Establishing Appropriate Physiological Baseline Procedures for Real-Time Physiological Measurement

Stephanie R. Fishel
Eric R. Muth
Adam W. Hoover
Clemson University

Abstract: The purpose of this study was to identify baseline procedures that are equally sensitive to negative and positive changes in heart rate variability, specifically in the context of moment-to-moment comparisons. Participants were engaged in a dual-task paradigm consisting of either a high- or low-workload primary task paired with a secondary task. Shifts between primary tasks occurred at set intervals. Average arousal was calculated for each primary task in relation to six different baseline procedures. A 2 x 6 repeated-measures ANOVA revealed a significant main effect for the baseline type, $F(1, 58) = 3.90$, $p = .04$, $\eta^2 = .08$. Planned comparisons revealed that certain procedures were biased to detect positive change and others negative change. The results of this study provide evidence that when developing physiologically based measures for detecting behavioral and/or cognitive state changes, the method used to calibrate these measures requires special consideration.

Introduction

The purpose of this study was to identify effective comparison conditions (i.e., baselines) for detecting state change in heart rate variability data. Researchers have struggled with change detection for many decades, and the problem has been defined and labeled in a variety of ways over time. Although the issue of physiological baselines is applicable across many disciplines and areas of study, the emphasis of this paper will be on how this issue is relevant to augmented cognition and the development of a closed-loop human-machine system. Specifically, this paper discusses the importance of appropriate physiological baseline, or calibration, procedures for use in paradigms in which accurate moment-to-moment, "real-time," comparisons of physiological data are made with the purpose of detecting cognitive and/or behavioral state changes.

The act of detecting physiological change traditionally involves comparing the information of interest (e.g., current physiological state) with a baseline period (Stern, Ray, & Quigley, 2001). The quality of the baseline against which a change is detected and interpreted is dependent on the methodology used to collect and establish the baseline data. As noted by Jennings, Kamarck, Stewart, Eddy, and Johnson (1992),

ADDRESS CORRESPONDENCE TO: Stephanie R. Fishel, Department of Psychology, Clemson University, sfishel@gmail.com.

Journal of Cognitive Engineering and Decision Making, Volume 1, Number 3, 2007, pp. 286–308. DOI 10.1518/155534307X255636.

"baselines contribute as much as response levels to change measures and thus require equal methodological care in their assessment" (p. 742). In general terms, the baseline period sets the standard against which the information of interest is compared. Setting the standard for any given system is typically referred to as *calibration*. We will use the term *calibration* throughout this paper to refer to the process of comparing the mean output of a given standardized period of data (i.e., baseline data) with the new data of interest (e.g., experimental data).

In psychophysiology literature, calibration usually consists of comparing data collected while an individual is at rest but awake for a given period with the data of interest – for example, during an experimental stimulus of interest (Hastrup, 1986; Jennings et al., 1992). Recently, some researchers have explored using a *vanilla* baseline technique, which replaces the resting baseline period with a simple, minimally demanding task to maintain consistent alertness and baseline stability (Jennings et al., 1992). Although both the resting and vanilla baseline techniques are commonly used when measuring physiological reactivity (e.g., an increase in heart rate), these methods of calibrating a physiological measure may result in biased state detection in comparison with other calibration methods when used to detect both increases and decreases in physiological activity that are required by augmented cognition systems.

The qualities of the calibration period will determine the level of bias present when the data of interest are interpreted (Jennings et al., 1992; Kamarck, Jennings, & Manuck, 1993) and the precision of the system as a whole in differentiating various states. The qualities refer to the content of the calibration period relative to the state changes of interest. For example, the calibration period could be the traditional resting baseline as described previously, or a period of combined states such as rest or light activity. Given the possibility of real-time wearable physiological measures, it could conceivably also consist of an individual's total physiological life history from when the system was first activated to the current time. The aforementioned calibration or comparison periods are very different; therefore, they have different uses, some of which are of interest to the area of augmented cognition.

Augmented cognition typically involves detecting positive and negative physiological state changes, particularly those in which performance may be compromised. For example, one physiological construct of interest, as it relates to performance, is autonomic arousal. Research has shown that changes in physiological indicators of arousal are often associated with changes in workload (Grossman, Stemmler, & Meinhardt, 1990; Porges, 1992). Changes in arousal in turn have effects on performance. Arousal that is too low has been associated with a lack of energy (Dickman, 2002), whereas high arousal has been associated with changes in attention and concentration (Dickman, 2002) as well as with increased muscle tension (Cathcart & Pritchard, 1998) and coordination difficulties (Robazza, Bortoli, & Nougier, 1998). Midrange arousal, on the other hand, has been associated with enhanced memory (Nielson, Yee, & Erickson, 2005) and accuracy (Dickman, 2002).

In augmented cognition scenarios, low physiological arousal (e.g., fatigue) and high physiological arousal (e.g., stress) may need to be monitored in order to prevent detrimental effects on performance. Because the two states differ dramatically and

are extreme opposites, calibration should allow for distinguishing those states from one another with accuracy. Accuracy, in this sense, entails both correctly identifying the presence of a state change and determining the actual magnitude and direction of that state change relative to the calibration or comparison period.

Researchers studying cardiovascular reactivity have noticed differences in their findings based on whether they used a resting baseline or a vanilla baseline in their calculations to assess the degree of cardiac reactivity (i.e., positive physiological change) when participants performed a variety of stimulating cognitive and psychomotor tasks (Jennings et al., 1992; Kamarck et al., 1993). Those results suggest that an even greater number of discrepancies may exist between calibration methods when trying to detect physiological change in both positive and negative directions. For example, if the mean heart rate during a resting baseline is compared with the mean heart rate during a period in which an individual is on the verge of falling asleep, the difference between the two would be rather small. The difference would increase slightly if a vanilla baseline were used instead.

If, on the other hand, the mean heart rate values obtained during the resting baseline and the vanilla baseline are compared with the mean heart rate during a period in which an individual is experiencing stress, the difference in mean heart rate between the resting baseline and the stress would be larger than the difference between the mean heart rate values obtained during the vanilla baseline and the period of stress (Jennings et al., 1992). Those differences point to the need to carefully consider what state changes are of interest in a closed-loop system and how to equally optimize detection of them so that the system is "honest" rather than biased in the sense of being more sensitive to either positive or negative physiological change. As Jennings et al. (1992) pointed out, a baseline should be conceptualized as "a comparison condition rather than a basal state" (p. 743), and "differences between baseline levels might be functionally significant if they modified the size of responses to tasks" (p. 746).

The problem of accurate magnitude detection is analogous to the well-known physiological Law of Initial Values (LIV; Wilder, 1957), which states that the magnitude of the physiological response to a stimulus will depend on the initial prestimulus state of the individual. This effect is present for those physiological phenomena, such as heart rate and respiration, that are subject to rapid antagonistic action by the autonomic nervous system (Hord, Johnson, & Lubin, 1964).

Keeping the LIV in mind, if the calibration period (i.e., prestimulus state) of the physiological data is not well balanced and is biased toward the low end of a physiological measure, for instance, then some environmental conditions experienced by an individual that result in physiological responses that are excitatory in nature (e.g., stress) may appear to be more severe than other inhibitory physiological conditions of equally detrimental nature (e.g., fatigue).

