3rd USA/Europe Air Traffic Management R&D Seminar Napoli, 13-16 June 2000

Investigating the Validity of Performance and Objective Workload Evaluation Research (POWER)

Carol A. Manning, Scott H. Mills, Cynthia M. Fox, and Elaine Pfleiderer
FAA Civil Aeromedical Institute, Oklahoma City, OK
and Henry Mogilka
FAA Academy, Oklahoma City, OK
Carol.Manning@FAA.gov

Summary

Performance and Objective Workload Evaluation Research (POWER) was developed to provide objective measures of ATC taskload and performance. POWER uses data extracted from National Airspace System (NAS) System Analysis Recording (SAR) files to compute a set of objective measures. A study was conducted to investigate the relationship of POWER measures with measures of sector complexity, controller workload, and performance. Sixteen instructors from the FAA Academy in Oklahoma City, OK, watched eight traffic samples from four en route sectors in the Kansas City Center using the Systematic Air Traffic Operations Research Initiative (SATORI) system. POWER measures were computed using the same data. Participants made three estimates of the workload experienced by radar controllers and provided two types of assessments of their performance. Sector complexity was determined using information about sector characteristics and the traffic samples. Some POWER measures were related to sector complexity and controller workload, but the relationship with performance was less clear. While this exploratory study provides important information about the POWER measures, additional research is needed to better understand these relationships. When the properties and limitations of these measures are better understood, they may then be used to calculate baseline measures for the current National Airspace System.

Introduction

Need for measuring ATC workload, taskload, complexity, and performance

It is necessary to measure workload, taskload, complexity, and performance in air traffic control (ATC) to evaluate the effects of new systems and procedures on individual air traffic controllers and on the ATC system as a whole (Wickens, Mavor, Parasuraman, & McGee, 1998). The effects of using different display designs or alternative procedures on controllers' workload and performance must be assessed before they are implemented. When new ATC systems are introduced in field facilities, it is necessary to document their effects on individual and system performance, both soon after implementation and after controllers have become accustomed to using them. Computing measures of taskload and performance on a system level, while accounting for sector complexity, may also contribute to better prediction of overloads at specific sectors.

Defining controller workload, taskload, sector complexity, and performance

While many methods have been used to measure ATC workload, taskload, sector complexity, and performance, definitions of the terms are not widely agreed upon. In general, workload typically refers to the physical and mental effort an individual exerts to perform a task. In this sense, ATC workload may be differentiated from taskload in that taskload refers to air traffic events to which the controller is exposed, whereas workload describes the effort expended by the controller to manage those events.

Sector complexity describes the static and dynamic characteristics of the air traffic environment that combine with the taskload to produce a given level of controller workload (Grossberg, 1989). In that sense, complexity can mediate the relationship between taskload and workload.

Federal Aviation Administration (FAA) *Air Traffic Control* (Order 7110.65M, 2000) states "The primary purpose of the ATC system is to prevent a collision between aircraft operating in the system and to organize and expedite the flow of traffic." Thus, measurement of

controller performance involves determining the effectiveness with which an individual controller's activities accomplished these goals.

Measures of ATC workload, taskload, sector complexity, and performance

Many methods have been developed to measure workload, taskload, sector complexity, and performance (see Hadley, Guttman, & Stringer, 1999, for a database containing 162 measures). The dynamic nature of ATC (encompassing both movement of an individual aircraft and constant changes in relative positions of multiple aircraft) makes it necessary to take the passage of time into consideration when measuring these constructs. Even when time is considered, it is more difficult to measure controller performance and workload than it is to measure taskload and sector complexity. The reason is that taskload can be measured by counting recorded ATC events and sector complexity can be measured by recording observable sector characteristics and other factors about the ATC situation. Controller workload and performance, on the other hand, include factors that cannot be easily observed, and are, therefore, not easy to measure. For example, controllers constantly review aircraft positions, directions, and speeds, and mentally project aircraft positions but may only occasionally take observable actions. It is possible to count or otherwise evaluate certain observable activities, such as making keyboard entries and marking or moving flight progress strips. However, the relationship between these measures (taskload) and the amount of cognitive effort expended (workload) or the effectiveness of the results (performance) is unclear. Even actions that appear to be easily interpretable (e.g., commission of operational errors resulting in losses of separation) may not be very meaningful because they occur so infrequently as to be of little value in assessing individual performance.

