version of the SAFTE model (Hursh, Redmond, et al., 2004). The SAFTE model includes 16 parameters. These are listed in Table 1, along with their general mechanism within the model and effects on the output of the model when modified.

Psychomotor Vigilance Test

The PVT is one of the most widely used tasks to assess fatigue due to its sensitivity to sleep decrements and robustness to learning effects (Arsintescu et al., 2017; Balkin et al., 2000; Basner & Dinges, 2011). In the PVT, participants wait for a rolling reaction time indicator in milliseconds to appear on a computer screen in a known location. When this indicator appears, the participant must respond as fast as possible. The PVT is traditionally 10 minutes in length and has a random inter-stimulus interval (ISI) of 2 to 10 seconds. Mean and median reaction time and number of lapses (reaction times greater than 500 ms) are the most common metrics examined to assess alertness or fatigue, but there are several other metrics that are also sensitive to fatigue (e.g., mean 1/RT, slowest 10% 1/RT, etc.) (Basner & Dinges, 2011).

Current Effort

In the remainder of the paper, we will describe our process for individualizing the SAFTE biomathematical fatigue model using PVT data. First, we will describe the archival sleep deprivation dataset used to develop and test the algorithm. We will then describe the process to scale model outputs to the PVT outcomes and how specific SAFTE parameters were chosen. We will then demonstrate the predictive capability of fitting the chosen parameters to individuals. Lastly, we will discuss implications of this work and future plans.

Table 1: SAFTE Model Parameters

Par	Rep	Effects	DV	RE
p	24h acrophase	Shifts effectiveness curve left and right	18	[1,24]
pp	12h acrophase	Changes shape of effectiveness curve	3	[1,12]
beta	Relative amplitude of 12h rhythm	If both circadian peaks are at the same height	0.5	[0,1]
m	Sleep propensity mesor	Positive values increase sleep inertia	0	[-5,10]
as	Sleep propensity amplitude	Higher values increase effectiveness	.55	[-5,5]
a1	Performance rhythm amplitude (fixed %)	Height of peak of circadian component	7	[0,20]
a2	Performance rhythm amplitude (variable %)	Height of peak of circadian component	5	[0,20]
f	Feedback amplitude	How gradually sleep increases reservoir	.0026243	[0,1]
k	Performance use rate	Depletion rate while awake	.5	[0,1]
k1	Down-regulation time constant	Only during sleep	.22	[0,5]
k2	Reference level for SI regulation	Only during sleep	0.5	[.01,5]
k3	Recovery time constant	Only during sleep	.0015	[0,5]
SI max	Max sleep accumulation per minute	Only effect when sleep <=3 hours	3.4	
I	Sleep inertia time constant	Only effect 2 hours following awakening	.04	
I max	Max inertia following awakening	Only effect 2 hours following awakening	5	
RC	Reservoir capacity	Kept constant across participants	2880	

Note. Par = Parameter; Rep = Represents; DV = Default Value; RE = Range Explored

Method

Dataset

To test fits from the model we utilized PVT data from two 62-hour sleep deprivation studies run at Washington State University (Tucker, Whitney, Belenky, Hinson, & Van Dongen, 2009; Whitney, Hinson, Jackson, & Van Dongen, 2015). The first dataset (Whitney, Hinson, Jackson, & Van Dongen, 2015) included 26 participants (M_{age} = 25.92, SD_{age} = 4.05, $Range_{age}$ = 22-37, 16 males and 10 females) from the surrounding Washington State University community. Participants were randomly assigned to a sleep deprivation (n = 13) or control group (n = 13). The second dataset (Tucker, Whitney, Belenky, Hinson, & Van Dongen, 2009) included 23 participants ($Range_{age}$ = 22-38, 12 males and 11 females) also from the surrounding Washington State University community. Participants were randomly assigned to a sleep deprivation (n = 12) or control group (n = 11).

