Challenges in Research Generalizability: The Need for Standardization of Performance Metrics and Methodology

Kathryn A. Feltman[1,2(✉)], Kyle A. Bernhardt[1,2], and Amanda M. Kelley[1]

[1] U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL, USA
{kathryn.a.salomon2.ctr,kyle.a.bernhardt2.ctr,
amanda.m.kelley.civ}@mail.mil
[2] Oak Ridge Institute for Science and Education, Oak Ridge, TN, USA

Abstract. The proliferation of psychological and psychophysiological metrics, data collection techniques, and data analysis strategies used throughout psychological research of operator performance presents cross-study synthesis complications. Currently, the lack of defined and established standardizations in psychological and psychophysiological research continues to present challenges to researchers studying an array of interrelated constructs. Without standardizations, differences in measurement implementation, data reduction techniques, and the interpretation of results make it difficult to directly compare studies and reach unequivocal conclusions while synthesizing literature and transfer laboratory findings to field-ready applications.

Keywords: Psychophysiological · Subjective · Standardization · Generalizability

1 Introduction

Scientists and researchers in all fields require, to some degree, standardization of the methodologies and measurements used in order to generalize and compare findings. The field of psychology, including psychophysiology and cognitive science, is challenged by the often times unobservable, latent constructs studied. While psychophysiological measurements and methods have been well established, their integration into research on emerging technologies, particularly those studied in applied environments, is ongoing. Of particular interest to the military aviation research community are measurements that may be integrated into the vehicle to provide real-time monitoring of the operator's state (e.g., fatigued, cognitive overload). In order to make advancements in this area, standardization of cognitive and subjective measures for purposes of determining construct validity as well as standardization of methodology for psychophysiological measures (e.g., data reduction) is needed. This will allow the community of researchers to more easily interpret findings relative to their own work and ultimately drive towards solutions to the shared mission.

Presently, several researchers and journals have begun examining the issues surrounding a lack of standardized methods, to include problems in replicability, the need for adequately powered studies, and increased openness and transparency in research (e.g., [1–3]) within the fields of psychology, neuroscience, and

F.F.-H. Nah and C.-H. Tan (Eds.): HCIBGO 2017, Part I, LNCS 10293, pp. 78–89, 2017.
DOI: 10.1007/978-3-319-58481-2_7

psychophysiology. These, and other articles are a great reference for researchers looking to follow research best practices. The present paper will focus more closely on the problems of standardization and generalizability in regards to applied research. Researchers working in applied settings, such as military and government funded laboratories, rely highly on the ability to integrate other researchers' findings into research that can be used to solve specific and applied problems. However, the ability to do so with accuracy requires that the literature drawn upon follows some degree of standardization and has generalizable results. This paper examines this issue in the specific context of incorporation of cognitive and psychophysiological measures into operator monitoring in military settings.

2 Need for Similarity of Subjective and Cognitive Measures

The volume of valid and reliable cognitive tests and measures available for researchers is substantive. Often, the decision of which instruments to employ is determined by a number of factors in addition to psychometric properties: setting of data collection/experiment, limitations on time, equipment required, limitations on physical space, availability of trained test administrators, and cost. While there are certainly benefits to an expansive library of assessments to choose from, the degree of comparability across studies can be compromised thus resulting in misleading conclusions or seemingly contradictory results between studies. An example of this is the focal point of a recent publication on the operational definition of mild cognitive impairment following transient ischaemic attack and stroke [4]. The authors illustrate how different valid and accepted methodologies for determining mild cognitive impairment (i.e., three different cut-off scores from a neuropsychological test battery) led to varied results and conclusions including a diverse set of resultant incidence rates and relative risk ratios. Similarly, the International Collaboration on Mild Traumatic Brain Injury (mTBI) Prognosis published its recommendations with respect to methodological challenges in research [5]. Their comprehensive and critical review of the literature from 2001–2012 found 66 different operational definitions of mTBI in 101 articles regarding mTBI prognosis. The interchangeable use of the parallel terminology given to this vast expanse of definitions ultimately impedes effective communication among researchers as well as overall knowledge advancement.