In order to have a truly accurate augmented cognition system, the calibration period must be well balanced such that both the low- and high-end states can be equally and successfully detected. Achieving a well-balanced calibration period requires careful consideration of a principle often encountered in statistics and probability, *regression to the mean*. Regression to the mean is the idea that over time, repeated sampling

from a given population will ultimately lead to the leveling out of a given sample's mean. In the case of physiological measurement, each data point collected makes up a sample of an individual's physiological population for a particular measure. In general, when a greater number of cases are used to compute the mean of a sample taken from a given population, the sample mean tends to be more accurate, less error prone, and, in the end, more statistically powerful (Rosenthal & Rosnow, 1991).

It is useful to think of the calibration period as sampling from an individual's physiological population. Ideally, to detect phasic physiological changes, a measure would have access to many data points. Using several years of data, for example, one would find that the individual's physiological measure would have unsurpassable sensitivity and accuracy for detecting state changes. However, in practical terms, this is not currently possible.

Researchers studying cardiovascular reactivity have previously demonstrated that resting and vanilla baseline values obtained by averaging several values, such as multiple blood pressure recordings, have increased reliability over using a single value for a given period at both a resting or vanilla state when used to derive reactivity (Jacob & Shapiro, 1994; Kamarck et al., 1993). Specifically, the reactivity scores obtained were less inflated when aggregated baseline values were used across a variety of typical resting and vanilla baseline settings. Based on these results, which examined only positive changes relative to a calibration period, it is reasonable to believe that calibration periods used to examine both negative (e.g., fatigue) and positive (e.g., excitement) state changes should contain as many data points as possible. Further, those data points should be from the entire range of values that might be encountered so that the calibration value represents the median state from which change is measured.

For example, if we are interested in using the same calibration period to differentiate a fatigue state and a stressful state, then we might have an individual carry out tasks such as sleeping, perform light active tasks, and even exercise to push his or her heart rate to both ends of the continuum so that we can, in effect, obtain a thorough sampling of that individual's heart rate population. If we use tasks to evoke low, middle, and high heart rates equally, and if we use a sufficient number of data points, then our comparison sample mean (i.e., calibration period) should level out. Also, the mean value should be truly representative of a medial state of being because it involved sampling heart rates under a variety of conditions.

Currently, it appears that the aforementioned mixed method is being applied offline by some augmented cognition researchers using electroencephalography (EEG; Belyavin, 2005; Berka et al., 2005). Essentially, they have developed a procedure consisting of a variety of tasks and then use baseline data collected during specific tasks to train the EEG so that classification of incoming data can occur based on the training data. This calibration method appears to serve well for that purpose and allows a high degree of sensitivity and accuracy as long as the training data used for calibration tap into the full range of variability experienced in the actual experimental session. In fact, Belyavin (2005) argued in favor of this type of calibration. He approached his argument from a cognitive perspective rather than the physiological perspective

presented in this paper. He stated that calibration tasks should be representative of relatively "pure" cognitive states that were readily identified and associated with the states that the measurement is trying to differentiate. In this paper, we state that the physiology evoked from these cognitive states should be representative of the range of physiology expected during the states of interest. Belyavin's argument can be extended even further into the idea that the calibration period should contain tasks that represent all the pure cognitive states, behavioral states, workload states, or other states to which one would expect the measure to be sensitive.

It is important to note that these researchers employ one of two approaches to calibrating physiological data. Their approach is similar to that of a cluster analysis, in which classifiers are developed to determine whether incoming data can be classified as having properties of high-workload situations or low-workload situations. Although that approach differs from the "calculate and compare with baseline value" approach presented in this paper, the same concerns exist, as the difference in technique simply results in a variation of the same issues. From the perspective of this paper, if one manipulates the physiological measure directly in a way that adequately represents the population of physiological responses expected, one will adequately calibrate the measurement for the state one is trying to detect regardless of the analysis approach to data calibration that has been employed.

Some researchers, on the other hand, who are also interested in cognitive state measurement, appear to continue using biased calibration periods when attempting to detect both low and high cognitive states. Backs, Shelly, and Lenneman (2005), for example, used a resting baseline for comparison when examining the operation of the different autonomic control modes of the heart in response to a variety of information-processing tasks, including driving. Utilizing only a resting baseline for calibration purposes may have limited the magnitude of the resulting inhibitory changes of the autonomic nervous system that they found, resulting in lower accuracy in the detection of state changes. Interestingly, the authors acknowledged the baseline calibration period issue as a concern when examining the physiological correlates of information processing in real time.

In addition, consider the work of Prinzel, Freeman, Scerbo, Mikuka, and Pope (2003), who examined adaptive automation by a system based on levels of engagement. Using EEG technology, the researchers calculated an engagement index to reflect changes in arousal and attention and used the index to drive automated changes in experimental conditions in an effort to improve participant performance. At any given moment in the experiment, the participant's current engagement index was compared with a baseline engagement index mean. The baseline engagement index was determined by having participants sit quietly for 5 min with their eyes closed and another 5 min with their eyes open. The data obtained during those periods were averaged to create a baseline engagement index. Based on the behaviors performed by the participants during the baseline, it appears that the mean engagement index during that time would be quite low and any comparison with that value will be sensitive to positive changes in cognitive state more so than negative changes in cognitive state.

The authors stated that in determining adaptive automation, "an EEG index above baseline was taken to indicate that the participant's engagement level was high, whereas an EEG index below baseline was taken to indicate that engagement level was low" (p. 605). Essentially, the authors were saying that attention and arousal at any moment above the attention and arousal evoked by 5 min of eyes-open data averaged with 5 min of eyes-closed data equates to high engagement. This logic seems flawed in that the engagement of the participant during the baseline is already quite low in comparison with the task requirements. In other words, most of the time, a participant may appear to be "highly engaged" when, in fact, he is only engaged more than he was when compared with the very low baseline.

A similar phenomenon may have also been present in the work by Freeman, Mikulka, Scerbo, and Scott (2004), who used a 12-min practice session to establish their baseline engagement index. For their adaptive automation paradigm, shifts in the experimental conditions occurred when the engagement index at any moment was more or less than 0.2 *SD* from the baseline engagement index. Although the baseline engagement index values were not reported, it appears that the calibration period may have been biased toward detecting negative changes in the cognitive state attributable to the calibration period, which consisted entirely of an engaging practice session in which participants were trying to learn about the task.

Overall, it appears that the quality of the calibration period methodology used for detecting state changes in the positive and negative directions in augmented cognition systems varies. We use heart rate variability data, as measured by the Clemson University-developed Arousal Meter (Hoover & Muth, 2004), to illustrate the complexity and level of involvement necessary to address the problem of optimizing physiological state detection through the use of an appropriate calibration method for just one physiological measure. Given that closed-loop augmented cognition systems intend to function with a battery of physiological measures, the complexity of the problem can quickly multiply.

Using the Arousal Meter, a study was conducted to examine the issue of physiological state detection with respect to the calibration method used. Six calibration methods were used to examine arousal state detection in response to experimentally created low- and high-workload states. We compared five of the calibration methods with what we feel to be the gold standard calibration method based on the quality and length of the calibration period. The against-self method was considered the gold standard and involved using all the data collected during an experimental session to calculate a mean calibration value for that data series. All the data were then compared with that value sequentially. The large sample size and the wide range of variance in the sample give this method a reliability advantage (Jacob & Shapiro, 1994; Kamarck et al., 1993) and should have maximized the likelihood that the method would detect both negative and positive state changes equally.

