THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, *17*(4), 371–390

Workload State Classification With Automation During Simulated Air Traffic Control

David B. Kaber and Carlene M. Perry
Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC

Noa Segall
Human Simulation and Patient Safety Center, Duke University Medical Center, Durham, NC

Mohamed A. Sheik-Nainar
Research and Development, Synaptics, Inc., Santa Clara, CA

Real-time operator workload assessment and state classification may be useful for decisions about when and how to dynamically apply automation to information processing functions in aviation systems. This research examined multiple cognitive workload measures, including secondary task performance and physiological (cardiac) measures, as inputs to a neural network for operator functional state classification during a simulated air traffic control (ATC) task. Twenty-five participants performed a low-fidelity simulation under manual control or 1 of 4 different forms of automation. Traffic volume was either low (3 aircraft) or high (7 aircraft). Participants also performed a secondary (gauge) monitoring task. Results demonstrated significant effects of traffic volume (workload) on aircraft clearances ($p < .01$) and trajectory conflicts ($p < .01$), secondary task performance ($p < .01$), and subjective ratings of task workload ($p < .01$). The form of ATC automation affected the number of aircraft collisions ($p < .05$), secondary task performance ($p < .01$), and heart rate (HR; $p < .01$). However, heart rate and heart rate variability measures were not sensitive to the traffic manipulation. Neural network models of controller workload (defined in terms of traffic volume) were developed using the secondary task performance and simple heart rate measure as inputs. The best workload classification accuracy using a genetic algorithm (across all forms of ATC automation) was 64%,

Correspondence should be sent to David B. Kaber, Edward P. Fitts Department of Industrial & Systems Engineering, North Carolina State University, Raleigh, NC 27695–7906. E-mail: dbkaber@ncsu.edu

comparable to prior work. Additional neural network models of workload for each mode of ATC automation revealed substantial variability in predictive accuracy, based on the characteristics of the automation. Secondary task performance was a highly sensitive indicator of ATC workload, whereas the heart rate measure appeared to operate as a more global indicator of workload. A limited range of cardiac response might be sufficient for the demands of the brain in ATC. The results have applicability to design of future adaptive systems integrating neural-network-based workload state classifiers for multiple forms of automation.

In cognitively complex tasks, such as air traffic control (ATC), changes in workload can have significant impacts on operator performance. Increases in future air traffic volume and, consequently, individual air traffic controller workload, have been projected in the human factors literature (e.g., Parasuraman, Sheridan, & Wickens, 2000). As a result, various forms of advanced automation have been designed for ATC to reduce controller load and support improved performance. Unfortunately, high-level, static automation (e.g., fully autonomous systems) has historically been found to negatively affect performance as a result of operators being removed from system control loops (Endsley & Kaber, 1999) and experiencing complacency, vigilance decrements (Parasuraman, Molloy, & Singh, 1993), and loss of situation awareness (SA; Endsley & Kaber, 1999). Consequently, automation research has identified a need to monitor operator functional states in real time as a basis for determining the type and level of automated (computer) assistance that may be most appropriate for operators to complete tasks (e.g., Wilson, Monett, & Russell, 1997). Specific continuous measures of operator functional state, such as physiological and task performance variables, may have benefits for determining when flexible forms of automation, or adaptive automation (AA), should be applied to particular information processing (IP) functions of ATC to facilitate workload management and maintain operator familiarity with the current system control situation.

Historical studies of AA have demonstrated the utility of dynamic system function allocations for managing operator workload or maintaining levels of operator involvement in control loops to promote SA (Bennett, Cress, Hettinger, Stautberg, & Haas, 2001; Hilburn, Jorna, Byrne, & Parasuraman, 1997; Kaber & Riley, 1999). These studies have compared high-level static automation or completely manual control conditions with AA in laboratory simulations of real-world tasks. Approaches to AA used in these studies (i.e., "what" and "when" to automate) were based on predetermined control allocation schedules, secondary task performance measures, and electroencephalographic (EEG) indexes of workload. Other research has investigated the use of cardiocirculatory measures (e.g., heart rate [HR] and heart rate variability [HRV]) for assessing operator workload in real time and to serve as a basis for dynamic task allocation to automation or manual control (Scerbo et al., 2001; Wilson, 2001; Wilson, Lambert, & Russell, 2000). These studies have revealed that cardiac measures may also be useful for both purposes. Using peripheral physiologic variables for real-time workload measurement is also more practical than direct measures of arousal, including EEG, because of the complexity and lack of portability of measurement systems. There remains a research need for identification of effective and practical means of real-time operator workload assessment for AA systems. Related studies have provided evidence that a combination of physiological variables (e.g., EEG, electro-oculographic [EOG], and electrocardiographic [ECG]) may improve the accuracy of workload classification over the use of one type of variable (Wilson, 2001; Wilson et al., 1997). However, such approaches are very complex from a measurement perspective and may be impractical for actual applications in aircraft cockpits or ATC workstations. To our knowledge, there have been no real applications of EEG-based AA in aviation systems. More practical approaches may involve using simple physiological measures in combination with, for example, secondary task performance measures, for describing operator functional states.

One way to collectively consider a number of real-time measures of operator states in decisions about adaptive aiding is to use an artificial neural network (NN) for establishing nonlinear relations among multiple state variables with actual task workload conditions. Previous AA research has developed NNs for operator workload classifications (e.g., Prinzel, Freeman, Scerbo, Mikulka, & Pope, 2003; Wilson & Russell, 2003a, 2003b). The manner in which AA has been implemented in these studies is that specific tasks in a multitask scenario are either turned on or off. They are assigned to the human operator or automation depending on the operator's current workload condition. For example, Wilson et al. (2000) successfully used a NN to integrate observations on EEG signals, blink rate, and heart period to classify (with 85% accuracy) the task workload levels to which operators were exposed in the Multi-Attribute Task Battery (MAT–B; Comstock & Arnegard, 1992). This research also demonstrated that NN-based AA systems might be effective for managing operator workload.