Inconsistencies in subjective and cognitive measures across the literature produce difficulties in creating a standardized approach to studying phenomenon of interest to the military community. For example, the different branches of the military often face similar research questions, such as how to counteract fatigue in sustained operations. While different laboratories may utilize different approaches in studying the topic (e.g., one laboratory looking into medications to promote sleep, another laboratory looking into medications to promote wakefulness) inconsistency of measures used to determine fatigue levels will create difficulties in applying and comparing results between laboratories. By using a standardized set of subjective or cognitive measures when assessing a construct such as fatigue or cognitive workload, comparisons between laboratories become possible. Standardized approaches to research regarding physiological

monitoring are also lacking, particularly in regards to applied settings such as military research. For example, a number of different researchers utilize different procedures in physiological data collection, which can also result in inconsistencies in findings and an inability to generalize results.

3 Methodological Differences in Psychophysiological Research

While several articles related to best practices of psychophysiological measurement exist (e.g., see: [6, heart rate variability; 7, electroencephalography; 8, respiration]), different methods for collecting psychophysiological data continue to persist throughout the literature. While there are often several practical reasons for using different methodology, such as different electrode placement to reduce the likelihood of movement artifacts in a study where participants are ambulatory or in a vehicle (e.g., [9]), the different methods used create inconsistency in research practices, particularly when examining the same underlying concept. For example, three separate articles each examined cognitive workload through cardiovascular activity to assess physiological changes in response to changes in task demands [10–12]. The three articles each reported either a different electrode lead placement, or did not report electrode placement at all. The most commonly recommended lead placement for psychophysiological research is a three-lead placement, based on Einthoven's triangle theory [13]. However, different lead placements are frequently observed in research articles, such as leads applied to the sternum or leads applied to the clavicles and lower left or right rib.

Furthermore, lead placement should be determined with consideration of the type of data analyses planned, such as a researcher planning to examine heart rate variability (HRV) data, which is frequently seen within the literature for a means of assessing operator cognitive state. The ability to obtain meaningful data for HRV analyses is dependent on the integrity of cardiac signal collected [6]. The quality of signal that is detected is influenced by where the leads are placed [14]. When considering the transition of laboratory monitoring into field-deployable monitoring, standardized methods of data collection, including lead placement, will assist in the interpretability and proper analyses of the data that is collected from any given location and thus increase the generalizability of the results. Improper measurement techniques may result in the adequate collection of meaningful data, which can then obscure the results and reduce the generalizability of the findings to other settings [15]. This is a point that researchers who are looking to move physiological monitoring from within the laboratory to field settings should keep in mind. For example, one study compared three mobile ECG recording devices for measuring R-R intervals and HRV, and found that the HRV analyses obtained by the devices were inconsistent and not recommended for use within research applications [16]. Thus, care should be taken in determining methods to be used for ECG data collection, including determination of electrode placement and recording devices, and standardized methods should be used as the science of identifying operator state through physiological monitoring is still in its infancy.

Similarly, several studies of workload and engagement using electroencephalography (EEG) have reported the use of different electrode sites for data analyzed. Some examples of different electrode sites used included the following: F3, F4, C3, and C4 [17]; Cz, Pz, P3,

and P4 [18]; and Fz, Pz, O1, and O2 [9]. Each of these studies provide valuable information and insight into brain activity in response to various tasks; however, with a goal of moving towards psychophysiological measures that can be used to monitor an operator's state in real-time, examining the same EEG sites is paramount to progress. For example, in a recent article Cohen [19] discusses that researchers should strive to find a balance between replicating previous findings and producing new ones. The reproduction of existing findings will provide further support for the use of real-time monitoring of operators, when researchers can demonstrate that specific electrode sites reliably result in changes in response to certain cognitive activity, which can then be transitioned into practice.

The effects that can result when different methods are used in psychophysiological research were highlighted in an article by Caccioppo and Tassinary [20]. In this classic article discussing the use of physiological measures in psychological research, the authors highlight a study where the psychophysiological measurement was electrodermal response. Here it was shown that the conclusions drawn from the data differed depending on how the electrodermal response was expressed. Specifically, "when the electrodermal response was expressed in terms of the change in skin resistance, one individual (Subject A), appeared to show a response equal to that of another (Subject B). When the electrodermal response was expressed in terms of the change in skin conductance, however, Subject B appeared to show the stronger response to the stimulus. Thus, conclusions about the physiological effects of the stimulus were completely dependent on the measurement procedure used" (p. 17). This is similarly seen within EEG research, where the placement of the reference electrode can impact the quality and subsequently the interpretations and waveform analyses of the data recorded [21].

