Cardiac-Activity Measures for Assessing Airport Ramp-Tower Controller's Workload

Miwa Hayashi, NASA Ames Research Center
Victoria L. Dulchinos, San Jose State University Foundation

Heart rate (HR) and heart rate variability (HRV) potentially offer objective, continuous, and non-intrusive measures of human-operator's mental workload. Such measurement capability is attractive for workload assessment in complex laboratory simulations or safety-critical field testing. The present study compares mean HR and HRV data with self-reported subjective workload ratings collected during a high-fidelity human-in-the-loop simulation of airport ramp traffic control operations, which involve complex cognitive and coordination tasks. Mean HR was found to be weakly sensitive to the workload ratings, while HRV was not sensitive or even contradictory to the assumptions. Until more knowledge on stress response mechanisms of the autonomic nervous system is obtained, it is recommended that these cardiac-activity measures be used with other workload assessment tools, such as subjective measures.

INTRODUCTION

Background

New air traffic control operation concepts, novel cockpit technologies, and human-factors research hypotheses are sometimes first tested in human-in-the-loop (HITL) simulation in a safe and controlled laboratory environment. In these simulations, the measurement of an operator's mental workload is often desired. There is a wide variety of workload assessment methods (Lysaght, et al., 1989).

Subjective measures, which directly survey the participants' subjective psychological experience, are the ones most commonly used. These measures include self-reported subjective ratings, such as NASA Task Load Index (Hart & Staveland, 1988), and open-ended comments. Subjective measures are popular because they are simple to implement, noninvasive to the body, and straightforward to interpret (i.e., a higher score means higher workload reported).

On the other hand, subjective measures have shortcomings. The measures are subjective by definition and normally contain large individual biases as well as noise. Furthermore, the data are typically sparse, because subjective ratings are probed at relatively large intervals, such as every 5 minutes, or after each run. Such sparse probing may miss a critical moment, when a certain traffic event of interest occurred. Also, the sensitivity of the ratings may be reduced if a participant chose to use only the lowest or the highest part of the scale, e.g., 1 or 2 in a 7-point scale. Lastly, in case of real-time workload ratings, asking the participants to assess their workload while they are on task may distract them and, in turn, affect their workload level. This distraction could be a major concern when the test goes to the field. Normally, real-time self-reporting is not an option if the operation is safety-critical or high-workload.

Wierwille and Eggemeier (1993) recommend using multiple workload assessment methods to mitigate the issues of using one type of method, and to take advantage of different methods. The goal of the present study is to examine mean heart rate (HR) and heart rate variability (HRV) measures as potential tools to supplement the subjective measures in our future HITL simulations and field trials.

These cardiac-activity measures were chosen because of their simplicity. An electrocardiography (ECG) sensor can be attached to the participant's body in a relatively non-intrusive manner and provide a continuous stream of inter-beat interval measurements. The measurement does not distract the operator, and the resulting data are objective. Thus, at least in theory, some of the subjective measures' issues are addressed. However, the issue of large individual biases and noise still remain or are perhaps even more problematic, as is described later in the paper. Also, the interpretation of the cardiac-activity measures is far from straightforward.

Mean Heart Rate (HR)

Mean HR (the average number of beats per minute) is derived from the heart's beat-to-beat (or R-to-R) intervals called *RR intervals*. HR is considered to reflect an overall level of general arousal, physical work, task demands, and emotional response (Wierwille & Eggemeier, 1993). Hankins and Wilson (1998) used HR to assess shifts in the pilots' workload during a real flight. Their study showed HR was sensitive to general task demands but poorly fit for diagnosing what type of work was causing the high workload. Roscoe (1987) points out that HR works better with the pilot-flying performing relatively demanding manual-flight task than the pilot-not-flying who is undertaking a purely monitoring task.