The previously discussed and frequently used traditional resting baseline was used in the study and compared with and against the self-calibration method. In addition, we included two less commonly used calibration methods – a comprehensive baseline and a practice baseline. The comprehensive baseline included data from the

tasks used in the study as well as resting data. The practice baseline included only task practice data. Last, we included two running calibration methods, *running with baseline* and *running without baseline*, which, unlike the other calibration methods, allow for uninterrupted, real-time data collection, analyses, and interpretation. These running calibration methods involve the physiological measure constantly refining and updating the calibration value as new data enter the computational system, which in this case was the Arousal Meter. The running-with baseline calibration method used incoming data from the comprehensive baseline and the experimental task data to continually refine the calibration value. The running-without baseline calibration method used only task data to continually refine the calibration value. Specific details about each calibration method can be found in the Method section.

Based on the previously discussed literature and the statistical properties of the calibration periods used, we formulated hypotheses regarding the ability for each of the calibration methods to detect the low and high arousal states that were experimentally created by manipulating workload in comparison with the ability of our gold standard against-self calibration method. Specifically, we first hypothesized that the resting baseline method would be biased to detect high arousal states and would inflate those effects while having decreased sensitivity to detect low arousal states. This prediction was based on the resting baseline period being composed of low arousal states only and having little variance.

Second, we hypothesized the opposite effect for the practice baseline, in that it would be biased to detect low arousal states and would inflate those effects while having decreased sensitivity to high arousal states. This prediction was based on the practice baseline period being composed primarily of medium to high arousal states and having little variance.

Third, we hypothesized that the state detection capabilities of the comprehensive baseline for low and high arousal states would not differ significantly from the gold standard against-self because the comprehensive baseline period was most similar to the gold standard. Presumably, it was composed of a wide range of arousal values from practicing the low- and high-workload tasks as well as a resting baseline and therefore had more representative variability.

Fourth, we hypothesized that the state detection capabilities of the running-with baseline for low and high arousal states would not differ significantly from the gold standard because calibration began at the start of the comprehensive baseline period. The running-with baseline method may have equal sensitivity to state changes in the negative and positive directions provided that the individual's physiology is sufficiently manipulated in a controlled and comprehensive manner prior to the data collection and analysis period of interest. Based on some of our previous pilot work, we determined that approximately 20 min of data are needed to stabilize the mean and standard deviation of an individual's arousal measure when performing a given task or set of tasks. Thus, for the Arousal Meter, the running-with baseline calibration method should allow for adequate accuracy and sensitivity for detecting changes after 20 min of data have been collected prior to experimental manipulation. In this

case, more than 20 min of comprehensive baseline data were used, motivating the prediction that this calibration method would perform well.

Our fifth hypothesis was that the running-without baseline calibration method would be biased toward detecting low arousal states and would inflate those effects while having decreased sensitivity to high arousal states. This prediction was based on the fact that the running calibration always began at the start of the high-workload task in the experimental paradigm. Therefore, the incoming high-workload data at the beginning of the experiment were being compared with the high calibration values. The low-workload task data, collected later in the experiment, were also being compared with the high calibration values. Hence, the running-without baseline calibration method would be biased toward detecting low arousal states.

Summary of Hypotheses

1. The resting baseline method would be biased toward detecting high arousal states and would inflate those effects while having decreased sensitivity to detect low arousal states (positive directionality).
2. The practice baseline would be biased toward detecting low arousal states and would inflate those effects while having decreased sensitivity to high arousal states (negative directionality).
3. The state detection capabilities of the comprehensive baseline for low and high arousal states would not differ significantly from the gold standard against-self (nondirectional).
4. The state detection capabilities of the running-with baseline for low and high arousal states would not differ significantly from the gold standard against-self (nondirectional).
5. The running-without baseline calibration method would be biased toward detecting low arousal states and would inflate those effects while having decreased sensitivity to high arousal states (negative directionality).

Method

Participants

Fifty-six participants, 27 males and 29 females, with a mean age of 19.29 years (*SD* = 1.04 years) were recruited using Clemson University's online participant pool. All participants read and signed a consent form approved by the Clemson University Institutional Research Board prior to participating.

Participants were monetarily compensated for their time. They were given $4.00 for simply participating. Bonus compensation was commensurate with performance in the form of a monetary composite score on the primary and secondary tasks as a whole. Participants received half the amount earned on the primary and secondary tasks combined. This was attributable to participants performing better than expected in pilot testing. In order to preserve the ratio of money gain to money loss that was observed during pilot testing, the amount earned was halved.

Apparatus

Physiological recording equipment and software. Arousal was objectively measured by connecting participants to a fetrode-based EZ-IBI device (UFI, Morro Bay, CA). The EZ-IBI derived inter-beat intervals (IBIs) as they naturally occurred in time (i.e., asynchronously). IBIs are the distance in time (milliseconds) between the R-waves of the electrocardiogram (EKG) series. The asynchronous IBIs were recorded and analyzed using the desktop version 2.4b of the Arousal Meter (Hoover & Muth, 2004).

The Arousal Meter algorithm first passes the IBIs through a buffer, and errors in the IBI series are automatically detected and corrected. The asynchronous IBIs are then read into a 64-s data window as they are sampled 4 times every second (4 Hz; i.e., synchronously). Hence, the data window moves forward in time every 0.25 s. It is important to keep in mind, however, that on average, IBIs occur once per second (1 Hz). So, although the data window may theoretically contain new data from a computational point of view, from a physiological point of view, new IBIs must be produced by the human physiology before they can be detected by the EZ-IBI and used by the Arousal Meter. Until new IBIs occur, the previous IBIs are over-sampled at a rate of 4 Hz.

As the data window moves, the IBIs are plotted over time. When plotted over time, the IBIs form a wave-form. The IBI wave-form is broken into its component frequencies using spectral analysis. The high-frequency component between 0.15 and 0.5 Hz, also known as respiratory sinus arrhythmia (RSA), has been established as an index of parasympathetic nervous system activity, the branch of the nervous system responsible for relaxing the body (Eckberg, 1983; Grossman, Karemaker, & Wieling, 1991; Katona & Jih, 1975). RSA data are continuously derived from the IBI data. To standardize the physiological data, log RSA data are converted into standardized Z-scores that are updated at a rate of 4 Hz. Standardizing the log RSA data allows for comparisons both between and within participants. The standardized Z-score represents the individual's arousal at that given point in time. When RSA is low, arousal is high and vice versa. Therefore, standardized scores are negated to account for this inverse relationship, such that positive standardized arousal corresponds to increases in ANS arousal.

In order to obtain the aforementioned physiological data, ConMed Cleartrace electrodes (ConMed Corp., Utica, NY) were applied to participants' skin following skin preparation using Omni Prep abrasive cleansing paste (D.O. Weaver & Co., Aurora, CO). Participants were connected to the EZ-IBI module using three electrodes. Electrode placement was as follows: positive below the lowest left false or vertebrochondral rib, negative in the upper center region of the sternum (i.e., breastbone), and reference below the lowest right false or vertebrochondral rib.

