DOT/FAA/AM-02/4

Office of Aerospace Medicine
Washington, DC 20591

Using Air Traffic Control Taskload Measures and Communication Events to Predict Subjective Workload

Carol A. Manning[1]
Scott H. Mills[2]
Cynthia M. Fox[1]
Elaine M. Pfleiderer[1]
Henry J. Mogilka[3]

[1]Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

[2]SBC Technology Resources, Inc.
Austin, TX 78759

[3]FAA Academy
Oklahoma City, OK 73125

April 2002

Final Report

This document is available to the public
through the National Technical Information
Service, Springfield, VA 22161.

U.S. Department
of Transportation

Federal Aviation
Administration

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents thereof.

Technical Report Documentation Page

1. Report No. DOT/FAA/AM-02/4	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle

Using Air Traffic Control Taskload Measures and Communication Events to Predict Subjective Workload

5. Report Date

April 2002

6. Performing Organization Code

7. Author(s)

Manning, C.A.[1], Mills, S.H.[2], Fox, C.[1], Pfleiderer, E.M.[1], and Mogilka, H.J.[3]

8. Performing Organization Report No.

9. Performing Organization Name and Address

[1]FAA Civil Aerospace Medical Institute P.O. Box 25082 Oklahoma City, OK 73125	[2]SBC Technology Resources, Inc. 9505 Arboretum Blvd. Austin, TX 78759	[3]FAA Academy, Air Traffic Division P.O. Box 25082 Oklahoma City, OK 73125

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

12. Sponsoring Agency name and Address

Office of Aerospace Medicine
Federal Aviation Administration
800 Independence Ave., S. W.
Washington, DC 20591

13. Type of Report and Period Covered

14. Sponsoring Agency Code

15. Supplemental Notes

Work was accomplished under approved subtask AM-B-01-HRR516.

16. Abstract

A study was conducted to determine whether air traffic control (ATC) communication events would predict subjective estimates of controller workload as well as measures of controller taskload. We compared different regression models' predictions of subjective workload estimates made by 16 subject matter experts on 5 occasions during 8 samples of air traffic activity. The predictors were different combinations of four taskload principal components computed from routinely recorded ATC data, two principal components representing the number and duration of voice communication events, and two principal components representing the content of voice communications. Several regression model comparisons were computed to identify "reduced" regression models containing fewer predictors that would predict the workload ratings as well as a full model containing all predictors. Several reduced models predicted ATWIT (Air Traffic Workload Input Technique) ratings as well as the full model but all of these contained the Activity component. These reduced models were a model containing only the Activity component, a model containing the Activity and Instructional Clearances components, and a model containing the Activity, Instructional Clearances, and All Communications Number and Duration components. The results suggest that routinely recorded ATC data provide a good estimate of subjective workload. However, if recordings of voice communications are available and researchers want to invest the time required to analyze the transcripts, they may be able to improve slightly their estimate of subjective workload. The researcher must consider whether the information gained is worth the additional time investment required for analysis.

17. Key Words Air Traffic Control, Workload, Taskload, Communications	18. Distribution Statement Document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 19	22. Price

Using Air Traffic Control Taskload Measures and Communication Events to Predict Subjective Workload

Introduction

Sensitive and valid workload measures are needed for en route air traffic control (ATC) to identify potential negative effects on controllers of using new forms of automation or ATC procedures (Wickens, Mavor, & McGee, 1997) and to ensure that intended benefits for controller productivity have been achieved. ATC workload can be influenced by many factors, including numbers and configurations of aircraft moving through a sector, the activities the controller performs to control those aircraft, and the controller's reaction to the air traffic situation.

Measures of ATC workload are typically based on controllers' subjective ratings, made either while controlling air traffic or just afterwards. One problem with using workload ratings obtained while controlling traffic is that their values may be influenced by the effort required to generate and record the ratings. On the other hand, workload ratings provided after the controller finishes controlling traffic may be influenced by extraneous factors such as remembering only events that occurred early or late in the traffic sample (e.g., due to proactive or retroactive inhibition).