Unfortunately, there has only been limited research examining the potential benefits of flexibly applying automation to specific IP functions in complex control tasks (Kaber, Perry, Segall, McClernon, & Prinzel, 2006; Kaber, Wright, Prinzel, & Clamann, 2006). This work has also been limited to investigation of AA driven by operator workload assessments using secondary task performance measures. There remains a need to establish the effectiveness of using specific physiological indicators of cognitive states and other secondary task measures of workload for triggering dynamic allocations of specific forms of IP assistance in adaptive systems, including information acquisition, information analysis, decision making, and action implementation, in real time. Prior research has demonstrated differential effects of these forms of automation on operator performance and workload (Kaber, Wright, et al., 2006). Information acquisition and action implementation automation target psychomotor tasks and appear to have a more direct impact on operator workload responses, and lead to higher SA and levels of performance, than information analysis and decision-making automation. Based on the characteristics of the adaptive aiding (e.g., summarizing information for operators, recommending decision alternatives, implementing system control actions), certain measures of operator workload may be more or less sensitive to specific automation manipulations. For example, operator workload fluctuations, as a result of decision-making automation, may best be revealed by HRV measures, which have been demonstrated to be sensitive indicators of high-level cognitive task demands. Aasman, Mulder, and Mulder (1987) and Byrne and Parasuraman (1996) found that HRV measures were sensitive to high levels of effortful processing, such as those required during information acquisition, with suppression of HRV under high workload. Consequently, the predictive accuracy of any workload classifier tool, like a NN, may vary depending on the characteristics of the AA system to which it is applied.

There is a need for further investigation of the use of HR and HRV measures of cognitive load in approaches to AA in laboratory simulations and in real-world systems. Furthermore, no studies have looked at combinations of HR and HRV with secondary task performance measures, as inputs to a NN for workload state classification, and to use the output of the model as a basis for triggering dynamic control allocations in an adaptive system. It is desirable for real-world applications of adaptive systems to identify a small, but powerful, set of variables, like HR and secondary task performance (Kaber & Riley, 1999), that can be easily captured for input into a NN classification tool during pilot or controller performance to accelerate prediction of workload states. This may also promote system responsiveness in applying AA. Of course, the accuracy of workload state prediction is critical as well, and novel combinations of simple physiological and secondary task performance measures may prove useful relative to complex EEG-based approaches. From an applications perspective, measurement implementation and accuracy issues may need to be considered tantamount.

With these research needs in mind, the objectives of the present project were to (a) develop an artificial NN for classifying operator functional states in an ATC-related task simulation on the basis of cardiovascular activity and secondary task performance data; (b) train the network for classifying operator states in terms of levels of workload in the ATC task (given specific modes of automation); (c) validate the network; and (d) make assessment of the potential utility of the NN for driving AA in the simulation under different modes of automation. We sought to quantify the differential effectiveness of the NN tool for workload classification, depending on the ATC IP functions to which automation is applied.

As a first step, an experiment was conducted to generate a data set for use in training and validating the NN to classify operator workload in the ATC simulation. We assessed the effects of static automation of various ATC IP functions on operator performance and workload using a battery of measures. An ancillary goal of the experiment was to validate heart measures as reliable indicators of mental workload in the specific task by comparing HR and HRV to the secondary task, and subjective measures of workload across different ATC workload settings (numbers of aircraft). Some prior research (e.g., Jorna, 1992) has found HRV (particularly the midband frequency) to be sensitive to cognitive load manipulations. Other work has proposed that HR may be more of a global measure of load (Wilson et al., 2000) reflecting both physical and mental demands. On the contrary, research (Scerbo et al., 2001) reviewing a large body of evidence on the efficacy of psychophysiological measures for implementing AA has concluded that HR may be more diagnostic than HRV. Our own review of the literature has lead to the same conclusion.