This section discusses some of the methods that have been used to measure controller workload, taskload, sector complexity, and performance. The advantages and disadvantages of using these methods will also be discussed.

Workload measures. Subjective workload, controllers' reactions to the taskload they experienced, is expected to include components that cannot be explained by taskload alone. Measures of subjective workload in ATC may be obtained either during a simulated scenario or after its completion. For example, the NASA Task Load Index (TLX; Hart & Staveland, 1988) is obtained after the completion of a scenario. Controllers provide separate ratings for each of six scales: mental demand, physical demand, temporal demand, effort, frustration, and performance.

In contrast, the Air Traffic Workload Input Technique (ATWIT) measures workload in "real-time" (Stein, 1985). The ATWIT presents auditory and visual cues (a tone and illumination, respectively) that prompt a controller to press one of seven buttons within a specified amount of time to indicate the amount of workload experienced at that moment. The Workload Assessment Keypad (WAK) device records each rating as well as the time it took to respond to the prompt.

The primary advantage of using a real-time workload measure is that the respondent can report the experience soon after it occurs. However, the process of providing a real-time rating may increase the controller's perceived workload or, worse yet, may interfere with the performance of certain tasks. On the other hand, a workload rating obtained after a scenario is complete may be overly influenced by early or more recent events or the rater may forget to consider certain events altogether.

Taskload measures. Several measures describing controller taskload have been derived from recordings of either operational National Airspace System [NAS] activities or simulation data. For example, Buckley, DeBaryshe, Hitchner, & Kohn (1983) developed a set of computer-derived measures obtained during ATC simulations. They identified four factors that summarized the measures: conflict, occupancy, communications, and delay. Galushka, Frederick, Mogford, & Krois (1995) used counts of controller activities (and Over-the-Shoulder [OTS] subjective performance ratings) to assess en route air traffic controller baseline performance during a simulation study.

Using data extracted from the Log and Track files generated by the Data Analysis and Reduction Tool (DART; Federal Aviation Administration, 1993), Mills (see Mills, Manning, & Pfleiderer, 1999) developed an extensive set of computer-derived taskload measures. Performance and Objective Workload Evaluation Research (POWER) software measures information about controlled aircraft, handoffs, number of altitude changes, number of controller data entries and data entry errors, and variations in aircraft headings, speeds,

and altitudes. Mills (2000) described the computation of these measures in more detail.

Complexity measures. Several measures of sector complexity have also been developed. These typically include physical characteristics of sectors and factors specific to the AT situation. For example, Grossberg (1989) identified three groups of factors (control adjustments such as merging, spacing, and speed changes; climbing and descending flight paths; and mix of aircraft types) that contributed to the complexity of operations in different sectors. Mogford, Murphy, Roske-Hofstrand, Yastrop, & Guttman (1994) identified 15 complexity factors using multidimensional scaling techniques.

The complexity construct has also been found useful in research. Rodgers, Mogford, & Mogford (1998) found a significant multiple correlation between the overall rate of operational errors at Atlanta Center and Mogford et al.'s (1994) 15 complexity factors.

If sector information is available, it should be relatively easy to measure these factors. The number of factors necessary to describe sector complexity is not clear (though the constructs proposed should be closely related). Nevertheless, it appears that the complexity construct may provide information beyond what is available from the taskload construct.

ATC Performance measures. One of the challenges associated with measuring controller performance is evaluating the different approaches controllers use to control traffic. Most approaches used by a controller to maintain aircraft separation and a smooth flow of air traffic would be considered acceptable. However, such individuality of technique makes it difficult to evaluate the effectiveness of an individual controller's actions to move a set of aircraft through a sector.