Dual-task paradigm. The dual-task paradigm designed for the purposes of this study consisted of both a primary and a secondary task (see Figure 1). The primary task switched between a shooting task and a surveillance task. A modified mental arithmetic task based on the one developed by Turner et al. (1986) served as the secondary task.

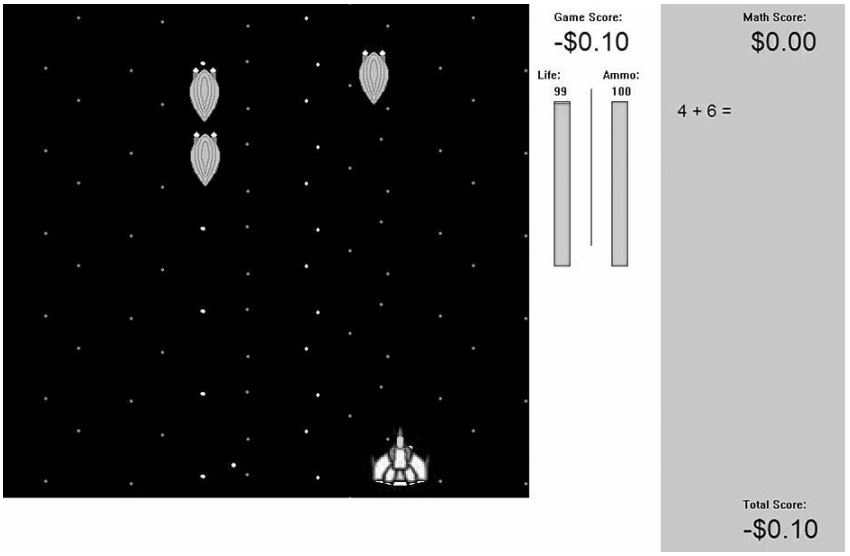

Figure 1. Screenshot of the high-workload scenario containing the shooting task and mental arithmetic.

The shooting task was designed to be a high-workload task. During the task, a spaceship was visible at the bottom of the screen. The spaceship could only move horizontally. Waves of enemy spaceships containing 3 to 15 ships appeared, as well as ammunition and life packets. The enemy ship pacing varied such that each wave of enemies contained fast-moving ships and slow-moving ships. The fast-moving ships took approximately 1–2 s to move across the screen. The slow-moving ships took approximately 6 s to move across the screen. The enemies also varied in approach such that some were heat seekers (i.e., followed the player ship) and some simply moved vertically down the screen, horizontally across the screen, or diagonally across the screen. The enemy fire rate was high throughout the task such that at least one enemy, if not more, was firing at any given time. The goals of the task were to shoot the enemy spaceships using the left mouse button, keep from getting hit with enemy fire, and sustain ammunition and life by collecting the falling ammunition and life packets.

Participants earned $0.02 on the task by hitting enemy spaceships and lost $0.02 when struck by enemy fire or an enemy ship. In addition, participants lost $5.00 when a life was lost. A life was lost if the participant ran out of life points. One life point was lost each time the participant's ship was hit with enemy fire and 10 points were lost when the participant's ship collided with an enemy ship. Once a life was lost, the life gauge automatically refilled to the original 100 points. Finally, participants lost money if they depleted their ammunition. Running out of ammunition did not result in the loss of life, but the participants lost $2.00 when the computer had to refill the gauge for them. Both life points and ammunition appeared on the screen routinely

throughout the game for the participants to collect. Upon collecting a life point or ammunition packet, the respective gauge was filled to 100 points. The shooting task monetary score was visible to the participants at all times and was located in the right-hand corner of the left-hand portion of the screen.

The surveillance portion of the primary task was designed to be a low-workload task and consisted of a warrior detection task. The participants were instructed to watch for a distinct red warrior ship with red engines among all the "regular" gray enemies. The warrior ship could appear randomly from any direction or angle on the screen. In fact, a regular enemy could transform into a warrior ship at any moment. Warrior ships appeared every 30 s to 60 s. Performance was measured as how long it took participants to detect the warrior ship (i.e., reaction time). When the ship appeared, two behind-the-scenes clocks started. The first was a monetary clock that counted down and continued to count down until the ship was noticed, as registered by a mouse button press. The clock started at $3.00. For every second that the participant took to recognize the warrior ship, the monetary clock decreased by $0.60. The resulting monetary amount was then added to or subtracted from the overall surveillance task monetary score. The second clock was a reaction-time clock that started at zero and continued to increase until the participant indicated that he or she saw the warrior ship by pressing the left mouse button.

The mental arithmetic task consisted of three levels of math problems. Level one consisted of 1-digit + 1-digit problems; level two, 2-digit + 1-digit problems; and level three, 2-digit + 2-digit problems. The participants were presented with one horizontal math problem at a time beginning with level one. A correct answer resulted in a jump of one level and an incorrect answer resulted in lowering of one level. Participants were given 5 s to answer level-one problems, 10 s for level-two problems, and 15 s for level three. The score for the mental arithmetic task was located in the far right-hand corner of the computer monitor right above the mental arithmetic task.

A green overall mental arithmetic monetary score indicated that the participant gave a correct answer, and a red overall mental arithmetic score indicated that the participant gave a wrong answer. To keep participants' motivation high, correct responses resulted in an increase in their mental arithmetic score, and monetary earnings and incorrect responses resulted in a decrease in their mental arithmetic monetary score. Failure to answer a problem in the allotted time resulted in a decrease in the monetary score. Level-one problems were worth $0.02, level two $0.04, and level three $0.06.

Both tasks performed by participants were locally developed and modified to meet the needs of the current study and were displayed on a PC with dual 800Mhz Pentium III processors, 1 GB of RAM, Windows 2000 Professional operating system, and a 15-inch LCD display. Participants' eyes were approximately 25 inches and no more than 34 inches from the monitor. They were instructed to focus most of their attention on the primary task without neglecting the secondary task. The primary task was located on the left-hand portion of the monitor and occupied 75% of the total screen space (see Figure 1). It was described to the participants as the most important and crucial task. If participants performed well on the primary task, they had the potential to earn significantly more money relative to the secondary task.

If they performed poorly, they could lose significantly more money relative to the secondary task.

When the primary tasks switched, the participant's representation in the game – a spaceship – changed in appearance from a solid red ship during the shooting task to a black-filled ship outlined in red during the surveillance task. In addition, the word "Shooting" or "Surveillance" temporarily appeared in the center of the screen in a large font the moment the tasks switched. With the exception of ship appearance, the two primary tasks were identical in appearance but differed in operation.

Experimental Design

This study examined the ability of six calibration types to differentiate the low-workload task (surveillance task) and the high-workload task (shooting task). There were two independent variables: the type of primary task (surveillance task or shooting task) and the calibration type (calibration against self, calibration against a comprehensive baseline, calibration against a resting baseline, calibration against a practice baseline, running calibration with a comprehensive baseline, and running calibration without a baseline). Hence, the design of the study was a 2 (task) × 6 (calibration type) repeated-measures design. There was one dependent variable – estimated arousal derived from heart rate variability data and calculated by the Arousal Meter.