METHOD

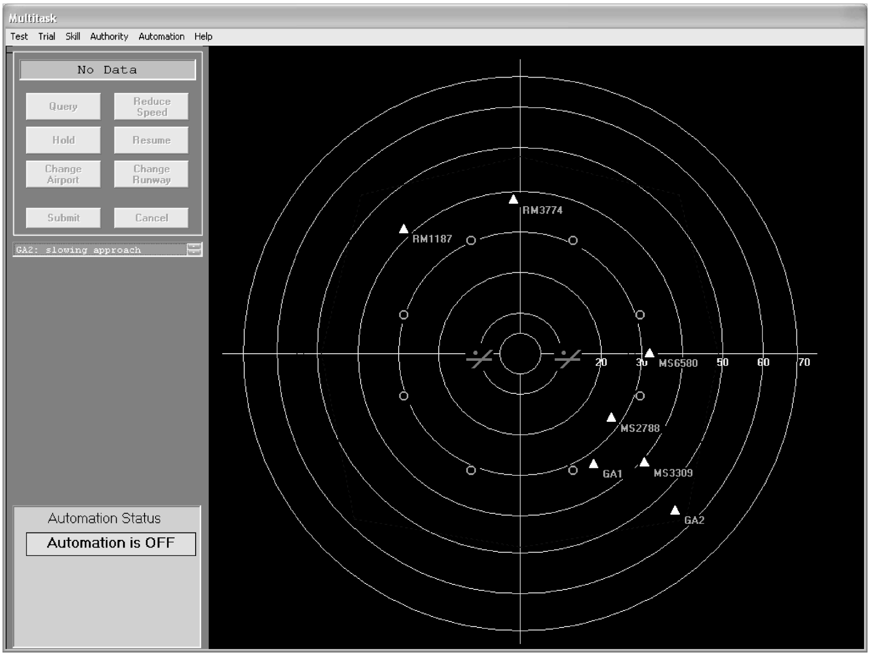
Tasks

The Multitask© Simulation (North Carolina State University, Raleigh, NC) is a PC-based simulation of an ATC-related task developed for studying workload-matched AA of various IP functions (see, e.g., Clamann, Wright, & Kaber, 2002). The task display (presented in Figure 1) includes a radarscope, an aircraft data box, a command and control box, an automation status box, and a menu bar. Near the center of the radarscope are two airports. During simulation run time, aircraft (white triangular icons) appear toward the perimeter of the display on one of eight approach trajectories and move toward one of the airports, destined for one of the two runways at an airport. The objective of the controller is to contact aircraft and make any necessary changes to preexisting clearances (based on their potential to cause a trajectory conflict) while maintaining landing efficiency.

The simulation is capable of operating under one of the following five modes of control:

1. *Manual control.* No automated assistance is provided to controllers. They must establish a communication link with each aircraft; query (virtual) pilots for aircraft flight information; decide whether to issue a revised clearance (e.g., reduce speed, hold, change runway); and implement the clearance using various interface controls.

2. *Information acquisition automation.* A scan line rotates around the radar display. As it passes over an aircraft icon, a trajectory projection aid (TPA) for that aircraft is presented for 2 sec. The TPA shows the aircraft destination and route, as well as its speed and destination airport and runway identifiers. The automation assists operators with acquisition of data on aircraft that would otherwise come from communication with pilots under manual control.



No Data

GA2: slowing approach

Automation Status

Automation is OFF

FIGURE 1 Multitask display in manual control mode.

3. *Information analysis automation*. Information on each aircraft on the radarscope is displayed in a table, including the aircraft's call sign, destination airport, destination runway, speed, and distance from the airport. An additional column presents the call signs of aircraft that are in conflict with each other. This form of automation assists operators with integration of aircraft information they would otherwise need to gather using working memory to make clearance revision decisions.

4. *Decision-making automation.* In addition to the conflict alerting capability provided under the information analysis mode, recommendations for conflict resolution are provided. Information on conflicting aircraft, the recommended clearance change, and which aircraft to advise of the change are displayed in the automation aid box as part of the Multitask interface. This form of automation assists operators with decision and response selection aspects of the task. It gives them some idea of what might be the most important actions to take to prevent collisions, but this advice only comes after conflicts have been detected.

5. *Action implementation automation*. This form of automation simulates the handoff of aircraft control from approach control to local-tower control, and the tower automatically maintains full responsibility for aircraft within 20 nm of the center of the radarscope. This type of automation prevents any conflicts after

handoff to tower control. It assists operators with the requirement of response execution as part of the ATC simulation. Operators do not need to monitor aircraft on final approach or make last-minute clearance revisions.

Performance in the Multitask simulation was measured in terms of the number of aircraft cleared, the number of trajectory conflicts, and actual collisions. Workload in the task was measured using a battery of objective and subjective responses discussed in the following sections.

In our experiment, at the same time participants performed the Multitask simulation, they also performed a gauge-monitoring task to serve as an objective index of workload in the ATC task. The gauge task included a fixed-scale, moving pointer display with a central acceptable region bordered on either side by two unacceptable regions. The user's goal was to detect and correct pointer deviations into either unacceptable region by using a keyboard. Gauge task performance was recorded as a hit-to-signal ratio (the number of unacceptable pointer deviations detected/the total number of deviations).

Participants

Twenty-five participants were recruited for the experiment. All were required to have 20/20, or corrected to normal, vision, to be physically fit (a body mass index less than 29), and to have personal computer experience. Additionally, all participants were required to be between the ages of 18 and 23 to limit substantial variations in heart function due to age. Participants were compensated at a rate of $7.50 per hr for their participation. According to physiological experiment procedures defined by Jorna (1992) and Porges and Byrne (1992), the participants were asked to refrain from smoking and taking caffeine for at least 1 hr preceding their experimental session.

Experimental Design and Dependent Measures

A mixed factorial design was used with the five levels of ATC task automation (LOAs) manipulated as a between-subject variable. The participants were randomly assigned to one of the modes, including completely manual control, information acquisition automation, information analysis automation, decision-making automation, or action implementation automation, to form five groups of five persons. Each participant trained on the ATC task for 20 min under the manual control setting. This was followed by training under the assigned mode of automation for 15 min. Participants then practiced the gauge-monitoring task for 5 min., followed by dual-task practice, under a low workload condition, lasting for 30 min. Our training trial times were based on the procedures of prior research (Endsley & Kaber, 1999) yielding Multitask performance data absent of learning effects. During experimental testing, the ATC task traffic volume (LOAD) was manipulated as a within-subjects variable with two levels, low and high. All participants completed a 30-min trial under each LOAD condition. The order of presentation of LOAD settings was balanced across all participants. Three aircraft appeared on the display at any given time for the low load and seven aircraft appeared on the display for the high load. Each trial consisted of all automated minutes (if a participant was assigned to one of the four automated conditions) or all manual minutes (if a participants was assigned to the manual control condition). There was no adaptation of the automation to operator workload states or cycling between manual and automated control modes because the objective was to quantify operator workload under the various LOAs and traffic volumes in terms of physiological responses and secondary task performance. This information was then to be used to train the NN for classification of operator functional states under specific task circumstances.