To accommodate these differences in technique, Subject Matter Expert (SME) observations are the most frequently used methods for measuring ATC performance. Several processes have been developed to record SME observations. The Behavioral Summary Scales (BSS) were developed as a criterion measure against which the Air Traffic Selection and Training (AT-SAT) selection battery (Caliber Associates, 1999) could be validated. The BSS scales included ten distinct performance categories and measured "typical" rather than "maximum" performance; that is, how well controllers performed consistently over time, rather than how well they could perform under peak traffic conditions.

Several other procedures have been developed to evaluate "maximum" performance (during high fidelity simulations.) For example, Bruskiewicz, Hedge, Manning, & Mogilka (2000) developed two other procedures for measuring controller performance that were used in a high-fidelity simulation study conducted to evaluate the AT-SAT performance measures. These were the Over-the-Shoulder (OTS) rating form and the Behavior and Event Checklist (BEC). The OTS rating form, used to evaluate controller performance across broad dimensions, was based in part on the BSS. The BEC was used to record specific mistakes made during the simulation exercises.

The advantage of using SME observations as a basis for evaluating controller performance is that SMEs (especially instructors involved in controller training) possess detailed knowledge about the job and, thus, can evaluate aspects of controllers' behavior beyond what can be obtained from counting events. They are also very accustomed to observing the actions of other controllers.

However, several problems may be associated with SME observations. First, determining appropriate performance ratings and identifying mistakes requires considerable interpretation on the part of the observer. To assure the reliability of these subjective ratings and error counts, extensive SME training and practice sessions are required. It is also not always possible to obtain SME observations because few controllers are available to participate in these activities.

Purpose of study

Our challenge was to develop a set of measures describing aspects of ATC activity that is objective, reliable, valid, and easy to obtain. Measures (such as SME observations or subjective workload ratings) that may have more apparent validity than taskload measures are frequently not available, while recorded ATC data usually are. On the other hand, POWER measures may be insufficient because (as discussed above) recorded ATC data may not sufficiently describe controller workload or performance.

The goal of this study was to determine whether taskload measures (specifically, the POWER measures) derived from routinely recorded ATC data could

sufficiently describe controller workload, sector complexity, and performance. While an extensive set of POWER measures has been computed, as yet, no empirical evidence is available to indicate whether these numbers actually measure the constructs they were intended to measure. The study was also intended to provide some preliminary information about the meaning of the POWER measures.

In particular, we hypothesized that some POWER measures may be related to measures of controller workload and/or performance for individual controllers (See Table 1). Likewise, some POWER measures may also reflect ATC complexity.

If the POWER measures are found to relate to measures of sector complexity, controller workload, or performance, it may be possible to use them in situations where it would not otherwise be possible to evaluate these variables (when SMEs are unavailable or controllers could not provide workload evaluations). For example, a validated set of POWER measures could provide information that would allow post-implementation evaluation of the operational effects of new ATC systems.

Method

Participants

Participants were 16 en route air traffic control instructors from the FAA Academy in Oklahoma City, OK. All were previously fully-qualified controllers at en route Air Route Traffic Control Centers (ARTCCs.) Two participants previously controlled traffic at some of the sectors represented in the traffic samples, though none worked all sectors included in the study.

Materials

Traffic samples. System Analysis Report (SAR) and voice communication tapes were obtained for 12 traffic samples obtained from four sectors in the Kansas City ARTCC. The traffic samples consisted of only routine operations and contained no accidents or incidents.

The SAR data used for the traffic samples were extracted by the DART and National Track Analysis Program (NTAP; Federal Aviation Administration, 1991) programs. Resulting files were processed both by SATORI (Rodgers & Duke, 1993) and POWER (Mills, Manning, & Pfleiderer, 1999) software. SATORI synchronizes information from DART and NTAP files with tapes containing the Radar (R) controller's voice communications, using the time code common to both data sources. POWER uses some of the same files from DART and NTAP to compute measures of sector and controller activity.

