



Longitudinal Classification of Mental Effort Using Electrodermal Activity, Heart Rate, and Skin Temperature Data from a Wearable Sensor

William Romine¹(✉), Noah Schroeder², Anjali Edwards¹, and Tanvi Banerjee³

¹ Department of Biological Sciences, Wright State University, Dayton, OH 45435, USA
romine.william@gmail.com, edwards.348@wright.edu

² Department of Leadership Studies in Education and Organizations, Wright State University,
Dayton, OH 45435, USA

noah.schroeder@wright.edu

³ Department of Computer Science and Engineering, Wright State University,
Dayton, OH 45435, USA

tanvi.banerjee@wright.edu

Abstract. Recent studies show that physiological data can detect changes in mental effort, making way for the development of wearable sensors to monitor mental effort in school, work, and at home. We have yet to explore how such a device would work with a single participant over an extended time duration. We used a longitudinal case study design with ~38 h of data to explore the efficacy of electrodermal activity, skin temperature, and heart rate for classifying mental effort. We utilized a 2-state Markov switching regression model to understand the efficacy of these physiological measures for predicting self-reported mental effort during logged activities. On average, a model with state-dependent relationships predicted within one unit of reported mental effort (training RMSE = 0.4, testing RMSE = 0.7). This automated sensing of mental effort can have applications in various domains including student engagement detection and cognitive state assessment in drivers, pilots, and caregivers.

Keywords: Cognitive load · Wearable sensor · Mental effort · Machine learning · Cognitive assessment

1 Introduction

Researchers often strive to measure how focused someone is on a task, or how much mental effort they are putting into it. One domain where this is an important question is education and the study of learning. For more than three decades many researchers interested in this question, or related questions, have relied on a prominent theory called Cognitive Load Theory (CLT; [1–3]). According to CLT, we can put mental effort towards learning the salient material, known as intrinsic cognitive load, or towards other features of the instruction that do not support the learning task, known as extraneous cognitive

load [1, 3, 4]. Researchers suggest that the complexity of the task and the learner's level of prior knowledge in the subject determine the amount of mental effort that will be needed to learn the material and thus determine the intrinsic cognitive load, whereas mental effort put into parsing non-supporting elements of the instruction, such as interesting but ultimately unrelated stories, or visually searching for references needed to understand components of the learning materials, determine the extraneous cognitive load [1–4]. Since the working memory is limited in both capacity [5, 6] and duration [1, 3], CLT suggests that it is important to minimize the mental effort learners have to expend on tasks that are not essential to learning the material [7].

Cognitive load theory is well-established in the education literature, with a number of highly cited papers centering on the theory (e.g., [7–9]). Unsurprisingly, CLT has been used to theoretically support a number of specific task design principles, such as the worked example effect, the redundancy effect, and the split-attention effect [2, 4], and has become widespread outside of the educational psychology literature, appearing, for example, in the medical education literature as well [10–12]. The notion of cognitive load is an important theoretical paradigm in many types of educational research, but a lingering question persists in the CLT literature: how do we measure cognitive load?

The construct of cognitive load, as explained by CLT, is relatable to many; however, the measurement of such a construct has been a psychometric challenge for more than a decade. Researchers have used methods as varied as self-reports [13], eye-tracking measures [14], secondary task techniques [15], or physiological data [16]. Outside of the education literature researchers have measured a similar construct, mental workload, using similar methodologies like self-reports [17] or physiological measures [18] like facial skin temperature [19].

Recently, perhaps due to the increasing accessibility of wearable sensors or the psychometric issues associated with current methods for cognitive load assessment [20–22], researchers have been using physiological measurements and investigating their relation to learning relevant outcomes. Some of this work has shown promising results. For example, in relation to learning relevant processes, [23] used heart rate variability as an indicator of sustained attention. In addition, [24] measured electrodermal activity and examined these data in relation to self-reported emotional engagement. They found that students who were more engaged showed more frequently high levels of electrodermal activity. Taking a multimodal physiological approach, [25] differentiated between when students worked on high, moderate, and low mental effort activities, and further were able to predict a user's self-reported mental focus. It is noteworthy, however, that not all studies have shown such promising results. For example, examining task complexity in relation to physiological measures, [26] found that electrodermal activity and heart rate mean scores did not differ depending on the complexity of the task.