Procedure

Upon entering the laboratory, participants were given an informed consent form to read and sign. In addition, any questions the participants had were answered. They were then connected to the EZ-IBI recording device via three electrodes. The Arousal Meter was started and physiological data collection began. Participants were then introduced to an extensive tutorial that described in detail how to complete each task. The compensation method was also explained in the tutorial. Participants were instructed to complete the tutorial at a pace that would lead to a complete understanding of the tasks. The mean time to complete the tutorial was 7.78 min (*SD* = 1.84 min). After completing the tutorial, participants were asked several questions (orally) by the experimenter to ensure full understanding of the dual-task paradigm and the compensation method.

In order to collect resting physiological data, participants were instructed to sit quietly with their eyes closed for 10 min without falling asleep. They were also instructed to keep movement to a minimum. After the resting data were collected, participants were allowed to practice one of the primary tasks with the secondary math task for 5 min. Practice sessions were counterbalanced across all participants so that approximately half the participants received the shooting and math task practice first and the other half received the surveillance and math task practice first.

After practicing the first task combination for 5 min, participants read the NASA-TLX instructions and completed it accordingly, keeping in mind the task scenario they had just practiced (i.e., either the shooting task or surveillance task, paired with mental arithmetic). They were allowed as much time as necessary. The participants then practiced the remaining task combination for 5 min and completed the NASA-TLX,

again keeping in mind the task scenario they had just practiced. Completion of the tutorial, resting baseline, and the NASA-TLX marked the end of the pre-experimental physiological recording period. The mean duration of this period was 41.24 min (*SD* = 3.80 min) across all participants.

The participants were then told that the experimental portion of the study was about to begin and they were allowed to make any necessary adjustments or changes to the mouse and keypad setup. They were also given the opportunity to ask any clarifying questions. Participants were told that the experimenter would not be present in the room throughout the experiment and that they were to continue performing the tasks until the computer froze, marking the end of the study. They were then reminded for a final time that they would be receiving half the monetary amount in the bottom right-hand corner upon completion of the study.

Participants were instructed to begin. The mental arithmetic task and the primary task were presented simultaneously. The computer automatically switched the primary tasks at random paired time intervals of 30 s, 1 min, 2 min, 4 min, and 8 min. Only the 8-min intervals were used in the data analysis for reasons that will be discussed later. So that the timing intervals for each task duration could be preserved, the shooting task always came before the surveillance task. This resulted in the utilization of a partial Latin square method whereby only the starting and ending conditions were different and task order was kept the same. This method resulted in five possible ordering conditions that were duplicated. Each ordering setup was used for approximately 11 participants. Participants were assigned to the condition order based on when they participated in the study. An entire experimental session lasted no more than 2 hr including the pre-experimental physiological recording period. Following the experimental session, participants completed the computerized version of the NASA-TLX for each primary task. They were then compensated, debriefed, offered a consent form, and were allowed to ask any questions.

Data Screening

Prior to data analyses, all data were screened to determine inclusion in the final analyses. Initially, inclusion criteria were based on the participant's performance on the dual-task paradigm and the nature of the participant's IBI data. A third criterion was later added based on suspect RSA data.

Five participants were excluded from the final data set based on their performance scores alone. Of the five, three were because of experimenter error. The other two participants were low statistical outliers with regard to their overall performance scores of -$24.90 and -$36.26 and were excluded because of being more than two standard deviations from the mean performance scores (*M* = $33.30; *SD* = $20.68).

Three participants were excluded from the final data set based on their IBI files. Because the error detection and correction portion of the Arousal Meter algorithm is relatively immature, the asynchronous and synchronous IBI data for all 56 participants run in the study were visually examined for errors using Microsoft Excel as well as locally created programs AViewer and IBI Edit. IBI files were triaged into one of the following categories: (a) IBIs usable as is without any corrections, (b) erroneous

IBIs present that are not manually correctable, or (c) erroneous IBIs present that are potentially manually correctable. Two participants fell into the second category and were not used because of having a large number of uncorrectable errors in their raw asynchronous IBI file. Of those participants who fell into the third category, only one had unusable data. Even after manual correction, the synchronous IBI data file for that participant contained several errors that were not present in the raw asynchronous IBI file. This indicated that the error detection and correction functions of the Arousal Meter were correcting errors that were not present.

Finally, three participants were excluded from the final data set based on their RSA power. Those participants' raw RSA power was so low that the mean log RSA was negative, indicating that perhaps we were not getting an accurate measure of RSA and arousal from those participants.

The final data set used in the data analyses consisted of 45 participants (21 males, 24 females) with a mean age of 19.22 years (*SD* = 1.00 years).

Data Reduction and Analysis

Six calibration periods were used to calculate the arousal data for the 8-min primary-task intervals. Only the two 8-min intervals were used because the arousal state change between an 8-min low-workload task should be easily detectable compared with an 8-min high-workload task, as our previous work has shown the sensitivity of the Arousal Meter to be 2 min (Fishel, Muth, Hoover, & Gugerty, 2006). Further, the intended applications of the Arousal Meter are for periods greater than this interval. Because the purpose of this study was to compare different calibration techniques at a gross level, it was felt that the state change should be well differentiated so that if it were not detected, it would be clear that the calibration technique was flawed.

First, a gold standard against-self calibration period was created by taking the entire data stream for each participant from start to finish, including the comprehensive baseline as well as the experimental data, and calibrating the entire data file against that. In other words, the data stream was calibrated against itself. On average, this calibration period was 72.24 min in length (*SD* = 3.80). The experimental portion of the against-self calibration period was composed of all the task time intervals in an effort to simulate a calibration period that is both well balanced and has a large sample size.

The second calibration type was a comprehensive baseline. The entire baseline period including the tutorial, practice, and resting data was used as the calibration period against which to compare the physiological data from the 8-min task intervals. This calibration period had an average length of 41.24 min (*SD* = 3.80).

The third calibration type was a resting baseline. The 10-min resting baseline period served as the comparison condition for the experimental data.

The fourth calibration type was a practice baseline. The pre-experimental task practice session and NASA-TLX completions served as the comparison condition for the experimental data. This calibration type had an average length of 20.47 min (*SD* = 2.07).

The fifth calibration type was a running-with baseline calibration period. The entire pre-experimental physiological recording period (i.e., comprehensive baseline) along with the experimental data was run through the Arousal Meter. Essentially, the calibration value changed and continued to be refined throughout the data processing, and the new incoming data were compared with the calibration value at that point in time as they were run through the Arousal Meter. Because the calibration period changes with each new data point, the duration of the calibration period changes as well; therefore, calculating a length for this calibration period carries no meaning.

The sixth and final calibration type was a running-without baseline calibration period. Two minutes of IBI data, occurring immediately before the 8-min primary tasks, as well as the data from the primary tasks, were run through the Arousal Meter (18 min total). The 2 min of IBI data were necessary to fill the data-processing buffer of the Arousal Meter so that arousal values could be computed for the 8-min task intervals in their entirety. As with the running-with baseline calibration method, for this method, the calibration value constantly changed throughout data processing.

Following calibration, the arousal data for each participant were averaged across each of the 8-min shooting and surveillance task intervals for each calibration type, for a total of 16 arousal values per participant. A 2 (task) × 6 (calibration type) repeated-measures analysis of variance (ANOVA) was performed as a manipulation check to determine if the high- and low-workload tasks resulted in significantly different arousal values. In addition, the ANOVA was used to determine if detection of the high- and low-workload tasks was affected by calibration type and to determine the presence or absence of an interaction between calibration type and task. For hypothesis testing, planned comparison tests were performed comparing the arousal values obtained for each of the primary tasks using the gold standard against-self calibration method against the values obtained for each of the remaining five calibration types. Depending on the specific hypothesis tested, and as previously noted in the Summary of Hypotheses section, some of the planned comparisons were directional and some were nondirectional.