The following description of the protocol was common to both studies, except where noted. Participants spent 6 consecutive days (7 in the first study) and 6 nights in the lab. The first two days were a baseline period where participants had 10 hours time in bed from 22:00 to 08:00 each night. The control group continued this sleep schedule in the following days, but the sleep deprivation group was deprived of sleep for 62 continuous hours. The last two days were a recovery period where both groups had 10 hours time in bed each night. Participants completed several different tasks during the studies, but we only focus on the PVT task in the current effort. The PVT task was 10 minutes in length with a random ISI of 2 to 10 seconds. PVT bouts were collected about every 2 hours during scheduled time awake. This resulted in 8 baseline bouts for both groups, 24 bouts for the sleep deprivation group and 14 bouts for the control group during the sleep deprivation period, and 10 recovery bouts for both groups. We specifically focused on the sleep deprivation groups from both studies for this modeling effort. For fitting the model to the human data, we aggregated each participant's median RT by bout.

Scaling Model Outputs to PVT

Sleep schedule input into SAFTE followed the protocol described above. Output from the SAFTE model produces an effectiveness value on a scale of 0 to 1. As is the case in many biomathematical models, the output requires scaling and inversion to reflect the dependent measure of interest (in this case, the median RT per bout) (Van Dongen, 2004). We linearly transformed this value using the following formula: $Model = scale + scale * (1-EV)$, where $scale$ is determined for each participant and is the minimum median RT from all the bouts that went into fitting the model, EV is the effectiveness value output from SAFTE.

Finalizing Parameters to Modulate

After reviewing the effects of each SAFTE parameter on model output, we found that 12 of the parameters were good potential candidates for the individualization of model output. The culling of the original 16 parameters to 12 was done by visually inspecting the effects of each parameter independently with respect to a 62h sleep deprivation sleep schedule. We found that *SImax*, *i*, and *Imax* all had minimal effects on the Effectiveness values output by the model during periods of wakefulness. We chose to also keep the Reservoir Capacity (RC) constant across all participants as its magnitude is directly related to the *k* parameter which controls the rate at which the reservoir is depleted during wakefulness. Rather than vary both parameters, we chose the *k* parameter to vary. Table 1 also lists the ranges we used in exploring parameter effects.

The parameter space we wanted to explore in this work was fairly large (as seen in Table 1) and rather than try to run a brute force exhaustive search for each participant, we turned to genetic algorithms. Genetic algorithms have been used in many domains in order to find sets of parameters that minimize some fitness functions fairly efficiently (Fogel, 2006). We used the *GA* package in R to run a genetic algorithm with a population size of 50 and convergence determined by 50 generations with no change in fitness (Scrucca, 2013, 2017). The fitness function used in the GA was the root mean squared error (RMSE) between the human data and model output after scaling of the median RT. This initial parameter exploration was done using all of the bouts of data for each participant.

Although there are likely significant interactions between the various parameters, as a first pass at determining how much each parameter contributes to individually fitting the human data, we ran the genetic algorithm while varying 11 parameters and keeping the 12th constant. As a control, we used a model with default SAFTE parameters (green dotted line in all figures). Figure 1 shows the resultant average error across all participants when each parameter was held constant at its default value while the rest were explored. The red line in the figures indicates the error when all 12 parameters were varied. As the figure suggests, maintaining the *k* parameter at default had a considerable effect on the model's error. Using the results derived here, we compared the average error to the 12 parameter model and conservatively culled any parameters whose exclusion (keeping them at default values) either resulted in better performance than the 12 parameter model or were within .5

error units. From this point forward we kept the *p, pp, m, f,* and *k3* parameters at default values.

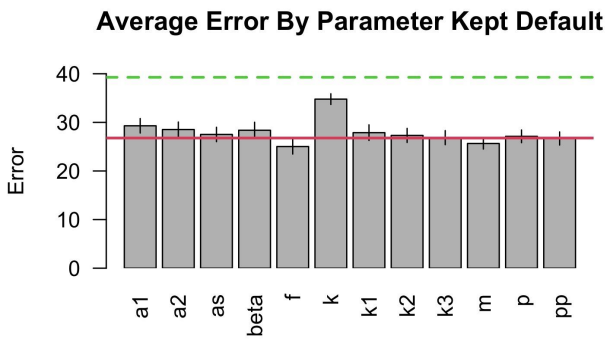

Figure 1: Average RMSE across all participants when fitting by keeping each parameter constant while the other 11 are varied.