In addition to consistent electrode placement, researchers must also be careful to ensure that they are indeed manipulating the psychological construct they wish to assess. This was noted early by Ekman [22] in an article discussing that reliable differentiation of emotions through physiological measurements has been difficult to obtain given that a variety of additional emotions were likely elicited in the attempt to assess physiological response of the target emotion. Indeed, this problem persists today if researchers are not careful in their manipulations. For example, in a study examining the physiological response of the vigilance decrement, Pattyn and colleagues [23] discuss the different findings in vigilance research where some studies demonstrate a physiological response similar to "cognitive overload" (e.g., a decrease in heart rate variability) whereas others, including themselves, find a physiological response similar to that of "underload." These differences have been attributed to differences in event rates of the vigilance task, such that studies with higher event rates show a more characteristic overload response. Determining and properly manipulating the construct that researchers wish to address through physiological monitoring becomes a key concern with the continued research interest of real-time monitoring of operator state. That is, various laboratories studying this topic need to be certain they are identifying the same operator state in order for developed countermeasure technologies to be effectively implemented in operational settings outside of the laboratory. The issue of proper manipulation is critical not only for the validity and reliability of the data collected during the testing period, but also for the data that is collected during the baseline period as well.

4 Baseline Data Collection

The nature of collecting psychophysiological measurements to determine the state of an operator requires a comparison from the time period of interest to some baseline state. For example, in order to determine via physiological sensors if an operator is exhibiting signs of overload, a comparison must be made between the condition in which the operator is said to be overloaded and one in which he/she is in a normal, non-overload state. This baseline measurement allows researchers to observe physiological changes in response to specified stimuli, conditions in a flight simulator, or field mission phases [24]. Classically, baseline measurements have been implemented in two forms: *resting* (e.g., [25]) and *vanilla* [26]. A resting baseline entails that the participant remains in a wakeful, but relaxed, state and not exposed to the stimuli of interest for a predetermined duration. Logically, a resting baseline seeks to measure the lowest physiological activity; that is, to record a "basal" or "tonic" state to which experimental condition data are compared [24]. A vanilla baseline refers to measurements that are made while participants are performing a low demand version of the task [26]. Some researchers utilize a practice session of the task as a vanilla baseline. Other baselining methods have been proposed on the principle of regression to the mean. The logic implies that over repeated sampling from an individual's "population" of potential physiological responses, a stable mean estimate of that individual's normative state can be obtained [27]. For example, a *comprehensive* baseline refers to a baseline period consisting of a resting period, task instruction period, and a task practice period. Moreover, an *against-self* baseline has also been proposed. The against-self method utilizes the entire set of data for a participant (baseline, practice, and experimental task) and calibrates the experimental data section of interest against these data [27].

Researchers must critically examine several issues when selecting an appropriate baseline technique. For example, participants may experience anxiety in anticipation of performing the experimental task, resulting in an elevated physiological state. Gramer and Sprintschnik [28] evaluated the cardiovascular activity of participants before having to give a 5-min public speech. For participants that were informed of the task, the anticipation of waiting to perform the speech increased blood pressure. Similarly, Davidson, Marshall, Tomarken, and Henriques [29] found elevated heart rate when individuals were in the anticipation stage of having to give a speech, with those possessing characteristics of social phobia exhibiting a larger increase in heart rate. Thus, depending on the individual, some may experience elevated physiological baseline activation prior to performing a task. This situation may very well extend into military laboratories using aviators as subjects. Flight simulations are often manipulated to induce stressful flight conditions. For instance, simulator weather modifications, such as high winds and reduced visibility, produce higher workload flights for pilots [30]. Consequently, if an aviator becomes aware of a potentially difficult flight, he/she may exhibit increased pre-flight physiological arousal and skew baseline measurements.