Results

A 2 × 6 repeated-measures ANOVA with Greenhouse-Geisser corrections for sphericity revealed a significant main effect for task and calibration type, $F(1, 44) = 46.97$, $p = .00$, $\eta^2 = .52$ and $F(1, 58) = 3.90$, $p = .04$, $\eta^2 = .08$ respectively. In addition, a significant task × calibration type interaction was found, $F(2, 79) = 3.95$, $p = .03$, $\eta^2 = .08$.

Planned comparison t-tests, which compared the against-self calibration method with the other five calibration methods for each of the tasks, revealed some significant differences. Figure 2 provides a graphical representation of the mean arousal values by task and calibration type. As shown in Table 1, the planned comparison *t* tests revealed three significant differences between the gold standard against-self calibration method compared with the other calibration methods. For the shooting task,

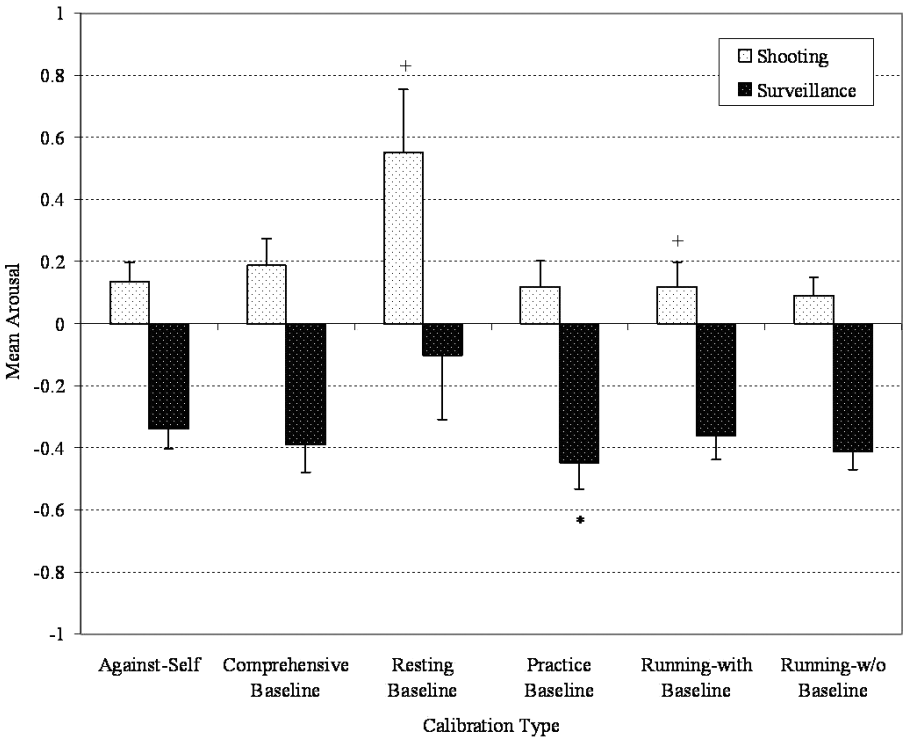

*Figure 2. Mean arousal values by calibration method and task. Note: + denotes statistical significance in the positive direction; * denotes statistical significance in the negative direction.*

the resting baseline calibration method resulted in a significantly higher mean arousal (M = .55; SD = 1.34) than the against-self method (M = .19; SD = .42). The running-with baseline calibration method, on the other hand, resulted in a significantly lower mean arousal (M = .09; SD = 0.54). For the surveillance task, the practice baseline calibration method resulted in a significantly lower mean arousal (M = -.45; SD = 0.57) than the against-self calibration method (M = –.34; SD = 0.44).

Discussion

The aim of this paper was to illustrate the effect that the chosen baseline calibration periods can have on the analysis of physiological data. Specifically, the calibration procedure used can lead to biased detection of some physiological state changes over others. Therefore, unique considerations and thorough forethought should be given to calibration periods used to compare data of interest in closed-loop systems, such as those discussed in the area of augmented cognition. Arousal data derived from heart rate variability were provided to illustrate the effects that different calibration

Table 1. Planned Comparison Results for Calibration Types Compared With Against-Self Calibration

Calibration Type Pairwise Comparisons	Shooting Task		Surveillance Task	
	t	*p*	*t*	*p*
Against-self – comprehensive baseline	–1.60	.12	–1.55	.13
Against-self – resting baseline	–2.13	.02	–1.42	.08
Against-self – practice baseline	1.58	.06	2.38	.01
Against-self – running-with baseline	2.34	.02	1.25	.22
Against-self – running-without baseline	1.03	.31	1.37	.18

methods can have on the sensitivity of a single physiological measure when detecting physiological state changes.

Six calibration methods were examined for effects on arousal values as determined by the Arousal Meter. The calibration methods possessed characteristics that caused them to vary on both the level of bias of the calibration period and the length (i.e., sample size) of the calibration period. The data support the idea that the output of a real-time physiological measure can vary depending on the calibration method utilized. In fact, 8% of the variance in the arousal values was accounted for simply by the calibration type used. Given that a methodological choice can account for that much variance in the measure of interest, great care should be taken to appropriately calibrate a given measure.

Our results indicated that all the calibration methods allowed for distinguishing the low-workload task from the high-workload task, but some of them distinguished the two tasks in a biased manner. Overall, three of the study predictions were supported and two were not. As predicted, the comprehensive baseline method was not significantly different from the gold standard against-self method on either of the primary tasks. The resting baseline method, on the other hand, resulted in significantly higher arousal values on the shooting task. In addition, the practice baseline method resulted in significantly lower arousal values on the surveillance task. It appears that the resting baseline method was biased toward detecting positive physiological changes and the practice baseline was biased toward detecting negative physiological state changes, as hypothesized.

One of the study predictions not supported by the data was that the running-with baseline calibration method would not differ significantly from the gold standard against-self method. The data, on the other hand, revealed that the running-with baseline method resulted in significantly lower arousal values on the shooting task. This unexpected finding will be discussed in more detail later.

The final prediction of the study was also not supported. The prediction was that the running-without baseline method would differ significantly from the gold standard calibration method. The results revealed that the running-without baseline method did not result in arousal values that were significantly different from the gold standard. Our suspected reason for this result will be discussed later.

Given that all three calibration methods that were successfully examined in the current study (although sometimes in a biased way) detected the differences in arousal associated with task workload changes, the advantages and disadvantages of each method from a statistical point of view as well as the potential sources of bias should be addressed. In examining each method, methodological and practical issues should also be considered.

As previously stated, the against-self method was utilized as the gold standard in this study against which all other calibration methods were compared. The calibration period that was utilized for this method was longer compared with the other methods and more or similarly balanced than the other methods, which gave it a sensitivity advantage when detecting physiological state changes. The against-self method involves using all the data collected during an experimental session to calculate mean and standard deviation calibration values for that data series. All the data can then be compared with that value sequentially.