Three traffic samples were re-created for each of the four sectors. One traffic sample (used for training) was eight minutes long. The two experimental traffic samples were both 20 minutes long.

Sector training materials. Computerized training sessions were developed that described characteristics and procedures applicable to each sector. Participants examined copies of sector maps on which important information was highlighted. These maps and a copy of the sector binder (containing additional sector information) were available for the participants to review while they watched the traffic samples. Participants also had access to flight plan information (derived from flight strip messages) for each aircraft controlled by the sector during the traffic sample.

Workload and performance measures. Participants provided three types of workload measures (the ATWIT, the NASA TLX, and an estimate of the traffic sample's activity level) and two types of performance measures (an OTS form and a BEC) for each traffic sample they observed. ATWIT ratings were elicited every four minutes during each traffic sample using the WAK. While controllers typically rate their own workload, in this study, the participants used the WAK to rate the amount of workload they thought the R controller experienced in reaction to the taskload that occurred during the traffic sample.

Participants also completed the NASA TLX after each traffic sample. Again, instead of rating their own workload, participants rated their perception of the workload experienced by the R controller. The TLX ratings were entered using a computerized form. The activity level rating was provided for each traffic sample using a 5-point scale ranging from "Not at all busy" to "Very busy."

Participants rated controller performance using a revised version of the OTS form originally developed for the AT-SAT high-fidelity validation study. The OTS form was revised because participants had access to only the R controllers' voice communications and, thus, were unable to evaluate all the events that occurred at the

sector during the traffic sample. Participants also used the BEC to record errors made by the R controller during the traffic sample.

Sector Complexity. The complexity measures used in the study were based on Mogford et al.'s (1994) 15 complexity factors. These factors were combined into three variables. Static complexity included numbers of adjacent sectors, transfer control points, sequencing functions, military operations, major airports, VORTACS, intersections, miles of airways, shelves, and airspace size. This information was derived from letters of agreement for each sector and from Kansas City ARTCC's Adaptation Control Environmental System (ACES) map files.

Dynamic complexity included numbers of pilot/controller transmissions, interphone communications, maximum number of Hs and Ls displayed (indicating high and low weather activity), amount of climbing/descending traffic, percentages of jets and VFR aircraft, number of military aircraft, percentages of arrivals/departures for St. Louis airport, clearances issued for traffic, altitude and speed restrictions issued, conversations about holding, and a variable reflecting traffic volume. This information was derived from the traffic samples. An overall complexity variable was computed by combining the static and dynamic complexity variables.

Procedure

Participants reviewed a description of the purpose and methods for the experiment, completed consent and biographical information forms, then viewed descriptions of the workload and performance measures. For each of the four sectors, participants 1) reviewed training materials, 2) observed one 8-minute training traffic sample, and 3) observed two 20-minute experimental traffic samples. To ensure continuity, all traffic samples for a sector were shown together as a block. The order in which the four blocks of traffic samples were observed was counter-balanced, as was the order in which the two experimental traffic samples within each block were presented.

As each traffic sample progressed, participants recorded any mistakes they observed on the BEC. The ATWIT aural signal occurred every four minutes. Participants responded by entering a number between 1 and 7 on the WAK keypad. At the end of the traffic sample, participants completed the NASA TLX using a computerized form, summed the errors they had marked on the BEC, then completed the OTS rating form. Finally, they rated the activity level for that traffic sample.

Completing the training process and observing the three traffic samples for each sector required about 1½ hours. After completing the observations for all four sectors, participants answered questions about their experiences during the evaluation process.