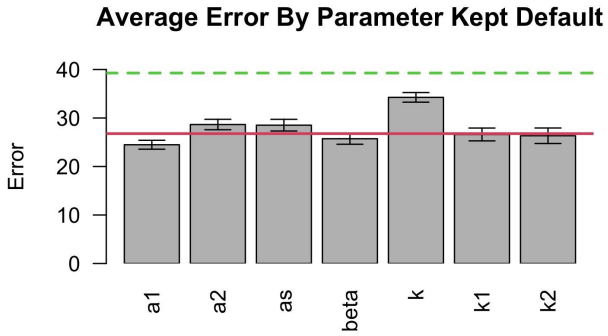

Figure 2: Average RMSE across all participants when fitting by keeping each parameter constant while the other 6 are varied.

We then repeated the above procedure with the 7 remaining parameters and ran the genetic algorithm while varying 6 parameters and keeping the 7th at default. The results are shown in Figure 2. Given these results, we found that we can further keep as default the parameters *a1* and *beta*. In an attempt to further reduce the number of parameters, we ran the GA only varying the *a2*, *as*, and *k* parameters as those seemed to provide the largest improvement to fit, as well as only varying the *k* parameter. Figure 3 shows the average error based on the number of parameters compared to the fully default model (green line) and the 12 parameter model (red line). Based on these results, the 5 parameter model which varies *a2*, *as*, *k*, *k1*, and *k2* produces the best individual fits to our dataset. These parameters correspond to how high the peak of the circadian component is (*a2),* the amplitude of sleep propensity with higher values resulting in higher effectiveness values (*as*), how quickly the reservoir is depleted (*k*), and how quickly the reservoir is refilled (*k1* and *k2).*

There was a statistically significant difference in error in the number of parameters in the model as determined by a linear mixed effects model, ($F(5, 836) = 224.68$, $p < .001$). Post-hoc tests indicated that both the 5-parameter model and the 7-parameter models have significantly less error than the 12-parameter or default models ($p < .05$). In the interest of simplicity, we used the 5-parameter model for predicting performance for each PVT bout based on all preceding bouts.

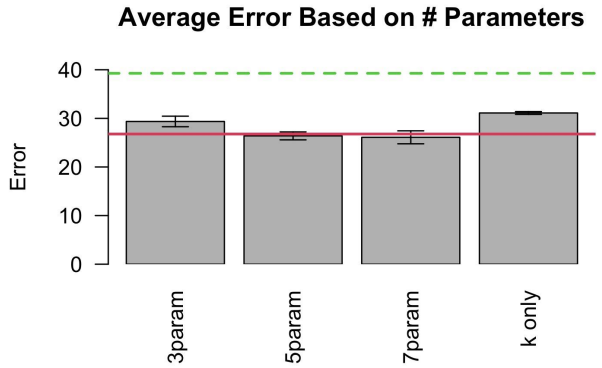

Figure 3: Average RMSE across all participants of the best fitting models by number of free parameters.

Results

Having settled on the 5 parameters we found to be the most appropriate for capturing individual differences in our dataset, we subsequently ran the genetic algorithm to find the best fitting parameters for each participant up to each bout time in order to predict the next bout's performance. Since the goal of the current work is to be able to adjust parameters in real-time to predict future performance, this approach should establish the validity of using the 5 parameters to fit individuals. Figure 4 depicts the average error across all participants during each bout based on the parameter set which minimizes the error of all previous bouts. For comparison, we also used the SAFTE model with default parameters and scaled each individual's performance as before. There was a statistically significant interaction between model type and hour on the error between the model's predicted median RT and the human data as determined by a linear mixed effects model, ($F(33, 1611) = 6.03$, $p < .001$) as well as both the main effect of type ($F(1, 1611) = 167.54$, $p < .001$) and hour ($F(33, 1611) = 20.28$, $p < .001$). Post-hoc tests indicated that there were significant differences between the default model and the 5-parameter individualized model in all hours between 105 and 141 into

the study ($p < .01$). These hours corresponded to being awake for 25 to 61 hours.