In general, the advantage of using this method is that the sensitivity at detecting state changes is high because there is a large sample size and more variance in the sample. In addition, high accuracy in detecting the magnitude of state changes can also be achieved the closer the physiologically evoked balance of low and high states is to 100%. A disadvantage of this method, particularly from an augmented cognition standpoint, is that this method is essentially useless for real-time data collection and analysis. In order to use this method, all the data have to be collected offline and then analyzed offline by comparing the data with the calculated calibration value.

The traditional resting baseline method is commonly used, simple to implement, straightforward, and not very time intensive. It is useful for simple examinations of the presence or absence of change as well as the direction of that change relative to the baseline. Problems occur with this method when precise measures of change magnitude are needed. Specifically, physiological changes associated with states below wakeful rest will be difficult to detect because of a floor effect analogous to those observed in studies examining the LIV. A resting baseline may suffice for some measures in which only excitatory states are of interest (e.g., cardiovascular reactivity), as a resting baseline is essentially at the lowest range of that measure. However, misinterpretations of evoked state changes of interest that are at the low and middle ranges of inhibitory and excitatory states compared with the resting baseline may occur. For example, as in some of the research previously discussed, the inhibitory responses of the autonomic nervous system in response to changes in information processing (Backs et al., 2005), as well as the level of engagement by an individual in response to a variety of experimentally created arousal and attention changes (Prinzel et al., 2003), may be under- or overestimated.

The practice baseline, consisting of practice with or exposure to the state changes of interest, is similar to the resting baseline in that it is fairly easy to implement and straightforward and has low temporal requirements. The practice baseline method may lead to biased and amplified state detection of negative state changes. Biased detection of negative states may be present in current literature, such as in the work by Freeman et al. (2004), in which a 12-min practice baseline was used in establishing

a calibration value to compare incoming EEG data against for adaptive automation. This effect may be attributable to the learning and/or excitement associated with exposure to a task or stimulus of interest for the first time, leading to a mean calibration value that is on the high side, even though it may contain values from low and high task periods of interest.

As was evident in our findings that the comprehensive baseline was not significantly different from the gold standard against-self calibration method (by incorporating resting physiological values into the calibration period mean), the likelihood of biased-state detection greatly decreases. In a sense, using a resting period in addition to practice periods may add balance to the physiological values obtained during the practice periods, which may be higher during the true periods of interest because of learning and first-exposure effects.

The comprehensive baseline method used in the study resulted in values similar to the against-self method. It involves calculating a calibration value from a calibration period that consists of tasks that manipulate an individual's physiology in the range of variability expected during the experimental tasks of interest. Some augmented cognition researchers have successfully employed this method (Berka et al., 2005); Belyavin (2005) even advocated the use of this method. In addition to the accuracy achieved when using this method, there are also several methodological advantages. First, this method is useful when the length of the calibration period needs to be short in contrast to the running-with baseline calibration method, which requires a longer calibration period. The time requirements to implement this method can be equal to those of the resting and practice baseline methods. For most measures, a higher degree of sensitivity for detecting state changes can be achieved using this method.

The comprehensive baseline method can essentially be thought of as a condensed version of the running-with baseline method in that a planned, comprehensive calibration period can lead to having an accurate calibration value against which incoming data can be compared. This method of calibration is measure-independent in the sense that one would not be constrained by physiological measure specifications (e.g., sampling rate) that might affect the time necessary for calibration value stabilization. In other words, if the running-with baseline method requires 20 min of data for calibration value stabilization to occur, then one could condense that period by using shortened versions of tasks and still end up with a comprehensive and accurate calibration value.

In addition, the comprehensive baseline method is not limited to offline data collection and analysis. Real-time data collection and interpretation using this method can occur by having a given physiological measure run continuously and then simply locking the calibration values at the end of the calibration period. Data collection can then resume.

It is important to note, as stated in the introduction, that the establishment of a comprehensive baseline can be done in several ways. One can choose representative workload or cognitive, behavioral, physiological, or other states to manipulate during the baseline. After collection of the baseline data, a variety of analysis techniques can be employed. These techniques include, but are not limited to, descriptive

statistics, classification techniques, and training methods. The strengths and weaknesses of these various analysis methods are beyond the scope of this paper. However, the critical factor is that the baseline data used in those measures should be comprehensive no matter which analysis technique is employed.

In addition to the calibration periods that are already being used by researchers, we introduced two new "real-time" running calibration methods that are not currently utilized. The running-with baseline calibration method involves the physiological measure constantly refining and updating the calibration value as more data are collected, beginning with a baseline and ending with the conclusion of the experimental period. The baseline used in this study to begin the running calibration was comprehensive in nature. Our results surprisingly revealed that this method blunted the high-workload shooting task arousal values compared with the gold standard. This may have been attributable to the comprehensive baseline period not being 100% balanced, in the sense that the resting baseline comprised only 10 min of the period, whereas, on average, over 30 min of the baseline was composed of learning about the tasks and practicing them. Perhaps if a more balanced comprehensive baseline period was used to start the running calibration, the high-workload arousal values may not have been blunted.

The advantage of this method is that it allows uninterrupted real-time data collection, analyses, and interpretation. In addition, a high degree of sensitivity and accuracy can be achieved if the physiological measure is allowed to run for an adequate amount of time and if an individual's physiology is sufficiently manipulated in a controlled and comprehensive manner prior to the data collection and analysis period of interest. The necessary calibration period length may vary depending on the physiological measure used, and thus needs to be determined for a measure on an individual basis. If the calibration values used in the running-with baseline method have not had a chance to stabilize, then the incoming data are essentially being compared with a calibration value that is still changing too greatly. This may lead to decreased accuracy of the output values, which may then be misinterpreted as a state change, or a given state change may be misclassified. Hence, a disadvantage of this method is that it requires a longer period in comparison with the traditional resting and vanilla baseline methods in order for stabilization of the calibration value to occur via regression to the mean. Although this may not be an issue for some, it may add unwanted time to an experimental procedure, thereby lengthening the time required of the participant.

The final calibration method utilized in the study, the running-without baseline, resulted in a finding that was unexpected at first glance. This method used only task data to continually refine the calibration value and was not found to result in significantly different arousal values compared with the gold standard against-self method. This may be explained by the fact that we had to use 2 min of data prior to the 8-min tasks in order to fill the data-processing buffer so we could obtain actual arousal values starting at the beginning of the first 8-min period rather than zeros. The 2 min of data prior to the first 8-min task consisted of data from the low-workload surveillance task for most participants. Those data, in combination with the incoming

high-workload data from the first 8-min period, most likely resulted in a seemingly well balanced calibration mean with which the data were being compared. This effect most likely continued when the second 8-min low-workload interval began, resulting in the low-workload values when compared with a seemingly balanced, yet constantly updating, calibration value.

The balance of the calibration period may have been maintained because of fatigue effects at the end of the high-workload 8-min period, which may have dropped the arousal values slightly. Without the inclusion of the 2-min buffer period of low-workload data, it is reasonable to predict that the incoming high-workload data would have been significantly blunted because of being compared with high values. The incoming low-workload surveillance data also would have been inappropriately amplified because of being compared with high values. Although not statistically significant, the running-without baseline method did result in lower arousal values for the high-workload task and lower values for the low-workload task compared with to the against-self method.