In addition to gauge-monitoring performance serving as an indicator of workload, participant heart interbeat interval (IBI) was continuously measured during trials. Data were collected using a Polar *S*810i Heart Rate Monitoring system (Polar Electro Oy, Finland). A wristwatch (receiver) was integrated with a chest strap containing an electrode (transmitter) to sense and record cardiac activity. The watch communicated data from the electrodes to a PC via an infrared connection. These data formed the basis for aggregate HR and HRV measures (low = 0.00–0.04 Hz, mid = 0.04–0.15 Hz, and high = 0.15–0.4 Hz frequency band HRV). The measures were calculated and filtered using the Polar Precision Performance Software and Microsoft Excel. All performance and continuous workload variables were averaged over 1-min periods for the duration of the 30-min trials. (The fixed recording period for the secondary task, as coded in the software application, was 1 min and the period for averaging the IBI data was synchronized with the secondary task.) Finally, the NASA-Task Load IndeX (NASA–TLX) was collected at the end of each experimental trial, as a subjective measure of workload for validating the heart measures (Hart & Staveland, 1988).

Three different baseline measures of HR were collected for each participant at three different times using slightly different procedures, including one prior to experiment trials, when participants were in a resting state; one prior to testing, when participants had been instructed on the tasks and were asked to sit and watch the displays; and one baseline measure after all experiment trials were completed. Of these baselines, the HR measures collected after test trials produced the lowest mean cardiac response and the observations on individual participants were more stable than those recorded at the times of the other two baselines. Relevant to this, Obrist (1981) and Turner and Carroll (1985) contended that, "immediately prestress cardiovascular measurements" might be inappropriate for baseline measurement purposes. For these reasons, only the postexperimental baseline HR measurements were used for our analyses.

Hypotheses

It was expected based on Perry, Segall, and Kaber's (2005) research that specific LOAs, including information acquisition and action implementation, would provide workload relief for operators, leading to primary and secondary task performance increases, as compared to the information analysis and decision-making modes. It was also expected that HR, HRV, and subjective workload measures would be sensitive to the automation manipulations. In general, it was expected that automation would significantly improve primary and secondary task performance, as compared to manual control. Clamann et al. (2002) presented findings that any form of AA of the Multitask simulation proved to be superior to manual control.

With respect to the traffic volume manipulation, it was expected that all workload and performance measures would be sensitive, with higher HR and subjective workload under the high traffic volume, as well as lower primary and secondary task performance and suppressed HRV for this condition, as compared to the low-traffic condition. Our expectations for the cardiocirculatory measures were based on similar findings in historical work (e.g., Veltman & Gaillard, 1996) using a high-fidelity flight simulation. Based on Scerbo et al. (2001), we did expect the simple HR measure to be more diagnostic than HRV measures.

In regard to the interaction of the automation and workload manipulations, it was expected that automation conditions providing assistance with psychomotor behaviors, including information acquisition and action implementation, would be the most effective in terms of supporting controllers in dealing with the high traffic volume. However, previous research did not provide insight into the sensitivity of ATC performance and workload measures to this type of interaction.

RESULTS

Secondary Task (Gauge-Monitoring) Performance

Analysis of variance (ANOVA) results on gauge-monitoring performance revealed a significant effect of the primary task LOAD manipulation, $F(1, 20) = 47.63$, $p < .0001$, and LOA manipulation, $F(4, 20) = 5.74$, $p = .0030$. The mean hit-to-signal ratios for the low and high traffic loads were 0.88 and 0.72, respectively. Duncan's Multiple Range (MR) tests on the LOA effect revealed that action implementation automation (mean hit-to-signal ratio = 0.86) led to significant reductions in workload (increases in gauge performance) ($p < .05$) compared to the information analysis mode of automation ($M = 0.73$). There was no significant interaction effect of the LOA and LOAD manipulations.

Physiological and Subjective Workload Measures

ANOVA results indicated the increase in HR (beats per minute) from baseline measurement to test trials was significantly affected by the LOA manipulation, $F(4, 20) = 7.81$, $p = .0006$. However, counter to our expectation based on Scerbo et al. (2001), neither the traffic load manipulation nor the interaction of the LOA and LOAD settings proved to be significant in effect.

Figure 2 presents a graph of the increase in HR during test trials from the postexperiment baseline for each mode of automation and the manual condition under the low- and high-load conditions. (The interaction plot is presented to also allow for inspection of the data relative to the lack of a LOAD main effect. The means are presented in conventional physiological units.) As can be seen in the plot, the response for information acquisition and action implementation automation was surprisingly higher across both traffic volume settings than all other modes of control. Contrary to our hypothesis, Duncan's MR test on the LOA effect on increase in HR confirmed that information acquisition and action implementation automation, providing assistance with psychomotor subtasks in the ATC simulation, produced significantly higher HR responses ($p < .05$) than decision making, information analysis, and manual control.