As noted, the use of wearable sensors to collect physiological data in relation to education-relevant outcomes is becoming more widespread in the literature. While there have been some promising results, the literature also shows some null results, highlighting the complexity of this area of work. When looking at recent studies [20–26], an important missing piece is understanding how we can use these data to track mental effort in an individual over extended periods of time, and the diagnostic utility of easily-obtainable physiological measures like EDA, skin temperature, and heart rate towards

this goal. In this case study, we utilize a longitudinal interpretable machine learning approach to understand how these data can be used to track the mental effort of an individual student in the context of both school activities and activities of daily living.

2 Methods

2.1 Study Design

In this study we sought to understand how EDA, skin temperature, and heart rate can be used to learn trends in mental effort for a single participant, and the extent to which we can model this in a robust way. We were first interested in using interpretable machine learning models to understand relationships between the participant's EDA, skin temperature, and heart rate measures and her reported mental effort. Second, we were interested in the diagnostic strength of these measures, and their efficacy in predicting mental effort in the context of future activities. To satisfy these goals, we used a longitudinal $n = 1$ case study design [27]. The goal of a case study is to generate rich description of a single case, which typically constitutes a single participant or entity [28]. Since our aim in this study was to evaluate the efficacy of a device for long-term monitoring of mental effort, it made sense to focus on a single participant over an extended time period. Researchers who place a premium on generalizability across contexts argue that a case study is disadvantaged by its focus within a single specific context [28]. However, [28] argues that this focus on a specific context is a strength in that it supports more accurate generalization to similar contexts. With the fields of psychology and medicine focusing less on giving general answers applying to everyone, and more on individualizing care, it is little surprise that the $n = 1$ design has increased in popularity in the medical research community [29, 30].

2.2 Description of the Case and Instrumentation

Since the focus of this study was to detect mental effort associated with school-related activities as well as activities of daily living, we chose an undergraduate university student as the case. This student was 19 years of age. She was a Psychology major with a concentration in Neuroscience in the second year of her undergraduate degree. Her primary hobbies included painting and spending time with her dog. Through the study, she identified her school-related activities, painting, learning to groom her dog with clippers and scissors, and watching brain games with her family as activities constituting high mental effort, and spent 48% of her time engaging in these types of activities. The remainder of her time was spent on low self-reported mental effort activities including eating, talking on the phone, watching television, driving, running errands, napping, and walking her dog. A total of 37 h, 33 min, and 34 s of data were collected. At a sampling rate of one sample per second, this constituted 135,214 total observations. These data were collected over approximately 3 weeks during the last half of the Spring 2020 semester.

The methodology relied on matching physiological data for EDA, skin temperature, and heart rate to self-reported data for mental effort dedicated to specific activities. EDA,

skin temperature, and heart rate data were collected using the Empatica E4 wristband. The E4 measures blood volume pressure, heart rate, interbeat interval, skin temperature, and 3-axis acceleration. The E4 sampled EDA at 4 Hz, skin temperature at 4 Hz, and calculated heart rate (1 Hz) based on the BVP signal (64 Hz). In order to minimize noise in the data, we elected to downsample the EDA and skin temperature signals to 1 Hz in order to match the heart rate signal.

The participant was asked to place the E4 band on her wrist approximately 3 cm from the base of the hand. She indicated that she wore the device on her right wrist since she was left-handed. As she engaged in different activities throughout the day while wearing the device, she logged them in a journal along with assigning a measure of mental effort to each activity. Mental effort was self-reported on a Likert scale of 1–4, where a “1” indicated very low effort, a “2” indicated low effort, a “3” indicated high effort, and a “4” indicated very high effort. Individual activities varied in length from under a minute to over an hour. During the course of her activities, the student’s data transitioned between low (1 and 2) and high (3 and 4) mental effort states 30 times.