It should be noted that there was considerable variability between participants in terms of how accurately the individualized model was capable of predicting performance. In particular, the majority of participants (n = 19) had an average error of 30 ms or less across all bouts. However, in some participants the later bouts which occurred during the sleep deprivation period were not as well fit by the model, as shown in Figure 4 in which bouts occurring during hours 120-140 have high variability and higher error. After inspecting the poorer fitting participants, we found that there was a difference in goodness of fit between the participants from the first study and those of the second. It was unclear why this would be the case as both studies used the same protocol. However, after filtering out the participants from the second study, we found a reduction in prediction error which mimicked that of the error we see when fitting the entire data set, see Figure 5. We again found a significant interaction between the model type and hour, ($F(33, 795) = 12.27$, $p < .001$) as well as both the main effect of type ($F(1, 795) = 318.88$, $p < .001$) and hour ($F(33, 795) = 20.68$, $p < .001$). As in the above analysis, post-hoc tests revealed significant differences between the two models for hours 105 to 141.

Taken together, our initial individualized modeling effort resulted in much better predictions of next bout performance than the default parameter model despite the large variability inherent during sleep deprivation bouts.

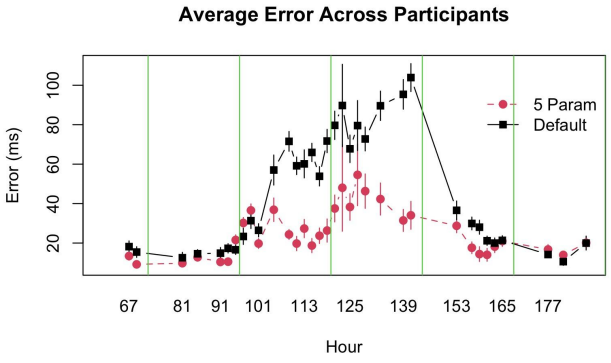

Figure 4: Predicted bout's median RT error based on all previous bout data, includes 25 participants in the sleep deprivation condition. Green lines are day boundaries.

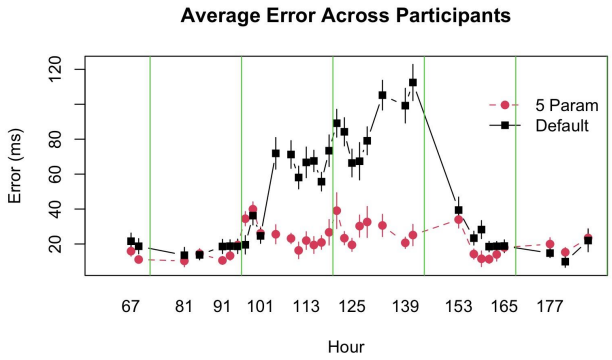

Figure 5: Predicted bout's median RT error based on all previous bout data, includes only the 13 participants from the first 62-h study who were in the sleep deprivation condition.

Discussion

In the current effort, we have demonstrated an initial attempt to develop an algorithm to individualize SAFTE biomathematical fatigue model estimates with PVT performance data. Although SAFTE includes 16 parameters which could theoretically all be manipulated in order to fit individual performance data, we attempted to cull the number of parameters down, both to avoid overfitting and to more efficiently find best fits, while still maintaining the ability to both fit the data and predict future performance. Out of the 16 parameters, 3 were negligible in their contribution to effectiveness values while awake and a 4th highly correlated with another parameter. Further exploration of the remaining 12 parameters found that 5 more could be culled without appreciably affecting the individual fits. Furthermore, a 5-parameter model was capable of fitting individual data as well as a 7-parameter model. Further reducing the number of parameters, however, produced worse fits. This suggests that these five parameters are associated with important individual differences regarding fatigue. The *a2* parameter is likely associated with differences in circadian typology, the *as* parameter is associated with how quickly individuals fall asleep and their ability to stay asleep, and *k*, *k1*, and *k2* parameters are likely associated with differences in sleep need.