Research is being conducted to develop objective workload estimates that could replace subjective workload ratings by computing variables from routinely recorded ATC data (Buckley, DeBaryshe, Hitchner, & Kohn, 1983; Galushka, Frederick, Mogford, & Krois, 1995; Mills, Pfleiderer, & Manning., 2002; Stager, Ho, & Garbutt, 2001). These taskload measures usually describe both aircraft and controller activities. It is desirable, from a research perspective, to use objective taskload measures rather than subjective workload ratings because it is often easier and less expensive to obtain access to routinely-recorded ATC data than it is to have air traffic controllers participate in research simulations. Another reason for using objective taskload measures instead of subjective measures is that they are not influenced by rater errors such as leniency/severity errors or errors of central tendency (Landry, 1989). Finally, computing objective measures from recorded data will not interfere with controllers' activities (thus, not affecting their perceived workload).

Although using objective measures has some obvious benefits, the argument has also been made that they do not provide a complete representation of ATC workload. While these measures capture variations in ATC activity, they cannot capture a controller's personal reaction to the air traffic situation (Stein, 1998). Stein contends that controllers' individual differences influence their perception of the effects of a particular taskload. Thus, subjective workload ratings are affected by a component that cannot be derived simply by analyzing recorded data. However, other research has found significant correlations between objective taskload and subjective workload measures (Stein, 1985; Manning, Mills, Fox, Pfleiderer, & Mogilka, 2001), suggesting that using taskload measures alone may provide sufficient information to evaluate the effects of new systems.

Communications between pilots and controllers and between controllers and other controllers are also recorded routinely and so may be included among a set of objective taskload measures. Communication events can include counts of the number of communications, the timing with which these events occur, and the content of the communications.

Pilot-controller communications are thought to be related to ATC workload because complicated communications can increase workload (Morrow and Rodvold, 1998). Bruce (1993) found that both traffic volume and traffic complexity (both frequently used indicators of ATC taskload) were significantly related to the number of pilot/controller communications. Cardosi (1993) examined numbers and timing of communication events as a function of message type in a descriptive study that analyzed timing of voice communications related to traffic avoidance. As part of this study, she used numbers of communications per hour to classify time periods as high- or low-workload.

Corker, Gore, Fleming, and Lane (2000) used communication time as an indicator of workload against which to assess alternative free flight conditions. Porterfield (1997) examined the relationship between the amount of time spent communicating and on-line workload ratings. He found a correlation of r = .88 and concluded that controller communication duration is a valid measure of workload.

Besides indicating the amount of activity, one advantage of using communication events as an indicator of workload is that their content and associated affective components may indicate the amount of effort the controller experienced at the time the event occurred. Thus, these measures may contribute at least part of the subjective component of workload that Stein (1998) argues is not accounted for by other taskload measures. Moreover, analyzing recorded voice communications does not interfere with the controller's task.

On the other hand, there are some disadvantages associated with the use of communications measures. First, determining the number and duration of communication events requires a considerable amount of time and manual labor, and coding their content and affect requires even more effort. Thus, the use of communication events would seem to be inconsistent with the goal of obtaining an easily-computed set of taskload measures, unless they add significantly to the prediction of subjective workload. Second, the timing of recorded communication events does not account for all communication activity because some exchanges (e.g., radar [R] controller to data [D] controller communications) are not recorded. Thus, analysis of recorded communications will provide, at best, a lower-bound estimate for subjective workload.