2.3 Markov Switching Regression Model

The goal of modeling was two-fold: (1) to generate longitudinal predictions for mental effort and evaluate their robustness, and (2) to understand the role of measured EDA, skin temperature, and heart rate in generating these predictions. In light of these goals, we utilized the Markov switching dynamic regression model [31], which is an interpretable machine learning model that describes how an outcome changes its state over time. At their most basic level, Markov models predict a current state based on the previous state and a transition probability matrix. Markov switching models build upon this by allowing incorporation of state-specific relationships, thereby improving our understanding of how the physiological parameters relate to mental effort within each state.

Given our interest in a device that is able to distinguish between high and low states of mental effort, we utilized a 2-state Markov switching model. We tested models with four hierarchical levels of complexity: (1) a 2-state intercept-only model, (2) a 2-state model which held the effects of EDA, heart rate, and skin temperature constant across state, (3) a 2-state model which allowed the effects of EDA, heart rate, and skin temperature to switch across states, and (4) a 2-state model allowing for switching effects and variances. The likelihood ratio test was used to test the null hypothesis that adding an additional level of complexity did not improve model fit (95% confidence level used). The generalized r-square was calculated from the ratio of deviance values from the null and alternative models as a measure of the extent to which the alternative model improved fit over the null model.

Upon arriving at the best model using the above procedure, our interest shifted to evaluating the model’s ability to provide robust temporal predictions. For validation, we fit the model to the first 22 h (58%) of the data, and tested that model on the final 16 h (42%) of the data. The root mean square error and mean absolute error were used to compare the fit of the raw output between the training and testing sets. We also discretized the reported mental effort and output to 2 states in order to evaluate the model’s strength as a classifier based on its precision, recall, and F1 measure for the training and testing sets.

3 Results

3.1 Descriptive Analysis

The participant spent 9 h 18 min and 29 s in activities requiring very low mental effort and 10 h 8 min and 6 s in activities requiring low mental effort. 12 h 48 min and 46 s were spent at high mental effort, and 5 h 18 min and 13 s were spent at very high mental effort. Small but significant differences in EDA, skin temperature, and heart rate were found between each level of mental effort (Table 1).

Table 1. Average values for EDA, skin temperature, and heart rate at each reported level of mental effort

Mental Effort	Time (s)	EDA (μ S)		Temperature ($^{\circ}$ C)		Heart Rate (bpm)	
		Mean	SD	Mean	SD	Mean	SD
1	33,509	0.12	0.08	32.55	1.56	94.74	17.42
2	36,486	0.12	0.06	31.70	2.64	97.79	21.50
3	46,126	0.15	0.18	32.30	1.80	91.65	17.59
4	19,093	0.13	0.04	32.40	1.93	92.55	16.14
Total	135,214	0.13	0.12	32.21	2.05	94.20	18.67

A MANOVA omnibus test indicated at least one significant difference in the multivariate mean across levels of mental effort (Wilk's $\Lambda = 0.94$, $F_{9,329061} = 948.97$, $p << 0.001$, $\eta^2_{\text{partial}} = 0.021$). Univariate ANOVA tests indicated that skin temperature exhibited the largest differences between levels of mental effort ($F_{3,135210} = 1159.92$, $p << 0.001$, $\eta^2_{\text{partial}} = 0.025$). EDA ($F_{3,135210} = 649.63$, $p << 0.001$, $\eta^2_{\text{partial}} = 0.014$) and heart rate ($F_{3,135210} = 810.04$, $p << 0.001$, $\eta^2_{\text{partial}} = 0.018$) also exhibited significant differences, but the effect sizes were less than that for skin temperature. Due to the large number of observations, Scheffe tests indicated that all differences between subsequent levels of mental effort were significant at the 99% confidence level. However, given the longitudinal nature of the data, it was difficult to specify how the physiological data support classification of high and low mental effort states over time using the MANOVA procedure.

3.2 Longitudinal Modeling of Mental Effort Using Physiological Data

Contribution of Physiological Measures: The log-likelihood tests (Table 2) suggested that the most complex model, allowing effects and variances to switch across states, provided the best fit to the data, and offered a significant improvement over the intercept-only null model ($R^2 = 0.023$, $\chi^2_{\text{partial, df=7}} = 4195.0$, $p << 0.001$).