Moreover, studies have reported the tendency of resting baseline measures to fluctuate over time. In their study, Gramer and Sprintschnik [28] reported slight increases in participant cardiac measures over time, even before the anticipation manipulation. It has also been shown that measures of resting baseline activity in the cardiovascular system can vary from day to day as well [31, 32]. Wet electrode electrodermal activity recordings may also display

a drift during baseline acquisition and may require an adaption period before any data recording begins [33]. With these fluctuations, there is significant variation as to how long a baseline period should last. Recommendations of at least 10 min [26], but upwards of 15 min [36], for a resting baseline have been reported. Vanilla baselines of 10 min have also shown relatively good stability [26]. Stern and colleagues [24] give the recommendation that the resting baseline period should be, "long enough to provide a stable pre-stimulus level and long enough to provide sufficient data for appropriate analysis" (p. 50). Moreover, Keil et al. [34] stated, "The choice of baseline period is up to the investigators and should be appropriate to the experimental design" (p. 5). Thus, when examining the literature, one may find an extensive range of baseline recording lengths making results somewhat difficult to interpret between studies.

Other than length of the baseline period, the choice of baseline procedure can influence conclusions researchers draw from their data. In research that is attempting to classify operator states accurately, the baseline procedure used will likely, to some degree, influence the outcome of augmented cognition systems (e.g., adaptive automation) to accurately detect changes in the operator state relative to baseline. More specifically, the selection of a certain baseline technique can overemphasize changes particular operator state and underemphasize others [27]. This point was communicated by Fishel and colleagues [27] in an examination of baseline techniques in relation to real-time physiological monitoring of operators. Take, for example, two operator states considered to be anchored at two different physiological poles: overload and fatigue. High workload situations are typically accompanied by physiological arousal, while fatigue is accompanied by physiological depression [35]. Baselines of the opposite physiological pole may exaggerate operator states that lie in the other direction. That is, a resting baseline would be more sensitive to detecting physiological changes associated with an overload state and a vanilla/practice baseline would be more sensitive to detecting physiological changes associated with a fatigued state. On the other hand, baselines that are of a similar polarity would tend to underemphasize a response. For instance, a resting baseline would tend to be relatively insensitive to detecting a fatigue state accompanied by physiological deactivation because of the already low physiological arousal of the resting baseline.

Indeed, Fishel et al. [27] empirically explored whether resting and practice (vanilla) baselines overemphasized high workload and low workload states compared to the against-self method. In their study, participants underwent several physiological baseline procedures before performing a shooting task (high demand) and a surveillance task (low demand). Results indicated that, compared to the against-self method, the resting baseline technique showed a significant bias for detecting cardiac arousal on the shooting task. In contrast, the practice baseline demonstrated a significant bias to detecting lower cardiac arousal during the surveillance task compared to the against-self method. Thus, this study demonstrates how the methodological selection of a baseline can bias data to detect certain operator states.

From the above discussion, it can be inferred that the selection of physiological baselining procedures can severely hamper the comparison of results across studies and laboratories. Assume that two hypothetical military laboratories are each using measures of the cardiovascular system to support detecting changes in workload during simulated flights. Further assume Laboratory A decides to use a resting baseline to calibrate their data and

Laboratory B decides to use a task-practice baseline procedure. In general, Laboratory A is more likely to detect positive changes in workload than Laboratory B. That is, Laboratory B may fail to detect cardiac changes associated with increased cognitive workload more so than Laboratory A. Even though each laboratory may be testing under similar flight parameters and independent variable manipulations, the outcome results may not be comparable across laboratories and appear to be fairly inconsistent. In an applied setting, this inferred inconsistency has the potential to misinform decision makers and policy writers.

5 Data Analyses and Data Reduction

How the data itself is examined can play a crucial role in the replicability and generalizability of the research. The differences in data analyses become most apparent when examining the use of baseline data. Keil et al. [34] stressed that the removal baseline activity may result in distortions of electrophysiological data especially if experimental groups show differential activation patterns. Many psychophysiological researchers agree that baseline data is necessary to collect in order to determine changes in physiological response; however, a brief review of psychophysiological research quickly yields differences in how the baseline data collected was actually used in analyses. For example, some studies report baseline data being used to normalize the physiological measures collected throughout the study, by calculating the ratio of the average processed recording data and the baseline data (e.g., [9, 12]), whereas others report the use of baseline data as a comparison point for the remainder of data collected (e.g., report a change in baseline [36]).