The HRV measures recorded during the experiment did prove to be less sensitive to the ATC automation manipulation than the simple HR measure. In agreement with Scerbo et al. (2001), both the low-frequency and high-frequency bands of HRV lacked sensitivity in terms of discriminating one mode of ATC task auto-

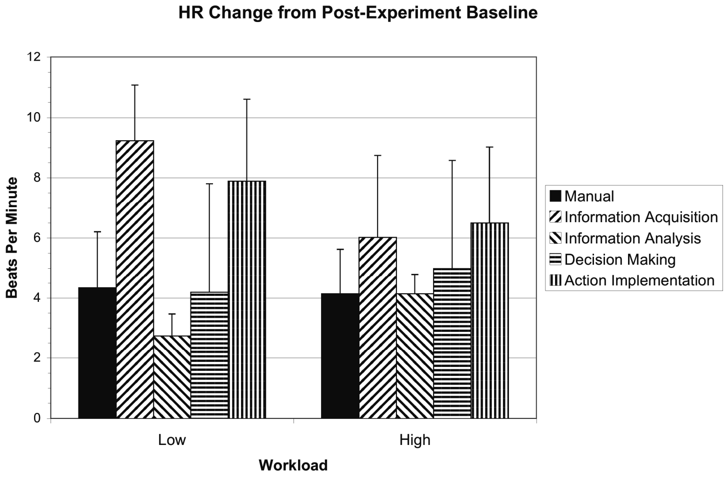


FIGURE 2 Actual change in heart rate from baseline to specific experimental conditions.

mation from another, or manual control. Furthermore, these frequency bands were not sensitive to the traffic volume manipulations. However, in line with Jorna's (1992) research, the power for the midfrequency band HRV response did prove to be marginally sensitive to the LOA manipulation, $F(4, 20) = 2.68$, $p = .0612$, but not to the LOAD setting. Duncan's MR tests revealed that the change in HRV (msec^2) under decision-making automation from the postexperiment baseline ($M = .0141$ [increase]) was significantly greater than for all other modes of automation (information acquisition = .0074 [decrease]; information analysis = .0071 [decrease]; action implementation = .0076 [decrease]) and manual control ($M = .0064$ [increase]). The interaction of the LOA and LOAD manipulations did not prove to be significant for any band of HRV.

The ANOVA results on the overall NASA–TLX workload measure revealed a significant effect of the traffic volume (LOAD) manipulation, $F(1, 20) = 177.37$, $p < .0001$, with the high volume (M rating = 60.6) perceived as significantly more difficult than the low volume ($M = 35.5$). The LOA setting and the interaction of LOA and LOAD did not prove to be significant.

ATC Task Performance Measures

ANOVA results on the ATC task simulation performance revealed a significant effect of traffic volume (LOAD) on the number of aircraft cleared, $F(1, 18) = 329.38$, $p < .0001$, and the number of conflicts, $F(1, 18) = 47.08$, $p < .0001$. Mean aircraft cleared for low and high workloads were 5.2 and 10.9 per trial, respectively. Mean conflicts for low and high workloads were 0.9 and 7.2 per trial, respectively. Aircraft collisions were significantly influenced by the level of ATC task automation (LOA), $F(4, 18) = 3.47$, $p = .0286$. Condition means included: manual control = 0.3/trial, information acquisition = 0.4, information analysis = 0; decision making = 0.1; and action implementation = 0.2. Duncan's MR tests revealed the information acquisition mode of automation to produce significantly worse performance ($p < .05$), compared to information analysis and decision making. Information acquisition also did not prove to be superior to manual control ($p > .05$) in terms of preventing aircraft collisions, nor did the other modes of automation. There was no significant interaction of the LOA and LOAD manipulations on any of the ATC performance measures.

DISCUSSION OF EXPERIMENT

Our results on workload in the ATC task simulation revealed agreement among the secondary task performance measure and subjective ratings of cognitive load. The high-traffic setting was perceived to be much more difficult. As more aircraft appeared on the display screen and required greater visual attention and participant working memory, this may have increased mental and temporal stress, leading to higher NASA–TLX ratings. Regarding the secondary task measure, there was also evidence supporting our contention that specific modes of automation in the ATC-related task would lead to workload relief. However, it was not apparent that action implementation (or information acquisition automation) led to greater gauge performance specifically under the high-traffic condition, as compared to the other LOAs.

We also investigated the use of cardio measures for cognitive workload assessment because of the practicality of implementation for data collection and conflicting results of historical studies (Scerbo et al., 2001; Veltman & Gaillard, 1996). However, there appeared to be limited sensitivity of the HR and HRV responses. Although participants perceived significant differences among traffic conditions, the heart measures did not confirm this. Our HR measure proved to be more diagnostic in terms of indicating differences among the LOA conditions than the HRV measures. This may be attributable to the specific characteristics of the automation conditions, including assistance with psychomotor versus cognitive task performance. Those modes of automation providing forms of decision aiding significantly reduced higher order cognitive processing and led to decreases in HR. The result on midband HRV also indicated decision aiding reduced cognitive load and led to an increase in HRV power. In general, the limited sensitivity of the HRV measure across frequency bands is in agreement with prior NN research (Wilson & Russell, 2003b) demonstrating limited accuracy of networks for predicting human workload based on ECG data. Through comparison of the pattern of results on the HR and HRV measures with those on secondary task performance and subjective ratings of workload, we concluded that the heart measures were less reliable indicators of cognitive load, particularly in terms of the air traffic volume manipulation.

The results on ATC task performance were logical, as the greater the number of vehicles in a sector, the greater the likelihood of conflicts. In agreement with our hypothesis, it was easier for participants to prevent conflicts under the lower traffic load condition. Based on the ANOVA results, there was support for our hypothesis that the various forms of ATC task automation would lead to differences in primary task performance. We expected that greater degrees of information aiding would lead to improvements in performance, including collision prevention. In line with the HRV results, it appeared that decision-making automation (and information analysis automation) supported operator performance in terms of addressing potential conflicts and preventing collisions. These modes of automation provided participants with warnings of conflicts and recommendations for how to deal with them, unlike the information acquisition mode.

With respect to the next step in our research, the experiment allowed us to develop a data set on a battery of objective and subjective measures of simulated ATC task workload and performance under various forms of automation and traffic volumes. The data set was considered to be sufficient in terms of the resolution of the response measures and number of observations for the development of a NN solution for classification of operator functional states. We investigated NNs because of the capability to develop models based on data unconstrained by rigid statistical assumptions, such as those associated with ANOVAs and regression models.