Heart Rate Variability (HRV)

HRV, also called *sinus arrhythmia*, is a measure of variability in the RR intervals. HRV is thought to reflect the balance between the sympathetic and parasympathetic activities of the autonomic nervous system (Task Force, 1996). In the frequency-domain, HRV's high-frequency power (HF; 0.15–0.4 Hz) is considered a marker of modulation of vagal tone (parasympathetic activity), whereas low-frequency power (LF; 0.04–0.15 Hz) is associated with both sympathetic and parasympathetic branches. Moreover, it has been reported that the mid-frequency power (MF; 0.08–0.15 Hz)—called *0.10-Hz component*—is suppressed during increased cognitive effort (Aasman, Mulder, & Mulder, 1987).

The MF have been successfully used to identify changes in operators' mental workload. For example, Vicente,

Thornton, and Moray (1987) found strong correlations between MF and the subjective ratings of effort recorded in a low-fidelity hovercraft course-tracking simulation. Rowe, Sibert, and Irwin (1998) had the participants play an air traffic control game, and reported that those with previous air-traffic-control experience exhibited reduction in MF as the number of free flyers increased. Tattersall and Hockey (1995) observed in a military long-haul flight simulation experiment that the flight-engineer trainees' MF showed suppression when they were working on problem-solving task rather than routine tasks, such as takeoff or landing.

Using the MF as a measure of cognitive workload has also met with skepticism. Nickel & Nachreiner (2003) demonstrated that the mental strain level inferred from the MF suppression when the participants were performing various types of work from the AGARD-STRESS battery were inconsistent with their perceived difficulty indices and task performance. Berntson and Cacioppo (2004) pointed out that the sympathetic and parasympathetic activations can be not only reciprocal, but also independent or even coactive. That is, individuals may respond differently to psychological stressors, e.g., one may increase sympathetic activation, whereas another may primarily withdraw parasympathetic activation. Kramer (1990) pointed out that speech and respiration increase blood pressure, thereby affecting the 0.1-Hz component.

Airport Ramp-Tower Simulation

The RR intervals were recorded as a part of (or "piggy backed" on) an airport ramp-tower HITL simulator evaluation conducted at NASA Ames Research Center in 2014. The main purpose of the simulation was to evaluate the Spot and Runway Departure Advisor (SARDA), a decision-support tool for ramp-tower controllers.

The ramp-tower controllers (*ramp controllers*) are responsible for overseeing aircraft traffic in the airport ramp area. They ensure aircraft-traffic separation and efficient taxi movements by giving proper and timely instructions to pilots via radio communication. Maintaining safe and efficient traffic flow requires the ramp controllers to engage in a multitude of high-level cognitive functions, such as monitoring, planning, calculating, problem-solving, multi-tasking, etc. Furthermore, the airport ramp area was divided into four sectors, and their duties included communicating and coordinating with the other sector controllers.

The simulation lasted three weeks and consisted of sixteen runs per week. Each run lasted for either 65 or 70 minutes depending on the traffic scenario. In total, six current ramp controllers from Charlotte Douglas International Airport (CLT) participated in the study. The four sectors were labeled North, East, South, and West. Each week, a new pair of CLT ramp controllers handled traffic in the East and South sectors, the most demanding sectors of the four. Traffic in the North sector, West sector, and the movement areas (i.e., the taxiways and the runways) were handled by the research team's confederate controllers.

The simulation demonstrated that the SARDA's departure metering advisory reduced, on average, one minute of departure taxi-time per flight, which resulted in 10-12% overall fuel saving (Hayashi, et al., 2015).

The present study compared the CLT ramp controllers' real-time workload ratings with their mean HR and HRV. The real-time workload ratings were recorded at every five minutes, starting at 10 minutes and ending at 65 minutes into the scenario. The rating scale was one to seven; one represented the lowest workload and seven the highest. When a beep sounded, the controllers indicated their score via a quick hand sign, and the researchers recorded them.

METHODS

Participants

All six CLT ramp-controller participants were male. All of them had been working in the CLT ramp tower for four to 25 years (mean = 9.4 years, standard deviation = 7.9 years). All of them signed a consent form for participation in the study.