Differences in methods used for data reduction and signal processing can also impede the generalizability of psychophysiological data relates to methods of data collection and data reduction. The sampling rate of psychophysiological signals can be found to vary from study to study within the literature. In an article examining different sampling rates when using respiratory sinus arrhythmia as a measure of heart rate variability, Riniolo and Porges [37] highlight that the importance of using the proper sampling rate, as the sampling rate chosen significantly affects "the ability to quantify accurately the amplitude of RSA because a slow sampling rate would be insensitive to small gradations when the amplitude of RSA is low" (p. 619). Thus, different sample rates can result in differences in the quality of data collected. Based on the Nyquist theorem [38] a sample should be taken at twice the maximum frequency expected to be encountered. Sample rates for different measures will naturally vary, such that changes in electrodermal or respiration activity are slow and can be sampled at lower rates, whereas changes in electrocortical or heart activity occur quicker and must be sampled at higher rates. Although it is not considered improper practice to sample physiological data at different rates (e.g., one researcher sampling HRV data at 256 Hz, with another sampling at 500 Hz), these differences will present alterations in the resolution and quality of the data [39], such that sampling HRV at 500 Hz would result in a greater resolution and more accuracy than a lower sampling rate. Weiergräber and colleagues [41] discuss some of the implications of differences in EEG sampling rates, and provide recommendations for best practices to follow. Additionally, they discuss that changes to sample rates can result in faulty frequency data and invalid results. Specifically, if sampling rates do not adhere to the Nyquist theorem, the frequency reconstruction becomes invalid

and interpretations of the data may become false. Weiergräber and colleagues [41] identify that a review of the literature of EEG studies revealed that EEG recordings were being done outside of the technical range of the equipment used, thus resulting in invalid analyses. Thus, it is crucial for researchers to understand the importance of sampling rates in regards to the variables of interest (e.g., examining gamma waveforms vs. alpha waveforms).

Filtering and data reduction practices are also integral not only to the replicability of research, but the quality of the findings reported. The methods that researchers use to preprocess data and remove artifacts can vary from study to study, thus changing the possible quality of data analyzed and presented. In Cohen's [19] recent article on replication and rigor in electrophysiology research, he discusses some of the problems surrounding artifact removal. Specifically, he highlights cautions to be considered when using algorithms for artifact removal, and recommends manual cleaning of the data over algorithms. However, the practicality of such manual methods may not be feasible in applied research where the goal is to develop field-ready devices that process data in real time. It may be the case that more research on the validity of electrophysiological artifact decontamination algorithm development needs to become available to the general research community.

6 Conclusions and Recommendations

While technology continues to advance at a rapid pace, with increased capabilities to monitor the physiological changes of an individual in a variety of settings, the need to maintain scientific integrity through standardized measurement techniques is paramount. Increased interest in continuous, real time monitoring of operators to either inform adaptive automation [41], monitor performance to assist in system design [42, 43], or to be used in training evaluation [44], requires first the ability to reliably identify the operator state through the desired metrics. This, of course, relies on the use of standardized measures and methods that can be applied across studies, scenarios, and laboratories. Through the use of standardized research practices, we will be able to advance from the laboratory to the field. Several other articles are available that provide thorough reviews and recommendations of how researchers can work towards conducting research that is rigorous and replicable, for which the reader is highly encouraged to peruse (e.g., [1–3]). However, in regards to conducting applicable and scientifically useful research that can be used for future implementation of real-time operator monitoring, a few suggestions are outlined below.

Researchers are encouraged to conduct thorough literature reviews, as well as engage in discussions with fellow researchers in the field to determine the best cognitive and subjective measures to use when assessing a specific cognitive construct. Researchers should do the same for determining how to properly design tasks that assess the cognitive construct they want to examine through physiological measurement. Indeed, both of these are encouragements for researchers to engage in some replicability of previous findings. In order for research to transition from strictly laboratory-based findings to technology that can be used in an operationally relevant manner (i.e., operator state monitoring through physiological assessment to determine pilot cognitive overload) there needs to be a consistency in the literature that reliably identifies that subjective measurement A, as well as psychophysiological measurement devices B and C, always produce X change in data when the

individual is placed in Y situation. Without such reproducible data, difficulties in transitioning research findings to applied settings will persist.