We then used the 5-parameter model to fit individual performance up to a particular bout and predict performance on the subsequent bout. The error between the model's predicted median RT and that of each individual participant's was within a range commensurate with using the entire data set to fit the parameters, suggesting that this approach may allow us to update parameter estimates with

limited data and provide a more individualized model of performance than the default SAFTE model. Although the individualized predicted fits get somewhat worse during the sleep deprivation period, they are still much better than using the default parameters.

The traditional PVT implementation is not practical in operational contexts due to the length of the task (10 minutes) and the hardware used to collect the reaction times (e.g., desktop computer or laptop) (Lamond et al., 2005). As a result, researchers have examined the validity of shorter PVT implementations (e.g., 5 or 3 minutes) on handheld devices (e.g., Basner et al., 2011, Grant et al., 2017; Lamond et al., 2005). Overall, these studies have found these implementations to be valid assessments of fatigue when the traditional PVT implementation is not possible. Three minute smart-phone based PVTs are an especially attractive alternative for operational environments as they are short in duration and operators commonly carry these devices on their person. As a result, real-time performance from a smartphone PVT can be used to individualize biomathematical fatigue models within FRM programs. The current effort is a first step in allowing us to use the output from these shorter duration PVT implementations to individualize predictions.

We will continue to improve upon the algorithm by testing with additional sleep deprivation, restricted sleep, and shift-work datasets to demonstrate performance in various sleep impairment-related contexts. Future work will also explore other scaling mechanisms as well as different dependent measures such as number of lapses and false alarms as those have also typically been used to evaluate PVT performance. We will also work toward being able to predict performance further than one bout in the future. Finally, the ultimate goal of this work is to provide efficient real-time parameter estimation on an individual basis allowing us to predict future performance.

Acknowledgments

The opinions expressed herein are solely those of the authors and do not necessarily represent the opinions of the United States Government, the U.S. Department of Defense, the U.S. Air Force, or any of their subsidiaries, or employees. This research was supported by funding from the 711th Human Performance Wing Studies and Analysis Intramural Proposal (Defense Health Program) program through the Aerospace Medicine Studies and Analysis Council. Distribution A. Approved for public release. Case number AFRL-2021-1848.

References

Arsintescu, L., Mulligan, J. B., & Flynn-Evans, E. E. (2017). Evaluation of a psychomotor vigilance task for touch screen devices. *Human Factors*, *59*(4), 661-670.

Balkin, T., Thorne, D., Sing, H., Thomas, M., Redmond, D., Wesensten, N., Williams, J., Hall, S., & Belenky, G. (2000). Effects of sleep schedules on commercial motor vehicle driver performance (Report No. MC-00-133). Springfield, VA: National Technical Information Service, U.S. Department of Transportation.

Basner, M., & Dinges, D. F. (2011). Maximizing sensitivity of the Psychomotor Vigilance Test (PVT) to sleep loss. *Sleep*, *34*(5), 581-591.

Basner, M., Mollicone, D., & Dinges, D. F. (2011). Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. *Acta Astronautica*, *69*, 949-959.

Caldwell, J. A., & Caldwell, J. L. (2016). Fatigue in aviation: A guide to staying awake at the stick. New York, NY: Routledge.

Dinges, D. F., & Powell, J. W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods Instruments and Computers*, *17*, 652-655.

Gander, P., Hartley, L., Powell, D., Carbon, P., Hitchcock, E., Mills, A., & Popkin, S. (2011). Fatigue risk management: Organizational factors at the regulatory and industry/company level. *Accident Analysis and Prevention*, *43*, 573-590.