RR Interval Recording

ECG data were recorded using a Firstbeat Bodyguard 2 (BG2) device. The BG2 was attached to the participant's chest via two electrodes. Each participant attached a BG2 in the morning and removed it after the last run of the day. The BG2's ECG sampling rate is 1000 Hz. Its internal RR-interval extraction algorithms automatically calculates the RR intervals from the ECG data. Parak and Korhonen (2013) demonstrated that BG2 was able to detect 99.95% of heartbeats.

In this study, the ramp controllers were not given any behavioral constraint. For instance, they were free to sit, stand up, and walk around as they wanted. Caffeinated drinks and smoking were allowed during break time. These conditions may have affected the quality of the cardiac data, but it ensured that the SARDA evaluation would not be affected by prohibition of those activities. Also, there was interest in testing the robustness of the HR and HRV assessment method under these types of conditions, since in the future field testing environment, to maximize safety of the operation, the participants' behaviors will likely be not constrained.

Computation of HR and HRV

The computation of these quantities was carried out in the following four steps.

1. *Artifacts* (or *ectopic beats*) in the RR intervals were detected using Saalasti's method (2004), which uses two criteria: a) hard limits (any intervals outside the minimum and maximum hard limits are marked as artifacts) and b) gradient (if the two successive intervals differ by more than the gradient threshold, the latter of the two intervals is marked as an artifact). The artifacts were simply skipped, without any value correction or interpolation performed to compensate for the skipped intervals.
2. The mean HR was calculated within each of the two-minute windows ending at the times when the real-time

workload ratings were recorded (e.g., 10, 15,…, 65 minutes), using only the non-artifact intervals.

3. MF and HF were computed within the same two-minute windows using only the non-artifact intervals. Lomb-Scargle Periodogram (LSP) algorithm was applied to estimate the power spectral density (PSD) (Scargle, 1982). Selection of the LSP algorithm is important, because Clifford and Tarassenko (2005) demonstrated that LSP tolerates up to 20% of missing data (i.e., the artifacts skipped in the Step 1) in terms of PSD estimation accuracy.
4. Lastly, MF and HF were normalized with the total power (0.04-0.15 Hz) to minimize the effects of moment-to-moment fluctuations of the total power during each run, and to emphasize the activity balance between the sympathetic and parasympathetic branches (Task Force, 1996).

Statistical Tests

Linear Mixed Model regression (LMM) (West, Welch, & Gałecki, 2014) was used to analyze correlation between the cardiac data and the real-time workload ratings. LMM was chosen because it is a repeated-measures analysis that can accommodate large between-subject biases, and it also can handle an unbalanced dataset (i.e., each of the seven workload rating scores contained different number of data points).

A linear model of mean HR or HRV was constructed with two main effects, Workload (WL) and Participant effects, and one two-way interaction of these effects (WL × Participant). WL effect was treated as a fixed, continuous effect, whereas Participant effect was treated as a random, categorical effect. A likelihood-ratio test was performed to examine whether the WL × Participant interaction term was significant. Next, under the most parsimonious model, the statistical-significance level for WL effect was calculated. *R* software (v. 3.1.2) and its packages, *lme4* (v.1.1-7) and *lmerTest* (v. 2.0-20), were used for analysis.

RESULTS

Mean HR

For the HR analysis only, the data point at 10 minutes into the scenario was excluded from the analysis in all the runs because of the slightly elevated heart rate trends observed in some participants' data. These trends were caused by the participants' climbing the staircase to the second-floor room at the beginning of each run, and large enough to affect the analysis of mean HR.

The LMM did not find statistical significance in WL effect. Figure 1 plots the means and standard errors of mean HR by workload rating. The plot shows that the HR did go down when the rating moved from 1 to 2 but that the ratings of 3 and 4 tended to result in higher mean HR than the rating of 1 or 2. (The rating of 5 comprised of only one data point. The ratings of 6 and 7 were never reported.) Overall, the correlation between the workload rating and the mean HR may be there, but is not strong enough to be statistically significant.