With respect to physiological baseline techniques, researchers should be cognizant of the underlying construct's essential nature. That is, some constructs are associated with physiological activation and some are associated with physiological deactivation. Researchers should carefully evaluate and justify their decision to utilize a certain baseline technique as opposed to using a technique out of convention. More importantly, in applied field research, the choice of baseline technique should reflect the operator's normal operating state. This prevents an introduction of positive or negative bias in operator state detection observed when polarized baseline techniques (e.g., resting) are used [27]. In general, applied researchers are less interested in changes from a resting state, but rather departures from a state in which the operator is under a normal operating progression. The former provides a basis for developing theory and generating research hypotheses, while the latter has direct implications for augmented system development and countermeasure deployment. Researchers should be explicit in their choice of baseline technique and provide a sound justification for employing the technique. Additionally, researchers should report and justify how experimental data were adjusted for baseline values (e.g., simple subtraction, change scores). Keil et al. [34] provides a thorough publication checklist for researchers using EEG methods, which includes baseline technique reporting.

Similar recommendations hold true for decisions in regards to sampling rates, data preprocessing, and data reduction. However, researchers are also encouraged to be open and detailed in their methodology used (see [19] for examples). Additionally, while researchers are sometimes constrained by either equipment or environment for the sampling rates they use in data collection, an explanation of why the decision was made to accept a lower sampling rate is encouraged, as well as a discussion on impacts that it may have had on the data quality, so that readers are fully aware of the reasoning behind such a decision.

The current state of the literature oftentimes shows divergent findings (e.g., variability in physiological response in measuring cognitive workload based on differences such as task length [45] or event rates [23]) on physiological measures of a cognitive construct, such as cognitive workload, which only further points to the need to follow similar research designs and protocols. This is also commonly seen in medical literature when an agreed-upon definition of a condition does not yet exist. This becomes essential as the field begins to transition into using devices that leave less of a "footprint" (e.g., reducing EEG measurements to just four electrodes). In order to be able to make the determination that only certain electrode sites are needed for detecting a change in cognitive workload, or that one psychophysiological measurement is enough to reliably detect a change in operator state, further work that demonstrates the reproducibility of this research is needed, and not only within laboratory settings, but also in operational settings.