Although this method appeared to perform well in this study, it is important to note that the calibration value was continually changing throughout the running calibration in a manner that was probably more dramatic than desired. Rather than fine tuning the mean calibration value, the mean was essentially still being established throughout the entire calibration time because the entire period lasted only 18 minutes. As stated before, the design of the study may have led to the mean calibration value being rather balanced.

On the surface, an advantage of this system may be that it requires no planning or forethought with regard to a calibration or baseline period. However, a large amount of accuracy may be sacrificed using this method because of the instability of the calibration values against which the incoming data are compared. Given the need for accuracy in augmented cognition systems, we recommend that this method not be used to evaluate incoming data simply because of the lack of an established mean calibration value against which to compare the data.

Conclusions

Overall, the issue of physiological measure calibration is a fundamental challenge associated with using physiological sensors in the closed-loop augmented cognition system. As previously discussed, a variety of factors must be considered when determining which appropriate calibration method to use. In essence, physiological measure calibration must be fully understood and correctly implemented for complete success and accuracy of augmented cognition systems, because calibration is a key factor in increasing a measure's sensitivity to state changes.

Future studies involving physiological measures in the augmented cognition realm and beyond will greatly benefit from choosing appropriate calibration procedures and specifying the calibration method used in documentation of those studies. Not fully understanding and accounting for this issue for all physiological measures that are involved in an augmented cognition "state detector" can lead to undesired

outcomes because of misclassification and interpretation of data, which in the end can result in an ineffective state detector. Given that closed-loop augmented cognition systems are intended to function with a battery of physiological measures, the challenge of calibration will no doubt be greater than the single physiological measure calibration presented in this paper.

Acknowledgments

We gratefully acknowledge the support of grant #N000140210347 from DARPA through the Office of Naval Research. This work was also supported by a subcontract from DARPA through Honeywell Corporation.

References

Backs, R.W., Shelley, J., & Lenneman, J. K. (2005). Using modes of cardiac autonomic control to assess demands upon processing resources during driving. In D. Schmorrow (Ed.), *Foundations of augmented cognition* (Vol. 11; pp. 101–109). Mahwah, NJ: Erlbaum.

Belyavin, A. (2005). Construction of appropriate gauges for the control of augmented cognition systems. In D. Schmorrow (Ed.), *Foundations of augmented cognition* (Vol. 11; pp. 430–437). Mahwah, NJ: Erlbaum.

Berka, C., Levendowski, D. J., Davis, G., Lumicao, M. N., Ramsey, C.K., Stanney, K., et al. (2005). EEG indices distinguish spatial and verbal working memory processing: Implications for real-time monitoring in a closed-loop tactical Tomahawk weapons simulation. In D. Schmorrow (Ed.), *Foundations of augmented cognition* (Vol. 11; pp. 405–413). Mahwah, NJ: Erlbaum.

Cathcart, S., & Pritchard, D. (1998). Relationships between arousal-related moods and episodic tension-type headaches: A biophysiological study. *Headache: The Journal of Head and Face Pain, 38*(3), 214–221.

Dickman, S. J. (2002). Dimensions of arousal: Wakefulness and vigor. *Human Factors, 44*, 429–442.

Eckberg, D. L. (1983) Human sinus arrhythmia as an index of vagal cardiac outflow. *Journal of Applied Physiology, 54*, 961–966.

Fishel, S. R., Muth, E. R., Hoover, A. W., & Gugerty, L. J. (2006). Determining the resolution of a real-time arousal gauge. In P. J. Gardner & A. W. Fountain III (Eds.), *Proceedings of the Society for Optical Engineering: Vol. 6218. Chemical and Biological Sensing VII.* Bellingham, WA: SPIE.

Freeman, F. G., Mikulka, P. J., Scerbo, M. W., & Scott, L. (2004). An evaluation of an adaptive automation system using a cognitive vigilance task. *Biological Psychology, 67*, 283–297.

Grossman, P., Karemaker, J., & Wieling, W. (1991). Prediction of tonic parasympathetic cardiac control using respiratory sinus arrhythmia: The need for respiratory control. *Psychophysiology*, *48*, 201–216.

Grossman, P., Stemmler, G., & Meinhardt, E. (1990). Paced respiratory sinus arrhythmia as an index of cardiac parasympathetic tone during behavioral tasks. *Psychophysiology*, *27,* 404–416.

Hastrup, J. L. (1986). Duration of initial heart rate assessment in *Psychophysiology*: Current practices and implications. *Psychophysiology*, *23*(1)*,* 15–18.

Hoover, A., & Muth, E. (2004). A real-time index of vagal activity. *International Journal of Human Computer Interaction*, *17,* 197–209.

Hord, D. J., Johnson, L. C., & Lubin, A. (1964). Differential effect of the Law of Initial Value (LIV) on autonomic variables. *Psychophysiology, 1*(1)*,* 79–87.

Jacob, R. G., & Shapiro, A. P. (1994). Is the effect of stress management on blood pressure just regression to the mean? *Homeostasis, 35*(3), 113–119.

Jennings, J. R., Kamarck, T., Stewart, C., Eddy, M., & Johnson, P. (1992). Alternate cardiovascular baseline assessment techniques: Vanilla or resting baseline. *Psychophysiology, 29*, 742–750.

Kamarck, T. W., Jennings, J. R., & Manuck, S. B. (1993). Psychometric applications in the assessment of cardiovascular reactivity. *Homeostasis, 34*(5-6), 229–243.

Katona, P. G., & Jih, F. (1975). Respiratory sinus arrhythmia: Noninvasive measure of parasympathetic cardiac control. *Journal of Applied Physiology*, *39*, 801–805.

Nielson, K. A., Yee, D., & Erickson, K. L. (2005). Memory enhancement by a semantically unrelated emotional arousal source induced after learning. *Neurobiology of Learning and Memory, 84*(1), 49–56.

Porges, S. (1992). Vagal tone: A physiologic marker of stress vulnerability. *Pediatrics, 90*, 498–504.

Prinzel, L. J., III, Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2003). Effects of a psychophysiological system for adaptive automation on performance, workload, and the event-related potential P300 component. *Human Factors, 45*, 601–613.

Robazza, C., Bortoli, L., & Nougier, V. (1998). Physiological arousal and performance in elite archers: A field study. *European Psychologist, 3*(4), 263–270.

Rosenthal, R., & Rosnow, R. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw Hill.

Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). *Psychophysiological recording.* New York: Oxford University Press.

Turner, J. R., Hewitt, J. K., Morgan, R. K., Sims, J., Carroll, D., & Kelly, K. A. (1986). Graded mental arithmetic as an active psychological challenge. *International Journal of Psychophysiology*, *3*, 307–309.

Wilder, J. (1957). The law of initial value in neurology and psychiatry. *Journal of Nervous and Mental Disease*, *125*, 73–86.

Stephanie R. Fishel is a doctoral candidate in the human factors psychology program at Clemson University. She earned her M.S. (2005) in applied psychology from Clemson University. Her current research focuses on physiological psychology, stress, and performance.

Eric R. Muth is a professor in the Psychology Department at Clemson University. He earned his Ph.D. (1997) in psychology from the Pennsylvania State University. His current research focuses on the effects of stress on human behavior, performance, and physiology.

Adam W. Hoover is an associate professor in the Electrical and Computer Engineering Department at Clemson University. He earned his Ph.D. (1996) in computer science and engineering from the University of South Florida. His current research focuses on tracking systems, embedded systems, and physiological monitoring.