Adding EDA, skin temperature, and heart rate as constant effects to the 2-state intercept-only model resulted in a significant improvement in the model ($R^2 = 0.012$,

Table 2. Hypothesis tests for significance of change in model fit as model complexity increased. Type 1 tests were used for calculating R^2_{partial} and χ^2_{partial} . R^2 and χ^2 were calculated with respect to the intercept-only null model

Two-state Markov model	R^2	χ^2	R^2_{partial}	χ^2_{partial}
Intercept-Only (k = 5)				
Intercept + Constant effects (k = 8)	0.012	2200.4*	0.012	2200.4*
Intercept + Switching effects (k = 11)	0.021	3838.2*	0.0090	1637.8*
Intercept + Switching effects and variances (k = 12)	0.023	4195.0*	0.0020	356.8*

* $p < 0.001$

$\chi^2_{df=3} = 2200.4$, $p < 0.001$). Allowing the effects of EDA, skin temperature, and heart rate to switch between states 1 and 2 resulted in a further improvement ($R^2_{\text{partial}} = 0.009$, $\chi^2_{\text{partial}, df=3} = 1637.8$, $p < 0.001$). Finally, allowing variances to switch across the two states resulted in a smaller, but nonetheless significant, improvement in model fit ($R^2_{\text{partial}} = 0.002$, $\chi^2_{\text{partial}, df=1} = 356.8$, $p < 0.001$).

The 2-state model with state-dependent effects and variances (Table 3) showed that State 1 was associated with low mental effort (Intercept = 1.508, SE = 0.002, $z = 800.4$, $p < 0.001$), and State 2 was associated with high mental effort (Intercept = 3.296, SE = 0.002, $z = 1823.7$, $p < 0.001$). With this qualification, we can begin to understand how this student's EDA, skin temperature, and heart rate changed with mental effort within these two states as well as across the two states. Within State 1, an increase in mental effort was accompanied by a decrease in skin temperature and EDA, and an increase in heart rate. Skin temperature provided the strongest diagnostic for mental activity (Coef = -0.089 , SE = 0.002, $z = -52.0$), followed by heart rate (Coef = 0.042, SE = 0.002, $z = 24.2$). EDA was significant (Coef = -0.034 , SE = 0.003, $z = -11.0$), but nonetheless had a weaker effect size than skin temperature and heart rate. This ordering of importance matched the conclusions from the MANOVA test.

Upon transition to State 2, EDA retained its negative relationship with mental effort (Coef = -0.025 , SE = 0.001, $z = -17.1$), and heart rate retained its positive relationship (Coef = 0.016, SE = 0.002, $z = 8.1$). However, skin temperature switched to a being positive indicator of mental effort (Coef = 0.018, SE = 0.002, $z = 8.9$) in State 2. The ordering of importance also changed from State 1. When the participant entered State 2, EDA became the strongest diagnostic, followed by skin temperature and heart rate.

Utility for Prediction. From the perspective of correct classification, our data indicate that the Markov switching regression model has high predictive utility both on the training and testing sets. The model predicted whether the participant was in a high or low state of mental effort with high accuracy (Accuracy_{train} = 0.9995, F1_{train} = 0.9995, Accuracy_{test} = 0.9996, F1_{test} = 0.9996). However, much of this was due to the fact that reported mental effort in association with certain activities was stable and sustained over extended time periods. This is illustrated by the model probabilities: given an initial state, the probability of staying in the same state was 0.99975, and the probability of transitioning to the other state was 0.00025. This means that when the model encountered

Table 3. Parameter estimates for the 2-state Markov switching model with switching effects and variances

Mental Effort	Feature	Coef	SE	z
State 1				
	EDA	-0.034	0.003	-11.0
	TEMP	-0.089	0.002	-52.0
	HR	0.042	0.002	24.2
	Intercept	1.508	0.002	800.4
	Variance	0.488	0.001	
State 2				
	EDA	-0.025	0.001	-17.1
	TEMP	0.018	0.002	8.9
	HR	0.016	0.002	8.1
	Intercept	3.296	0.002	1823.7
	Variance	0.454	0.001	
Transition	p11	0.99976	0.00006	
Probabilities	p21	0.00025	0.00006	