Results

To simplify the analysis, the original 24 performance and workload variables were combined into 8 categories (see Manning, Mills, Pfleiderer, Fox, & Mogilka, 2000, for a description). The five performance categories included three factors derived from the BEC describing types of errors made: (1) Inactivity, 2) Disorganization, and 3) Inefficient but Safe), the Overall OTS rating, and the TLX Performance scale. The three workload categories were 1) a combination of the Mental, Physical, Temporal, and Effort TLX scales into a single scale called "Demand," 2) the TLX Frustration scale, and 3) the average of the ATWIT and SME activity level ratings. Values for the 8 categories (averaged across raters) were computed for each traffic sample.

The static, dynamic, and overall complexity factors were computed and POWER measures were obtained for the eight experimental scenarios. These data were matched with the nine performance and workload measures, and descriptive statistics were computed (see Table 1.)

During the 20-minute traffic samples, some POWER measures did not occur (e.g., immediate alerts; data entries, errors, and pointouts for A-side controllers; start track; and hold entries). These were not included in the analysis. Other measures (such as data controller, or D side, entries) were also eliminated from the analysis.

Table 2 shows correlations of the remaining POWER measures with the complexity, performance, and workload measures. Because this was an exploratory study, correlations significant at the .10 level or lower (italicized) were displayed as well as those significant at the .05 level or lower (bolded). Nevertheless, since the number of traffic samples analyzed was so small (N=8) and the number of correlations computed was so large (N=338), it is likely that some of the statistically significant correlations could have occurred due to

Table 1. Expected relationships and descriptive statistics for POWER measures (N=8).

	Expected Relationships			Descriptive Statistics	
Power Measure	Complexity	Performance	Workload	Mean	SD
Total N aircraft controlled	X		X	15.25	5.23
Max aircraft simultaneously controlled	X		X	6.88	2.47
Average time aircraft under control	X	X	X	389.75	97.62
Avg Heading variation	X	X	X	11.64	3.02
Avg Speed variation	X	X	X	1.28	.68
Avg Altitude variation	X	X	X	.84	.51
Total N altitude changes	X	X	X	12.13	5.38
Total N handoffs	X		X	19.50	7.37
Total N handoffs accepted			X	5.88	3.98
Avg time to accept handoff		X	X	39.22	19.54
Total N handoffs initiated			X	10.0	4.0
Avg time until initiated HOs are accepted			X	50.47	27.26
Total N data entries			X	66.38	23.37
Total N data entry errors			X	1.50	1.60
N Radar controller data entries		X	X	56.75	22.70
N Radar controller data entry errors		X	X	1.13	1.36
N Route displays		X	X	2.00	2.27
N Radar controller pointouts		X	X	0.38	0.74
N data block offsets		X	X	0.75	0.89
Total N Conflict Alerts		X	X	0.38	0.52
Number of Conflict Alert suppression entries		X		0.13	0.35
N Distance Reference Indicators requested		X	X	0.25	0.46
N Distance Reference Indicators deleted		X	X	0.13	0.35
N track reroutes			X	0.38	0.74
N strip requests			X	0.13	0.35

chance. However, this result is less likely if a POWER measure was correlated with more than one measure of a construct or if several similar POWER measures were correlated with the same construct.

Relationship with Sector Complexity. Several POWER measures were significantly related to the sector complexity measures. Higher speed variation was related to higher complexity using all three measures (static, dynamic, and overall). Fewer handoffs accepted, R controller pointouts, and data block offsets were also related to higher static complexity (but not to dynamic or overall complexity).

More track reroutes, fewer numbers of handoffs (initiated and accepted or just initiated), shorter times to accept a handoff, and longer times for handoffs to be accepted by an outside sector were related to both higher dynamic and overall complexity. More conflict alert suppressions were related to higher overall complexity (but not static or dynamic complexity).

Relationship with Workload. Many POWER measures were significantly related to most or all of the workload measures. More aircraft controlled, controlled simultaneously, and accepting more handoffs were related to higher workload using all three measures. More total handoffs and handoffs initiated, making more data entries (total and for the R controller), and the presence of more conflict alerts were also related to both higher demand and ATWIT/ Activity level. Furthermore, making more pointouts (R controller) and more data block offsets were significantly related to higher TLX Frustration. Finally, making more altitude changes was related to higher TLX demand.