Gertler, J., DiFiore, A., & Raslear, T. (2013). Fatigue status of the U.S. railroad industry. U.S. Department of Transportation Federal Railroad Administration (report num: DOT/FRA/ORD-13/06). Office of Research and Development Washington, DC.

Grant, D. A., Honn, K. A., Layton, M. E., Riedy, S. M., & Van Dongen, H. P. A. (2017). 3-minute smartphone-based and tablet-based psychomotor vigilance tests for the assessment of reduced alertness due to sleep deprivation. *Behavior Research Methods*, *49*(3), 1020-1029.

Fogel, D. (2006). *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence* (3rd ed.). Piscataway, NJ: IEEE Press.

Lamond, N., Dawson, D., & Roach, G. D. (2005). Fatigue assessment in the field: Validation of a hand-held electronic psychomotor vigilance task. *Aviation, Space, and Environmental Medicine*, *76*(5), 486-489.

Hursh, S. R., Balkin, T. J., Miller, J. C., & Eddy, D. R. (2004). The fatigue avoidance scheduling tool: Modeling to minimize the effects of fatigue on cognitive performance. *SAE Transactions*, *113*(1), 111-119.

Hursh, S. R., Redmond, D. P., Johnson, M. L., Thorne, D. R., Belenky, G., Balkin, T. J., Eddy, D. R. (2004). Fatigue models for applied research in warfighting. *Aviation, Space, and Environmental Medicine*, *75*(3), A44-A53.

Kancherla, B., Upender, R., Collen, J. F., Rishi, M. A., Sullivan, S. S., Ahmed, O., Peters, B. R. (2020). Sleep fatigue and burnout among physicians: An American Academy of Sleep Medicine position statement. *Journal of Clinical Sleep Medicine*, *16*(5), 803-805.

Liu, J., Ramakrishnan, S., Laxminarayan, S., Balkin, T. J., & Reifman, J. (2017). Real-time individualization of the unified model of performance. *Journal of Sleep Research*, *26*, 820-831.

Mallis, M. M., Mejdal, S., Nguyen, T. T., & Dinges, D. F. (2004). Summary of the key features of seven biomathematical models of human fatigue and performance. *Aviation, Space, and Environmental Medicine*, *75*(3), A4-A14.

National Academies of Sciences, Engineering, and Medicine. (2016). *Commercial motor vehicle driver fatigue, long-term health, and highway safety: Research needs*. Washington, DC: National Academies Press.

Rajdev, P., Thorsley, D., Rajaraman, S., Rupp, T. L., Wesensten, N. J., Balkin, T. J., & Reifman, J. (2013). A unified mathematical model to quantify performance impairment for both chronic sleep restriction and total sleep deprivation. *Journal of Theoretical Biology*, *331*, 66-77.

Ratcliff, R., & Van Dongen, H. P. A. (2018). The effects of sleep deprivation on item and associative recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(2), 193-208.

Scrucca, L. (2013). GA: A package for genetic algorithms in R. *Journal of Statistical Software*, 53(4), 1-37.

Scrucca, L. (2017). On some extensions to GA package: Hybrid optimisation, parallelisation and islands evolution. *The R Journal*, 9/1, 187-206.

Tucker A.M., Whitney P., Belenky G., Hinson J.M., Van Dongen H.P. (2010). Effects of sleep deprivation on dissociated components of executive functioning. *Sleep*, *33*(1), 47-57.

Van Dongen, H. P. A. (2004). Comparison of mathematical model predictions to experimental data of fatigue and performance. *Aviation, Space, and Environmental Medicine*, *75*(3), A15-A36.

Whitney, P., Hinson, J. M., Jackson, M. L., & Van Dongen, H. P. A. (2015). Feedback blunting: Total sleep deprivation impairs decision making that requires updating based on feedback. *Sleep*, *38*, 745-754.