DEVELOPMENT OF A FUNCTIONAL STATE CLASSIFICATION TOOL

The objective of the NN model development was to create an operator functional state classification tool that would determine when manual control of the ATC task should be allocated to operators in the case of low-workload states (reduced task engagement) under all forms of automation, or when automation should be invoked in the case of high-workload states under manual control. In the context of the model, a low-workload classification represented the experimental condition in which participants were required to manage three aircraft at any time, and a high-workload classification represented the experiment condition in which operators were presented with seven aircraft.

We initially used multiple linear regression for selection of input variables for the NN (see Chen, Kaber, & Dempsey, 2000, for a detailed example) among all workload measures recorded during the ATC simulation. We sought to investigate only those predictors that appeared to be significant in explaining objective workload states as inputs to the NN models. Results revealed both the secondary task performance measure and the change in HR from baseline to be significant predictors of actual controller workloads at an alpha criterion of .05. (None of the HRV measures proved to be significant, based on this analysis.) Consequently, these variables were used as inputs to all candidate NN models for classification of operator workload states across LOAs and traffic volumes. (It is important to note here that additional regression models, including LOA as a predictor, were examined and there appeared to be multicollinearity of LOA with the HR term in the models. This suggested that the importance of the cardio measure in predicting operator workload states may have been mediated by the LOA manipulation.) Subsequent to the input variable selection, we used part of the physiological and secondary task performance data to train the network for classifying controller states in terms of actual levels of workload (the traffic volumes). Finally, we attempted to validate the network for use in real-time prediction of workload states by using another portion of the experimental data set (not used in the training process).

Based on the experimental data, a single network output node was used for classification of workload as low or high. The network included HR or secondary task performance as inputs. A number of candidate networks were initially created and trained using NeuroSolutions© software (NeuroDimension Inc., Gainesville, FL). These candidates included networks in which inputs represented aggregate measures over each minute of the 30-min trials, as well as networks where inputs were aggregated over larger periods of time (2 min, 3 min, and 5 min) to investigate the potential for longer term trends in the data. In all cases, the networks were trained with 80% of the data for all participants, with the remaining 20% set aside for validation and testing. Genetic optimization (the Neural Expert module of the Neural Solutions© package) was used to determine the optimal number of hidden network layers, processing elements (PEs) in each layer, the network step size in processing data, and the front- and back-side momentums for each PE in a hidden layer.

Results of Neural Network Development

Initially, NNs with only one input were developed with data on either secondary task performance or HR being used to predict workload states. The predictive accuracy of the NN with the HR input ranged from approximately 39% to 53% for low- and high-workload states, respectively. The best candidate NN with secondary task performance as an input yielded prediction accuracies between 64% and 59% for low and high workload, respectively. As one might expect, based on the regression analysis, the predictive accuracies of the models were less than those for the best dual-input NN.

Table 1 presents the common confusion matrix (Johnson & Wichern, 1992) for the training performance of the dual-input NN producing the highest predictive accuracies among all candidate networks, ranging from approximately 66% for the low-traffic condition down to 59% for the high-traffic load. As a result of the genetic optimization procedure, two hidden neuron layers containing two and three PEs characterized this network. The front- and back-side learning momentums for the PEs across layers ranged from .70 to .86. The validation results for this network were close to the training performance, with approximately 70% network accuracy in predicting high-workload states and roughly 60% accuracy in predicting low-workload states.

Based on the preceding NN results, and the multiple regression analysis, we were concerned that the predictive accuracies of the NNs might be limited by significant individual differences in physiological responses among the participant groups used in the experiment, or among the forms of primary task automation. Al-

TABLE 1
Confusion Matrix for Best Dual-Input NN

	Predicted Workload	
Actual Workload	*Low*	*High*
Low	63.74	36.26
High	41.13	58.87

though the cardio measure used as an input in our NN models was a relative measure (i.e., the difference between the test HR and baseline readings), it is possible that between-group variability was substantially greater than within groups. With this in mind, the experimental data set was parsed (with observations on HR and secondary task performance aggregated over 1-min periods) into five subsets, with each subset including only the observations on one LOA. Data sets were created with physiological responses and secondary task performance observations for information acquisition, information analysis, decision making, and action implementation automation, as well as manual control, of the ATC task simulation. The data sets were then used to train and test additional NNs to predict operator workload states under each specific mode of automation. This approach also allowed us to address the objective of determining the sensitivity of the various workload measures for revealing fluctuations under specific LOAs, and whether there was differential effectiveness of our NN model depending on the automation conditions.

The inputs for the LOA-specific NNs still consisted of HR and secondary task performance. As in the previous network development, 80% of the experiment data were used for training and 20% for validation and testing. With respect to optimization of the architecture, all candidate networks were defined based on the results of the genetic optimization routine used in the prior model development step, with two hidden neuron layers containing two and three PEs, and front- and back-side learning momentums for the PEs ranging from .70 to .86.

Table 2 presents a summary of the prediction accuracies for each LOA and manual control in the validation step.