Relationship with Performance. Notably, only a few POWER measures were related to controller performance, and the relationships were not consistent across the different types of performance measures. Making more R controller pointouts, accepting more handoffs, lower heading variation, and making more

Table 2. Observed relationships of POWER measures with Sector complexity, Controller performance, and Workload (N=8).

	Observed Relationships										
	Complexity			Performance					Workload		
Power Measure	Static	Dyna-mic	Comp-lexity	F1: Inactiv	F2: D-org	F3: Inef-S	OTS	TLX P	DEMD	TLX F	AT/ AL
Total N aircraft controlled	-.61	-.55	-.47	.46	-.43	.62	.36	.23	**.82**	**.77**	**.80**
Max aircraft simultaneously controlled	-.46	-.26	-.22	.56	-.30	**.78**	.43	.26	**.89**	*.66*	**.87**
Average time aircraft under control	-.10	.29	.12	.47	.40	.60	-.05	.29	.43	.25	.41
Avg Heading variation	.59	-.02	.19	**-.71**	-.11	-.57	.45	**-.91**	.05	-.44	.08
Avg Speed variation	**.75**	**.92**	**.82**	-.29	.51	-.12	-.10	-.38	-.13	-.39	-.21
Avg Altitude variation	.52	.52	.36	-.25	**.72**	-.18	-.09	-.24	-.10	-.48	-.07
Total N altitude changes	.26	.46	.37	..04	.51	.38	.14	-.26	*.63*	.21	.52
Total N handoffs	-.56	**-.74**	-*.64*	.25	-.42	.35	.42	.09	*.70*	.58	**.75**
Total N handoffs accepted	-*.66*	-.15	-.27	**.79**	-.01	**.91**	-.02	*.65*	*.69*	**.82**	*.62*
Avg time to accept handoff	-.41	**-.74**	-*.64*	-.12	-.20	-.22	-.01	-.05	-.01	.19	.01
Total N handoffs initiated	-.55	**-.79**	-*.69*	.24	-.44	.29	.42	.11	*.63*	.50	**.72**
Avg time until initiated HOs are accepted	.57	*.69*	**.73**	-.42	-.06	-.38	-.01	-.18	-.52	-.48	-.61
Total N data entries	-.37	-.47	-.37	.17	-.30	.43	*.66*	-.03	**.88**	.45	**.89**
Total N data entry errors	.10	.19	.33	-.09	-.58	-.15	.14	.08	-.40	-.20	-.44
N Radar controller data entries	-.33	-.48	-.35	.21	-.35	.44	*.66*	-.08	**.85**	.41	**.91**
N Radar controller data entry errors	.02	.29	.36	.08	-.42	.02	-.00	.25	-.38	-.11	-.44
N Route displays	-.26	-.31	-.40	-.16	.20	-.04	.33	.18	.33	.03	.36
N Radar controller pointouts	**-.87**	-.34	-.44	**.95**	-.15	**.89**	-.34	**.80**	.42	**.93**	.35
N data block offsets	**-.83**	-.42	-.55	*.70*	-.01	**.72**	-.04	**.75**	.55	**.75**	.52
Total N Conflict Alerts	.18	-.04	.07	-.07	-.24	.24	**.83**	-.39	**.79**	.05	**.86**
Number of Conflict Alert suppression entries	.58	.59	*.64*	-.10	.16	.21	.26	-.58	.46	-.01	.38
N Distance Reference Indicators requested	-.18	.03	-.15	-.01	.51	.08	.05	.30	.15	-.03	.13
N Distance Reference Indicators deleted	.07	.32	.08	-.24	**.79**	-.21	-.35	.19	-.23	-.17	-.32
N track reroutes	.58	**.71**	*.65*	-.15	.52	.10	.08	-.46	.33	-.09	.21
N strip requests	.35	.50	.61	-.09	-.47	-.17	-.01	.00	-.55	-.33	-.58