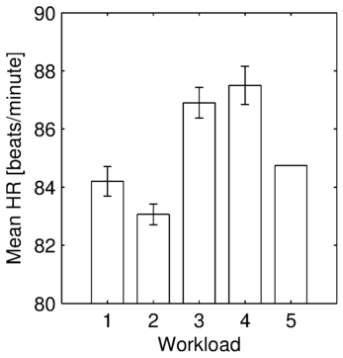

Figure 1. Means and standard errors of mean HR by real-time workload rating.

HRV

Of the 1,152 2-minute windows, seven resulted in an artifact ratio greater than 20%. Following the Clifford and Tarassenko's guidelines (2005), these seven windows were excluded from the HRV analysis. Unlike the mean-HR analysis, the HRV analysis was not affected by the slightly elevated values of the HRs at the beginning of each run. Thus, the data at all time-points were included, unless they fell in one of the aforementioned seven excluded windows.

The HRV results were either insensitive to the subjective workload ratings or sensitive but in an unanticipated direction. For the normalized MF, the LMM analysis revealed that WL effect was statistically significant ($p < 0.01$). However, its estimated coefficient suggested that the MF increased by 0.015 ± 0.006 (standard errors) per each workload score, rather than being suppressed as anticipated. Fig. 2 shows the means and standard errors of MF. The graph shows that MF was indeed higher when the participants reported a workload rating of 3 or 4 than 1 or 2, confirming the LMM results.

WL effect in the normalized HF was not significant. Figure 3 shows that HF was not sensitive to the real-time workload ratings. Figure 4 plots the means and standard errors of the total power in absolute values (i.e., no normalization), including LF, MF, and HF bands. It shows the total power increased when the participants reported scores of 3 or 4. This contradicts the general assumption that the total power is suppressed when sympathetic activity increases.

DISCUSSION

Mean HR showed only weak correlation with the participants' self-reported real-time workload ratings. This observation was based on only visual inspection of the means, and no corroboration from formal statistical testing was obtained. The direction of the trend observed in Fig. 1 (i.e., the scores 3 or 4 resulted in higher mean HR than the scores 1 or 2) was in agreement with the general assumption that mean HR increases with overall level of general arousal and task demands. The weakness of the association may not be solely due to low sensitivity of mean HR, but possibly also due to large noise in the self-reported workload ratings. The self-reported workload ratings are not the true state of the

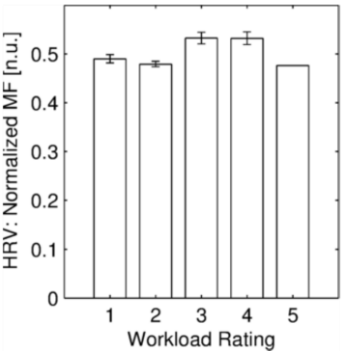

Figure 2. Means and standard errors of HRV MF by real-time workload rating.

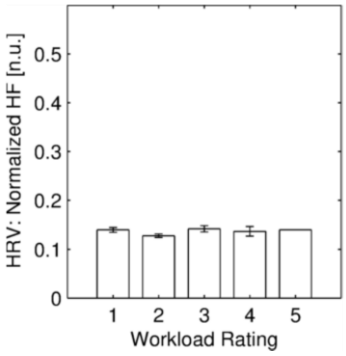

Figure 3. Means and standard errors of HRV HF by real-time workload rating.

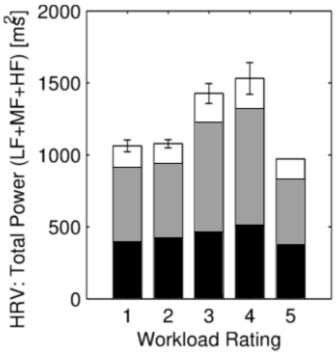

Figure 4. Means and standard errors of HRV total power (absolute values) by real-time workload rating. White is HF, gray is MF, and black is LF.

workload level, but, rather, noisy measurements of that. If both heart rate and self-reported ratings contain large noise, identification of correlations between them could be difficult.