TABLE 2
Neural Network Validation Performance in Classifying Workload (Number of Aircraft) Under Each LOA

		Predicted	
Primary Task LOA	*Actual*	*3*	*7*
Manual	3	63.15	36.84
	7	26.82	73.17
Information acquisition	3	55.55	44.44
	7	50.00	50.00
Information analysis	3	47.05	52.94
	7	59.52	40.47
Decision making	3	52.63	47.36
	7	58.53	41.46
Action implementation	3	52.63	47.36
	7	21.95	78.04

In general, several of the models produced higher prediction accuracies in high-workload state classification than the NN developed for workload prediction across all LOAs. Specifically, the networks for manual control and action implementation produced high workload classification accuracies between 73% and 78%. However, the accuracies of these networks in classifying low-workload states was relatively poor, ranging from as low as 47% for the information analysis mode of automation to about 63% for the manual control mode. These results are interesting because the classification accuracies support the notion of differential effectiveness of the workload measures for operator state classification under different modes of automation. More specifically, it appears that the simple HR measure and secondary task performance measure may have greater utility for establishing operator workload states under manual control or forms of automation in which computer assistance is provided with lower order psychomotor functions, as compared to networks for predicting operator states under information analysis or decision-making automation.

Discussion of NN Results

Wilson and Russell (2003b) used a NN for operator workload classification in the MAT–B under two levels of difficulty by considering several peripheral physiological variables, including HR, eye blinks, and respiration interval as inputs. Their classification accuracies are very similar to our results, with values ranging from 43.8% to 64.9%, using the three physiological inputs. We reported accuracies from 59% to 66% for an optimized network predicting controller workload based on HR and secondary task performance measures across various forms of simulated ATC task automation. We also observed workload classification accuracies of 47% to 78% for NNs developed based on data for specific modes of ATC task automation. We believe the similarity of these results among studies suggests that the simple HR measure (alone), or HR in combination with only a small number of other performance or physiological measures, may not be sensitive enough to indicate varying levels of operator workload in ATC tasks. Our results suggest that a small range of cardiocirculatory response may be sufficient for operators to deal with the various forms of ATC automation presented during the experiment, and that changes in HR induced by traffic volume manipulations may be even less pronounced. Our results support the contention that HR may be a more “global” measure of cognitive demand (Russell & Wilson, 1998), as compared to a well-designed secondary task or central physiological measures of workload (e.g., EEG).

These observations agree with other historical findings on cardiocirculatory measures of ATC workload under various task load conditions (Brookings, Wilson, & Swain, 1996; Costa, 1993). These studies also pointed to a lack of sensitivity of HR and HRV measures of real and simulated workload in ATC operations. In this study, the secondary task performance measure was more sensitive than the cardiac measures to ATC task load manipulations and was the most important input in the NN models for predicting actual workload states.

CONCLUSIONS

In general, the cardio measures we examined as workload indexes appeared to be limited in sensitivity for discriminating among the use of information acquisition, information analysis, decision making, and action implementation automation in the experimental ATC task or low and high air traffic volume conditions. By extending the investigation to examine the use of NNs to model nonlinear associations of physiological responses with actual ATC workload conditions, our conclusion was that heart measures, although highly practical and relatively easy to collect, would not be sufficient bases for accurate and reliable classification of operator states and triggering dynamic allocations in adaptive systems. These findings may have applicability to real-world ATC tasks as the simulation used in this experiment was more realistic in representation of air traffic approach control operations than prior laboratory simulations used to study NN-based approaches to triggering AA in complex systems control.

Through the NN development effort, the effectiveness of using a combination of different classes of human metrics as inputs to networks for controller workload state classification was studied. The resulting NNs yielded classification accuracies comparable to previous networks developed for simulated piloting tasks using inputs, including HR, EOG, and respiration measures. However, the prediction accuracies of the networks in this study were not as good as those for prior NNs developed to classify air traffic controller workload states using numerous EEG signal inputs (Russell & Wilson, 1998; Wilson & Russell, 2003a). There appears to be a trade-off between NN classification accuracy and the practicality of implementation of the physiological measurement approaches leading to useful NN inputs. Russell and Wilson's (1998) average NN classification tool accuracy was ~83% when they used between 8 and 88 physiological variables as inputs (specifically EEG signals). In general, our combination of secondary task performance and physiology-based measures of workload, which had not been previously explored, was not as effective for producing an accurate NN-based workload state classifier as developing networks trained based on one general type of input, such as EEG signals.

Through our NN development, the differential effectiveness of using network models for operator workload state classification for different modes of ATC automation was also established. The classification accuracies of networks trained on data on controller performance under automation of psychomotor functions and manual control of the ATC simulation for both low and high traffic volumes were higher than those for networks when automation was applied to information analysis and decision functions. It may be possible that the NN inputs investigated were more sensitive to operator workload changes in terms of psychomotor task performance versus cognitive task performance. This is an important new finding as it may be critical to carefully consider the types of workload measures for use with specific forms of automation in attempting to do real-time operator state classification in augmented cognitive systems of the future.

One caveat of this study is that in the experiment we did not recruit actual air traffic controllers. We studied trained university students. It is possible that the stress response of the student participants in operating the Multitask simulation may not have approached that of actual controllers in approach control, who are concerned with the potential for loss of life due to aircraft collisions or the impact of the occurrence of aircraft trajectory conflicts on their careers.

Related to this, the Multitask simulation is only a limited fidelity representation of air traffic approach control. There are a number of features that the simulation lacks with respect to the actual real-world task. Specifically, in this study, participants were not responsible for providing aircraft with altitude clearances. They were constrained to managing potential trajectory conflicts through other clearances, including speed changes, holds, and redirects or airport changes.

One interesting direction for future research would be to consider the use of secondary task indicators of workload in combination with other physiological responses that have proved to be highly sensitive to changes in operator arousal states in complex systems control. Substantial work has demonstrated the effectiveness of, for example, EEG indexes of arousal as a basis for facilitating dynamic function allocation in laboratory simulations of adaptive systems (Wilson & Russell, 2003a, 2003b). It would be interesting to explore combinations of EEG-based measures of cognitive state and secondary task performance measures as inputs to an NN workload classification tool for use in adaptive system control.