References

1. Open Science Collaboration: estimating the reproducibility of psychological science. Science **349** (2015) http://dx.doi.org/10.1126/science.aac4716
2. Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R.: Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. **14**, 365–376 (2013)
3. Larson, M.J.: Commitment to cutting-edge research with rigor and replication in psychophysiological science. Int. J. Psychophysiol. **102**, ix–x (2016)
4. Pendlebury, S.T., Mariz, J., Bull, L., Mehta, Z., Rothwell, P.M.: Impact of different operational definitions on mild cognitive impairment rate and MMSE and MoCA performance in transient ischaemic attack and stroke. Cerebrovasc. Dis. **36**(5–6), 355–362 (2013)
5. Kristman, V.L., Borg, J., Godbolt, A.K., Salmi, L.R., Cancelliere, C., Carroll, L.J., Holm, L.W., Nygren-de Boussard, C., Hartvigsen, J., Abara, U., Donovan, J., Cassidy, J.D. Methodological issues and research recommendations for prognosis after mild traumatic brain injury: results of the international collaboration on mild traumatic brain injury prognosis. Arch. Phys. Med. Rehab. **95**(3), S265–S277 (2014)
6. Berntson, G.G., Bigger, J.T., Eckberg, D.L., Grossman, P., Kaufmann, P.G., Malik, M., Nagaraja, H.N., Porges, S.W., Saul, J.P., Stone, P.H., van der Molen, M.W.: Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology **34**, 623–648 (1997)
7. Pivik, R.T., Broughton, R.J., Coppola, R., Davidson, R.J., Fox, N., Nuwer, M.R.: Guidelines for the recording and quantitative analysis of electroencephalographic activity in research contexts. Psychophysiology **30**, 547–558 (1993)
8. Ritz, T., Dahme, B., Dubois, A.B., Folgering, H., Fritz, G.K., Harver, A., Kotses, H., Lehrer, P.M., Ring, C., Steptoe, A., Van de Woestijne, K.P.: Guidelines for mechanical lung function measures in psychophysiology. Psychophysiology **39**, 546–567 (2002)
9. Ryu, K., Myung, R.: Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. Int. J. Ind. Ergon. **35**, 991–1009 (2005)
10. Bonner, M.A., Wilson, G.F.: Heart rate measures of flight test and evaluation. Int. J. Aviat. Psychol. **12**(1), 63–77 (2002)
11. Durantin, G., Gagnon, J.F., Tremblay, S., Dehais, F.: Using near infrared spectroscopy and heart rate variability to detect mental overload. Behav. Brain Res. **259**, 16–23 (2014)
12. Hsu, B.W., Wang, M.J.J., Chen, C.Y.: Effective indices for monitoring mental workload while performing multiple tasks. Percept. Mot. Skills **121**(1), 94–117 (2015)
13. Berntson, G.G., Quigley, K.S., Lozano, D.: Cardiovascular psychophysiology. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (eds.) Handbook of Psychophysiology, pp. 182–210. Cambridge University Press, New York (2007)
14. Kligfield, P., Gettes, L.S., Bailey, J.J., Childers, R., Deal, B.J., Hancock, W., van Herpen, G., Kors, J., Macfarlane, P., Mirvis, D., Pahlm, O., Rautaharju, P., Wagner, G.S.: Recommendations for the standardization and interpretation of the electrocardiogram. Circulation **115**, 1306–1324 (2007)
15. Strube, M.J., Newman, L.C.: Psychometrics. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (eds.) Handbook of Psychophysiology, pp. 789–811. Cambridge University Press, New York (2007)
16. Weippert, M., Kumar, M., Kreuzfeld, S., Arndt, D., Rieger, A., Stoll, R.: Comparison of three mobile devices for measuring R-R intervals and heart rate variability: Polar S810i, Suunto t6 and an ambulatory ECG system. Eur. J. Appl. Physiol. **109**(4), 779–786 (2010)

17. Di Stasi, L.L., Diaz-Piedra, C., Suárez, J., McCamy, M.B., Martinez-Conde, S., Roca-Dorda, J., Catena, A.: Task complexity modulates pilot electroencephalographic activity during real flights. Psychophysiology **52**(7), 951–956 (2015)
18. Fairclough, S.H., Venables, L., Tattersall, A.: The influence of task demand and learning on the psychophysiological response. Int. J. Psychophysiol. **56**, 171–184 (2005)
19. Cohen, M.X.: Rigor and replication in time-frequency analyses of cognitive electrophysiology data. Int. J. Psychophysiol. **111**, 80–87 (2017)
20. Cacioppo, J.T., Tassinary, L.G.: Inferring psychological significance from physiological signals. Am. Psychol. **45**(1), 16–28 (1990)
21. Pizzagalli, D.A.: Electroencephalography and high-density electrophysiological source localization. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (eds.) Handbook of Psychophysiology, pp. 56–84. Cambridge University Press, New York (2007)
22. Ekman, P., Levenson, R.W., Friesen, W.V.: Autonomic nervous system activity distinguishes among emotions. Science **221**, 1208–1210 (1983)
23. Pattyn, N., Neyt, X., Henderickx, D., Soetens, E.: Psychophysiological investigation of vigilance decrement: Boredom or cognitive fatigue? Physiol. Behav. **93**, 369–378 (2008)
24. Stern, R.M., Ray, W.J., Quigley, K.S.: Psychophysiological Recording. University Press Inc., New York (2001)
25. Pollak, M.H.: Heart rate reactivity to laboratory tasks and ambulatory heart rate in daily life. Psychosom. Med. **53**, 25–35 (1991)
26. Jennings, J.R., Kamarck, T., Steward, C., Eddy, M., Johnson, P.: Alternate cardiovascular baseline assessment techniques: vanilla or resting baseline. Psychophysiology **29**(6), 742–750 (1992)
27. Fishel, S.R., Muth, E.R., Hoover, A.W.: Establishing appropriate physiological baseline procedures for real-time physiological measurement. J. Cogn. Eng. Decis. Making **1**, 286–308 (2007). doi:10.1518/15553407X255636
28. Gramer, M., Sprintschnik, E.: Social anxiety and cardiovascular responses to an evaluative speaking task: The role of stressor anticipation. Personality Individ. Differ. **44**, 371–381 (2008). doi:10.1016/j.paid.2007.08.016
29. Davidson, R.J., Marshall, J.R., Tomarken, A.J., Henriques, J.B.: While a phobic waits: regional brain electrical and autonomic activity in social phobics during anticipation of public speaking. Biol. Psychol. **47**(2), 85–95 (2000). doi:10.1016/S0006-3223(99)00222-X
30. Hart, S.G., Bortolussi, M.R.: Pilot errors as a source of workload. Hum. Factors **26**(5), 545–556 (1984)
31. Miller, S.B., Ditto, B.: Cardiovascular responses to an extended aversive video game task. Psychophysiology **25**(2), 200–208 (1988). doi:10.1111/j.1469-8986.1988.tb00988.x
32. Miller, S.B., Ditto, B.: Individual differences in heart rate response during behavioral challenge. Psychophysiology **26**(5), 506–513 (1989). doi:10.1111/j.1469-8986.1989.tb00701.x
33. Boucsein, W., Fowles, D.C., Grings, W.W., Ben-Shakhar, G., Roth, W.T., Dawson, M.E., Filion, D.L.: Publication recommendations for electrodermal measurements. Psychophysiology **49**(8), 1017–1034 (2012). doi:10.1111/j.1469-8986.2012.01384.x
34. Keil, A., Debener, S., Gratton, G., Junghofer, M., Kappenman, E.S., Luck, S.J., Luu, P., Miller, G.A., Yee, C.M.: Committee report: publication guidelines and recommendations for studies using electroencephalography and magnetoencephalograph. Psychophysiology **51**(1), 1–21 (2014)
35. Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., Bablioni, F.: Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue, and drowsiness. Neurosci. Biobehav. Rev. **44**, 58–75 (2014). doi:10.1016/jneubiorev.2012.10.003