Another way to assess the validity of mean HR is to apply the same statistical model as the one used on the real-time workload ratings for the main study. In the main part of the study for the SARDA evaluation, a six-way LMM that included six main effects and five two-way interaction effects was applied to the real-time workload ratings. Table 1 lists those effects, and the "x" marks indicate which effects were found statistically significant in mean HR and the real-time workload ratings. The two columns show different sets of effects found in each analysis. Again, the perfect agreement between the two is not a necessary condition for the validity. Interestingly, most of the inference results in the mean HR analysis were actually consistent with what was observed in the simulation, but not found statistically significant in the real-time workload ratings. For instance, the mean HR analysis found Sector effect to be significant (the third item in the table). This was consistent with the observation that the East sector was generally regarded as the most challenging sector by the controllers. Yet, the Sector effect was not found to be significant in the real-time workload ratings results. Researchers should be cautioned against type-I errors (false positives). That being said, in general, having more effects detected as statistically significant often helps the explanation of performance results, connecting findings, and strengthening the research conclusions.

Table 1. LMM statistical inference results

Effects	Mean HR	Real-time workload
Advisory (Advisory vs. Baseline)	x	
Scenario (1 vs. 2)	x	
Sector (East vs. South)	x	
Phase (4 chronological phases in each run to capture effects of traffic volume shift)		x
Run Block (1st-4th runs, 5th-8th runs, 9th-12th runs, and 13th-16th runs in each week to account for any learning or fatigue effect)	x	x
Participant (1-6)	x	x
Advisory × Scenario		x
Advisory × Sector		x
Advisory × Phase		
Advisory × Run Block	x	
Advisory × Participant	x	

The HRV data did not produce conclusive results in this study. MF and the total power increased when higher workload ratings were reported (scores 3 or 4). Unless the participants actually felt relaxed when the task demand was higher, these HRV values must have sensed some phenomenon other than the cognitive workload. One possible explanation for this inconsistency is that, when traffic volume increased, the controllers had to speak on the radio much more frequently. This increase in speech may have raised MF, but then, HF should also see decrease. In our data, HF increased in the absolute values (Fig. 4, white parts). Thus, this explanation is not completely satisfactory.

As researchers have pointed out, the sympathovagal balance of the autonomic nervous system involves complex mechanisms, and stress reactions and heart activity are a part of them. Therefore, it may be unreasonable to expect consistent behaviors between them (Berntson & Cacioppo, 2004). Besides the act of speaking, certain body movements, caffeine intake, room conditions, etc., may have also affected HRV. For assessing workload in the air traffic control tasks in a field-like setup, where behavior and conditions are less controlled, HRV may not be a suitable tool.

The analysis found that HRV was less susceptible to the effects of high physical activity (stair climbing) than mean HR. In an application where intense physical activities take place, HRV may offer a potential alternative to mean HR.

In the present study, mean HR and HRV demonstrated different levels of sensitivity to the operator workload level. This discrepancy was observed in past research. Cases were reported where mean HR was sensitive, but HRV was not (Hankins & Wilson, 1998), vice versa (Harris, Bonadies, & Comstock, 1989), or neither mean HR nor HRV was sensitive (Casali & Wierwille, 1983).

HR and HRV are considered to represent different parts of the cardiovascular systems (Aasman, Mulder, & Mulder, 1987). It is difficult to predict which measure will work better for a given task set in a given situation. Until further research is conducted, researchers are advised to include both mean HR and HRV, along with other types of workload assessment measures, such as subjective ratings. Both mean HR and HRV

are calculated from the RR intervals post experiment; thus, there is no reason not to include both of them.

CONCLUSIONS

The study found mean HR weakly sensitive to the airport ramp-tower controllers' self-reported real-time workload rating in simulated operations. HRV measures were insensitive or even contradictory, and their utility in complex, field-like settings, such as high-fidelity laboratory simulation or field testing in an actual air traffic control facility, is questionable. The discrepancy of sensitivity levels between mean HR and HRV have been observed commonly in past research. It is difficult to predict which type of measures will work, or even whether either of them will work; thus, it is recommended to use HR and HRV measures along with other types of workload assessment measures, such as subjective measures.