ACKNOWLEDGMENTS

This research was supported by a grant from NASA Langley Research Center (No. NNL04AA18G), Airspace Systems Program, Human Measures and Performance Element. The technical monitor of this research project was Lance Prinzel, and we thank him for suggestions on the approach to the experiment and NN development in the proposal phase. The views and opinions expressed in this article are those of the authors and do not necessarily reflect the views of NASA. We thank the anonymous reviewers and the editor for their evaluation of our work.

REFERENCES

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

Bennett, K. B., Cress, J. D., Hettinger, L. J., Stautberg, D., & Haas, M. W. (2001). A theoretical analysis and preliminary investigation of dynamically adaptive interfaces. *International Journal of Aviation Psychology, 11,* 169–195.

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology, 42,* 361–377.

Byrne, E. A., & Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological Psychology, 42,* 249–268.

Chen, C.-L., Kaber, D. B., & Dempsey, P. G. (2000). A new approach to applying feedforward neural networks to the prediction of musculoskeletal disorder risk. *Applied Ergonomics, 31,* 269–282.

Clamann, M. P., Wright, M. C., & Kaber, D. B. (2002). Comparison of performance effects of adaptive automation applied to various stages of human–machine system information processing. In *Proceedings of the Human Factors Society 46th annual meeting* (pp. 342–346). Santa Monica, CA: Human Factors and Ergonomics Society.

Comstock, J. R., & Arnegard, R. J. (1992). *Multi-attribute task battery* (NASA Tech. Memo. No. 104174). Hampton, VA: NASA Langley Research Center.

Costa, G. (1993). Evaluation of workload in air traffic controllers. *Ergonomics, 36,* 1111–1120.

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics, 42,* 462–492.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA–TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam: North-Holland Elsevier.

Hilburn, B., Jorna, P., Byrne, E. A., & Parasuraman, R. (1997). The effect of adaptive air traffic control (ATC) decision aiding on controller mental workload. In M. Mouloua & J. M. Koonce (Eds.), *Human–automation interaction: Research and practice* (pp. 84–91). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Jorna, P. G. A. M. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological Psychology, 34,* 237–257.

Kaber, D. B., Perry, C. M., Segall, N., McClernon, C. K., & Prinzel, L. P. (2006). Situation awareness implications of adaptive automation for information processing in an air-traffic control related task. *International Journal of Industrial Ergonomics, 36,* 447–462.

Kaber, D. B., & Riley, J. (1999). Adaptive automation of a dynamic control task based on secondary task workload measurement. *International Journal of Cognitive Ergonomics, 3,* 169–187.

Kaber, D. B., Wright, M. C., Prinzel, L. J., & Clamann, M. P. (2006). Adaptive automation of human–machine system information processing functions. *Human Factors, 47,* 730–741.

Obrist, P. A. (1981). *Cardiovascular psychophysiology.* New York: Plenum.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation induced complacency. *International Journal of Aviation Psychology, 3,* 1–23.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model of types and levels of human interaction with automation. *IEEE Transactions on Systems, Man and Cybernetics, 30,* 286–297.

Perry, C. M., Segall, N., & Kaber, D. B. (2005). Measurement of situation awareness effects of adaptive automation of air traffic control information processing functions. In R. S. Jensen (Ed.), *Proceedings of the 13th International Symposium on Aviation Psychology* (pp. 457–462). Dayton, OH: Wright State University.

Porges, S. W., & Byrne, E. A. (1992). Research methods for measurement of heart rate and respiration. *Biological Psychology, 34,* 93–130.

Prinzel, L. J., 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.

Russell, C. A., & Wilson, G. F. (1998). Air traffic controller functional state classification using neural networks. In C. H. Dagli, M. Akay, A. L. Buczak, O. Ersoy, & B. Fernandez (Eds.), *Proceedings of the Artificial Neural Networks in Engineering (ANNIE '98) conference* (Vol. 8, pp. 649–654). New York: ASME Press.

Scerbo, M. W., Freeman, F. G., Mikulka, P. J., Parasuraman, R., DiNocero, F., & Prinzel, L. J. (2001). *The efficacy of psychophysiological measures for implementing adaptive technology* (Tech. Rep. No. NASA/TP-2001-211018). Washington, DC: NASA.

Turner, J. R., & Carroll, D. (1985). Heart rate and oxygen consumption during mental arithmetic, a video game, and graded exercise: Further evidence of metabolically-exaggerated cardiac adjustments? *Psychophysiology, 22,* 261–267.

Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology, 42,* 323–342.

Wilson, G. F. (2001). Real-time adaptive aiding using psychological operator state assessment. In D. Harris (Ed.), *Engineering psychology and cognitive ergonomics: Volume 6. Industrial ergonomic, HCI, and applied cognitive psychology* (pp. 175–182). Aldershot, UK: Ashgate.

Wilson, G. F., Lambert, J. D., & Russell, C. A. (2000). Performance enhancement with real-time physiologically controlled adaptive aiding. In *Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th annual meeting of the Human Factors and Ergonomics Society* (pp. 503–506). Santa Monica, CA: Human Factors and Ergonomics Society.

Wilson, G. F., Monett, C. T., & Russell, C. A. (1997). Operator functional state classification during a simulated ATC task using EEG. In *Proceedings of the 41st annual meeting of the Human Factors and Ergonomics Society* (pp. 1099–1102). Santa Monica, CA: Human Factors and Ergonomics Society.

Wilson, G. F., & Russell, C. A. (2003a). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors, 45,* 381–389.

Wilson, G. F., & Russell, C. A. (2003b). Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human Factors, 45,* 635–643.

Manuscript First Received: June 2006