36. Mehler, B., Reimer, B., Coughlin, J.F., Dusek, J.A.: Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. Transp. Res. Rec. **2138**, 6–12 (2009)
37. Riniolo, T., Porges, S.W.: Inferential and descriptive influences on measures of respiratory sinus arrthymia: sampling rate, R-wave trigger accuracy, and variance estimates. Psychophysiology **34**, 613–621 (1997)
38. Nyquist, H.: Certain topics in telegraph transmission theory. IEEE Trans. Commun. **47**, 617–644 (1928)
39. Bolek, J.E.: Digital sampling, bits, and psychophysiological data: A primer, with cautions. Appl. Psychophys. Biofeedback **38**(4), 303–308 (2013)
40. Weiergräber, M., Papazoglou, A., Broich, K., Muller, R.: Sampling rate, signal bandwidth and related pitfalls in EEG analysis. J. Neurosci. Methods **268**, 53–55 (2016)
41. Parasuraman, R.: Neuroergonomic perspectives on human systems integration: mental workload, vigilance, adaptive automation, and training. In: Boehm-Davis, D.A., Durso, F.T., Lee, J.D. (eds.) APA Handbook of Human Systems Integration, pp. 163–176. American Psychological Association, Washington, DC (2015)
42. Warm, J.S., Parasuraman, R., Matthews, G.: Vigilance requires hard mental work and is stressful. Hum. Factors **50**(3), 433–441 (2008)
43. RTO human factors, medicine panel task group: operator functional state assessment. Technical report, Research and Technology Organization (2004)
44. Borghini, G., Aricò, P., Graziani, I., Salinari, S., Sun, Y., Taya, F., Bezerianos, A., Thakor, N.V., Babiloni, F.: Quantitative assessment of the training improvement in a motor-cognitive task by using EEG, ECG, and EOG signals. Brain Topogr. **29**, 149–161 (2016)
45. Stuiver, A., Brookhuis, K.A., de Waard, D., Mulder, B.: Short-term cardiovascular measures for driver support: increasing sensitivity for detecting changes in mental workload. Int. J. Psychophysiol. **92**, 35–41 (2014)