a transition from one level of mental effort to another, it tended to misclassify the initial observation within the new activity. However, once the model observed that initial observation, it tended to classify the rest of the observations correctly until it encountered another transition. It is for this reason that the Intercept-Only model (RMSE = 0.48, MAE = 0.46) predicted nearly as well as the Switching Effects and Variances model (RMSE = 0.47, MAE = 0.45) despite its lack of explanatory utility. The Switching Effects and Variances model predicted the testing set (RMSE_{test} = 0.70, MAE_{test} = 0.61) slightly less accurately than the training set (RMSE_{train} = 0.40, MAE_{train} = 0.32), illustrating some deterioration in performance when predicting into the future. However, these measures of fit sat within one unit of reported mental effort, illustrating the model's usefulness for classification of discrete states of mental effort both in the training and testing sets.

4 Discussion and Conclusions

Our findings suggest that the Markov switching model is useful as an explanatory tool for understanding the diagnostic utility of EDA, skin temperature, and heart rate for measuring mental effort. Providing that information about the participant's previous state is available, we can expect this model to perform well in predicting the participant's state at the next time point. This means that for extended activities, we will be able to discern the participant's level of mental effort at the next time point with reasonable certainty.

However, the utility of Markovian assumptions reduces when we do not have knowledge of the previous state, or if that knowledge is highly tentative. The utility of this framework could be improved if it were combined with another machine learning approach which is less sensitive to prior states. Previous work suggests that machine learning models invoking the assumption that the data are independent and identically distributed (i.i.d.) may be useful for detecting transitions between states [25]. For example, a simple logistic regression model applied to this data set using EDA, skin temperature, and heart rate as main effects (Accuracy = 0.55, F1 = 0.50) was able to detect 2 of the 30 total transitions in the data despite performing relatively poorly as a classifier. In this sense, traditional machine learning approaches could be used to generate time-independent predictions, and then the Markov model could act as a smoother over the temporal dimension which would improve the coherence of predictions while a user is within a particular state of mental effort. Our next steps include exploring linear dynamical systems and variants that incorporate both the temporal information, as well as utilize the i.i.d. nature to be able to detect both stability and transitions with high certainty.

Previous work has shown the promise of using physiological data collected from wearable sensors to facilitate automated monitoring mental effort and cognitive load [23–26], and [25] proposed the application of this framework toward development of an Educational Fitness Sensor (EduFit) system to help students track the duration and quality of their studies in real time. However, for EduFit to have utility as a personal device, models have to work in less structured environments over relatively long time durations. This study shows that EDA, skin temperature, and heart rate have diagnostic utility in these types of less controlled settings. It has been argued that the EduFit system would enable building of personal understanding of one's study endeavors through interpretable biofeedback and enablement of personal accountability [25]. Beyond engagement in studies, we believe this type of system may also be useful in other contexts where mental effort is important such as fields involving high-stakes operation of machinery. Monitoring of mental effort may also be useful for detecting cognitive decline in gerontology contexts. Within any of these contexts, the ability to specify and train models which are accurate and robust over time is essential if EduFit is to be useful, and our data indicate that interpretable machine learning models specified for time series data provide a step in the right direction.

References

1. Paas, F., Sweller, J.: Implications of cognitive load theory for multimedia learning. In: *The Cambridge handbook of multimedia learning*, pp. 27–42. Cambridge University Press, New York (2014)
2. Sweller, J., Chandler, P., Kalyuga, S.: *Cognitive load theory*. Springer, New York (2011)
3. Sweller, J., van Merriënboer, J.J., Paas, F.: Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, pp. 1–32 (2019)
4. Sweller, J.: Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educ. Psychol. Rev.* **22**, 123–138 (2010). <https://doi.org/10.1007/s10648-010-9128-5>
5. Cowan, N.: The magical mystery four: how is working memory capacity limited, and why? *Curr. Directions Psychol. Sci.* **19**, 51–57 (2010)
6. Cowan, N.: The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* **24**, 87–114 (2001)