Note: All statistically significant correlations are shaded. Bolded correlations are significant at $p < .05$. Italicized correlations are significant at $p < .10$. Abbreviations for Observed Relationships (Performance: Inactiiv = Inactivity factor, D-org = Disorganization factor, Inef-S = Inefficient but Safe factor; OTS = Over-the-Shoulder rating; TLX P = TLX Performance Scale; Workload: DEMD = TLX Demand [Mental, Physical, Temporal, Effort scales], TLX F = TLX Frustration scale, AT/AL = ATWIT/Activity Level rating).

data block offsets were related to the Inactivity factor. Deleting more Distance Reference Indicators and higher altitude variation were related to the Disorganization factor. Several POWER measures (e.g., more aircraft controlled simultaneously, more handoffs accepted, R controller pointout entries, data block offsets) were related to the Inefficient but Safe factor. More conflict alerts and data entries (both total and R controller) were associated with higher OTS ratings. Finally, lower heading variation, making more pointouts and data block offsets, and accepting more handoffs were related to worse TLX performance ratings (note that the TLX performance scale was reversed [higher ratings indicate worse performance].)

Conclusions

The statistics in Table 2 provide some interesting information about the relationships between controller and sector activities and the constructs of complexity, workload, and performance. POWER measures seem to describe sector complexity and controller workload reasonably well. It appears that certain aircraft and controller activities, such as higher variation in speed across sectors, data block offsets, and pointouts, may reflect differences in sector structure and function that may influence the need for controllers to take certain actions more frequently.

Likewise, controller workload seems to be related to certain fairly straightforward aircraft and controller activities, such as numbers of aircraft, maximum aircraft controlled, and numbers of data entries. Handoffs and conflict alerts also seem to be relevant measures.

Interpretation of the relationships between POWER measures and controller performance is less clear. Measures predicting the Inefficient but Safe also predicted the TLX performance rating but were different than the measures predicting the OTS rating and Disorganization. Errors related to the Disorganization factor were marginally related to some measures not associated with other performance measures. The Inactivity factor was related to some POWER measures that predicted performance and some that predicted workload, just as the OTS rating was related to some POWER measures that predicted workload.

Several POWER measures appear to be unrelated to any measures of complexity, workload, or performance. These include average time aircraft are under control, number of data entry errors, route displays, distance reference indicators requested, and strip requests. While this exploratory study has provided important information about the POWER measures, additional research is needed to better understand the relationships observed here. Other analyses should investigate whether variation in sectors, time of day, or individual differences between controllers affect the use of POWER measures. Other analyses should identify a reduced set of POWER measures that may sufficiently account for differences in controller complexity, performance, and workload. When the properties and limitations of these measures are better understood, they may then be used to calculate baseline measures for the current National Airspace System.

References

Bruskiewicz, K. T., Hedge, J. W., Manning, C. A., & Mogilka, H. J. (2000, January). The development of a high fidelity performance measure for air traffic controllers. In C. A. Manning (Ed.) *Measuring Air Traffic Controller Performance in a High-Fidelity Simulation.* (Report No. DOT/ FAA/AM-00/2). Washington, DC: FAA Office of Aviation Medicine.

Buckley, E. P., DeBaryshe, B. D., Hitchner, N., & Kohn, P. (1983*). Methods and measurements in real-time air traffic control system simulation* (Report No. DOT/FAA/CT-83/26). Atlantic City, NJ: DOT/FAA Technical Center.

Caliber Associates. (1999). *Documentation of Validity for the AT-SAT Computerized Test Battery*. (Contract number DTFA01-95-C-00052). Alexandria, VA: Author.

Federal Aviation Administration. (1991). *Multiple Virtual Storage (MVS); Subprogram Design Document; National Track Analysis Program (NTAP).* (NASP-9114-H04). Washington, DC: Author.