ACKNOWLEDGMENT

The authors would like to thank Dr. Rollin McCraty and Jackie Waterman of HeartMath Institute for their guidance in heart-rate measurement.

REFERENCES

Aasman, J., Mulder, G., & Mulder, L. J. (1987). Operator effort and the measurement of heart-rate variability. *Human Factors, 29*(2), 161-170.

Berntson, G. G., & Cacioppo, J. T. (2004). Heart rate variability: stress and psychiatric conditions. In M. Malik, & A. J. Camm (Eds.), *Dynamic Electrocardiography* (pp. 57-64). Elmsford, NY: Wiley-Blackwell.

Casali, J. G., & Wierwille, W. W. (1983). A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communication load. *Human Factors, 25*(6), 623-641.

Clifford, G. D., & Tarassenko, L. (2005). Quantifying errors in spectral estimates of HRV due to beat replacement and resampling. *IEEE Transactions on Biomedical Engineering, 52*(4), 630-638.

Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation, Space, and Environmental Medicine, 69*(4), 360-367.

Harris, R. L., Bonadies, G. A., & Comstock, R. J. (1989). Usefulness of heart measures in flight simulation. *Third Annual Workshop on Space Operations, Automation, and Robotics* (pp. 353-359). Houston, TX: NASA Johnson Space Center.

Hart, S. G., & Staveland, L. E. (1988). Development of a NASA-TLX (task load index): results of empirical and theoretical research. In P. S. Hancock, & N. Meshkati, *Human Mental Workload* (pp. 139-183). Amsterdam: Elsevier Science Publishers B. V.

Hayashi, M., Hoang, T., Jung, Y. C., Malik, W., Lee, H., & Dulchinos, V. L. (2015). Evaluation of pushback decision-support tool concept for Charlotte Douglas International Airport ramp operaetions. *11th USA/Europe Air Traffic Management Research and Development Seminar.* Lisbon, Portugal.

Kramer, A. F. (1990). *Physiological metrics of mental workload: a review of recent progress.* San Diego, CA: Navy Personnel research and Development Center.

Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., Linton, P. M., Wierwille, W. W., . . . Wherry, R. J. (1989). *Operator workload: comprehensive review and evaluation of operator workload methodologies.* Willow Grove, PA: US. Army Research Institute for the Behavioral and Social Sciences.

Nickel, P., & Nachreiner, F. (2003). Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. *Human Factors, 45*(4), 575-590.

Parak, J., & Korhonen, I. (2013). *Accuracy of Firstbeat Bodyguard 2 beat-to-beat heart rate monitor.* White paper by Firstbeat Technologies Ltd.

Roscoe, A. H. (1987). In-flight assessment of workload using pilot ratings and heart rate. In A. H. Roscoe (Ed.), *The Practical Assessment of Pilot Workload* (pp. 78-82). Essex, UK: Specialised Printing Services Ltd.

Rowe, D. W., Sibert, J., & Irwin, D. (1998). Heart rate variability: indicator of user state as an aid to human-computer interaction. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 480-487). Los Angeles: ACM Press/Addison-Wesley Publishing Co.

Saalasti, S., Seppänen, M., & Kuusela, A. (2004). Artefact correction for heart beat interval daeta. *Advanced Methods for Processing Bioelectrical Signals.* Jyväskylä, Finland.

Scargle, J. D. (1982). Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal, 263*, 835-853.

Task Force. (1996). Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *European Heart Journal, 17*, 354-381.

Tattersall, A. J., & Hockey, R. J. (1995). Level of operator control and changes in heart rate variability during simulated flight maintenance. *Human Factors, 37*(4), 682-698.

Vicente, K. J., Thornton, D. C., & Moray, N. (1987). Spectral analysis of sinus arrhythmia: a measure of mental effort. *Human Factors, 29*(2), 171-182.

West, B. T., Welch, K. B., & Gałecki, A. T. (2014). *Linear Mixed Models: a Practical Guide Using Statistical Software* (2nd ed.). Boca Raton, FL: CRC Press.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors, 35*(2), 263-281.