7. Sweller, J., Van Merriënboer, J.J., Paas, F.G.: Cognitive architecture and instructional design. *Educ. Psychol. Rev.* **10**, 251–296 (1998)
8. Sweller, J.: Cognitive load theory, learning difficulty, and instructional design. *Learn. Instr.* **4**, 295–312 (1994)
9. Chandler, P., Sweller, J.: Cognitive load theory and the format of instruction. *Cogn. Instr.* **8**, 293–332 (1991)
10. Young, J.Q., Van Merriënboer, J., Durning, S., Ten Cate, O.: Cognitive load theory: implications for medical education: AMEE Guide No. 86. *Med. Teach.* **36**, 371–384 (2014)
11. Leppink, J., Duviolier, R.: Twelve tips for medical curriculum design from a cognitive load theory perspective. *Med. Teach.* **38**, 669–674 (2016)
12. Van Merriënboer, J.J., Sweller, J.: Cognitive load theory in health professional education: design principles and strategies. *Med. Educ.* **44**, 85–93 (2010)
13. Paas, F.G.: Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J. Educ. Psychol.* **84**, 429–434 (1992). <https://doi.org/10.1037/0022-0663.84.4.429>
14. Cook, A.E., Wei, W., Preziosi, M.A.: The use of ocular-motor measures in a convergent approach to studying cognitive load. *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*, pp. 112–128 (2018)
15. Korbach, A., Brünken, R., Park, B.: Differentiating Different Types of Cognitive Load: a Comparison of Different Measures. *Educ. Psychol. Rev.* **30**, 503–529 (2018). <https://doi.org/10.1007/s10648-017-9404-8>
16. Antonenko, P., Paas, F., Grabner, R., van Gog, T.: Using electroencephalography to measure cognitive load. *Educ. Psychol. Rev.* **22**, 425–438 (2010). <https://doi.org/10.1007/s10648-010-9130-y>
17. Hart, S.G.: NASA-task load index (NASA-TLX); 20 years later. Presented at the Proceedings of the human factors and ergonomics society annual meeting (2006)
18. Charles, R.L., Nixon, J.: Measuring mental workload using physiological measures: a systematic review. *Applied Ergonomics.* **74**, 221–232 (2019)
19. Or, C.K., Duffy, V.G.: Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. *Occupational Ergonomics.* **7**, 83–94 (2007)
20. De Jong, T.: Cognitive load theory, educational research, and instructional design: some food for thought. *Instr. Sci.* **38**, 105–134 (2010)
21. Schroeder, N.L., Ceneci, A.T.: Do measures of cognitive load explain the spatial split-attention principle in multimedia learning environments? A systematic review. *J. Educ. Psychol.* (2019)
22. Zheng, R.Z.: *Cognitive load measurement and application: a theoretical framework for meaningful research and practice*. Routledge (2017)
23. de Avila, U.E.R., de França Campos, F.R., Leocadio-Miguel, M.A., Araujo, J.F.: 15 Mins of attention in class: variability of heart rate, personality, emotion and chronotype. *Creative Educ.* **10**, 2428 (2019)
24. Di Lascio, E., Gashi, S., Santini, S.: Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. In: *Proceedings ACM Interactive Mobile Wearable Ubiquitous Technology*, vol. 2, pp. 1–21 (2018). <https://doi.org/10.1145/3264913>.
25. Romine, W.L., et al.: Using machine learning to train a wearable device for measuring students' cognitive load during problem-solving activities based on electrodermal activity, skin temperature, and heart rate: development of a cognitive load tracker for both personal and classroom use. *Sensors.* **20**, 4833 (2020)
26. Larmuseau, C., Vanneste, P., Cornelis, J., Desmet, P., Depaeppe, F.: Combining physiological data and subjective measurements to investigate cognitive load during complex learning. *Frontl. Learn. Res.* **7**, 57–74 (2019)

27. Lundervold, D.A., Belwood, M.F.: The best kept secret in counseling: Single-case (N=1) experimental designs. *J. Counsel. Dev.* **78**, 92–102 (2000)
28. Stake, R.: *The Art of Case Study Research*. Sage Publications, New York (1995)
29. Morgan, D.L., Morgan, R.K.: Single-participant research design: bringing science to managed care. *Am. Psychol.* **56**(2), 119–127 (2001)
30. Foster, L.H.: A best kept secret: single-subject research design in counseling. *Counsel. Outcome Res. Eval.* **1**(2), 30–39 (2010)
31. Hamilton, J.D.: A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica J. Econ. Soc.* 357–384 (1989)