Federal Aviation Administration. (1993). *Multiple Virtual Storage (MVS); User's Manual; Data Analysis and Reduction Tool (DART).* (NASP-9247-PO2). Washington, DC: Author.

Federal Aviation Administration. (2000, February). *Air Traffic Control.* Order 7110.65M. Washington, DC: Author.

Galushka, J., Frederick, J., Mogford, R., & Krois, P. (1995, September). *Plan View Display Baseline Research Report.* (Report No. DOT/FAA/CT-TN95/45). Atlantic City, NJ: Federal Aviation Administration Technical Center.

Grossberg, M. (1989, April). Relation of airspace

complexity to operational errors. *Quarterly Report of the Federal Aviation Administration's Office of Air Traffic Evaluation and Analysis.*

Hadley, G. A. Guttman, J. A., & Stringer, P. G. (1999, June). *Air traffic control specialist performance measurement database.* (Report No. DOT/FAA/CT-TN99/17). Atlantic City, NJ: William J. Hughes Technical Center.

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

Manning, C. A., Mills, S. H., Pfleiderer, E., Fox, C., & Mogilka, H. J. (2000). Relationships among observer ratings of air traffic controller performance and workload based on routinely recorded data. To be presented at 71st Annual Scientific Meeting of the Aerospace Medical Association, Houston, TX.

Mills, S. H. (2000). Performance and Objective Workload Evaluation Research (POWER.) Technical report in preparation.

Mills, S. H., Manning, C. A., & Pfleiderer, E. M. (1999, May). Computing en route baseline measures with POWER. Poster presented at Tenth International Symposium on Aviation Psychology, Columbus, OH.

Mogford, R. H., Murphy, E. D., Roske-Hofstrand, R. J., Yastrop, G., & Guttman, J. A. (1994, June). *Research techniques for documenting cognitive processes in air traffic control: Sector complexity and decision making* (Report No. DOT/FAA/CT-TN94/3). Atlantic City, NJ: Federal Aviation Administration Technical Center.

Rodgers, M. D., & Duke, D. A. (1993). SATORI: Situation Assessment Through Re-creation of Incidents. *The Journal of Air Traffic Control*, *35*(4), 10-14.

Rodgers, M. D., Mogford, R. H., & Mogford, L. S. (1998). *The relationship of sector characteristics to operational errors.* (Report No. DOT/FAA/AM-98-14). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Stein, E. S. (1985). *Air traffic controller workload: An examination of workload probe*. (Report No. DOT/FAA/CT-TN84/24). Atlantic City, NJ: Federal Aviation Administration Technical Center.

Wickens, C. D., Mavor, A. S., Parasuraman, R., McGee, J. P. (Eds). (1998). *The future of air traffic control: Human operators and automation* (pp. 216-217). Washington, DC: National Academy Press.

Carol A. Manning, Ph.D.
FAA Civil Aeromedical Institute
Human Factors Research Laboratory
Human Resources Research Division

Carol Manning is an Engineering Research Psychologist in the Human Factors Research Laboratory at the FAA's Civil Aeromedical Institute (CAMI), located in Oklahoma City, OK. Carol has a Ph.D. in Experimental Psychology from the University of Oklahoma (awarded in 1982), with an emphasis in Decision Theory. She has been with CAMI since 1983. Carol has conducted research on validation of Air Traffic Control Specialist (ATCS) selection procedures, evaluation of ATCS field training programs, and identification of aptitude requirements for ATCSs who will operate future automated systems. She has participated in a number of studies that investigated the use of flight progress strips in en route air traffic control. More recently, Carol was involved in developing criterion performance measures for validation of the AT-SAT selection battery. She is currently involved in a project to develop objective measures of ATCS taskload and performance using available System Analysis Report (SAR) data that will be used to evaluate the effectiveness of new ATC system concepts. She is also working on additional projects to identify possible replacements for flight progress strips.