Previous research suggests that certain communications measures, such as number and duration, are associated with workload. However, a related question that must be answered is whether distinguishing between pilot/controller and controller/controller communications or coding the content of communications will add a unique component to the prediction of subjective workload over and above that contributed by other types of objectively measured controller and aircraft activities. If counts and durations of communication events measure something different than ATC taskload measures, as evidenced by low correlations between the variables, and they contribute a unique component to the prediction of subjective workload, then it would be useful to expend the effort required to obtain and analyze them. If, on the other hand, communication events are highly correlated with other objectively-measured ATC activities and subjective workload, then they will contribute little unique variance to the prediction of subjective workload, and expending the effort required to extract them would be of little value. Given the results of research that suggest that communication events are related to taskload, we expect that the communication events measured here will be so highly correlated with our taskload measures that they will not make a unique contribution to the prediction of subjective workload.

The purpose of this study was to examine the relationship between communication events, subjective workload, and objective taskload measures. The communication events analyzed were total number of communications, total time spent communicating, average time spent for an individual communication, and communication content. The number of communication events and time spent communicating were analyzed separately for each speaker (controller, other). The number and timing of a controller's communications should be related to subjective workload, but having to attend to other speakers could also be a component of workload.

We proposed several hypotheses about the relationships between these measures. First, we expected that the total number and duration of communication events would be significantly related to busyness—as measured both by subjective workload and objective taskload measures. As the traffic situation gets busier, more communication events should occur, and more time should be spent communicating, both by the controller and other speakers.

Second, we expected that the average time for an individual communication event should be negatively related to both workload and taskload. As the traffic situation gets busier, the amount of time spent on a single communication should decline. The time spent on an individual communication event is likely to be related to the identity of the speaker; that is, controllers are likely to reduce the amount of time they spend on an individual communication while other speakers are unlikely to be as affected by activity occurring in the sector.

Third, we expected that the content of communication events may be related to sector activity. As the situation gets busier, there should be more clearances issued, readbacks, and pilot requests. However, the number of advisories or unrelated remarks may not change. We also expected that the number of clearances issued to pilots should be related to subjective workload, while radio frequency changes issued should be unrelated.

Fourth, if communication events are significantly related to sector activity, we expected that they would not contribute uniquely to the prediction of subjective workload, over and above the contribution of the taskload measures. Thus, we expected that taskload measures alone would account for most of the variance in a set of subjective workload ratings and this prediction would not improve by adding communication measures to the set of predictors.

If any measures derived from communication events do indeed add a unique component to the prediction of subjective workload, then it would be worth taking the time to analyze the transmissions. On the other hand, if they do not add a unique component to the prediction of subjective workload, it would not be necessary to analyze them.

Method

This study examined statistical relationships between communication events, objective taskload measures, and subjective workload measures. The communication events and taskload measures were obtained from samples of routinely-recorded ATC data. The workload measures were provided by subject matter experts (SMEs) who observed graphical displays of the same ATC data samples (hereafter called "traffic samples") and rated the workload they thought the R controller responsible for the sector had experienced. Each component of the study is discussed in more detail below.

Traffic Samples

System Analysis Report (SAR) data and voice communication tapes were obtained for 12 traffic samples recorded during January, 1999, at four sectors (sectors 14, 30, 52, and 54) in the Kansas City Air Route Traffic Control Center (ARTCC). The ATC data were extracted by the Data Analysis and Reduction Tool (DART; Federal Aviation Administration, 1993) and the National Track Analysis Program (NTAP; Federal Aviation Administration, 1991). The resulting files were processed both by the Systematic Air Traffic Operations Research Initiative (SATORI; Rodgers & Duke, 1993) and Performance and Objective Workload Evaluation Research (POWER; Mills, Pfleiderer, & Manning, 2002) software programs. SATORI synchronizes extracted data from DART and NTAP files with tapes containing the R controller's voice communications, using the time code common to both data sources, while POWER uses data from a subset of the DART files to compute taskload measures. Three traffic samples were re-created for each sector. One traffic sample (used to train the SMEs) was eight minutes long. The two remaining experimental traffic samples were both 20 minutes long.

Participants

Participants were 16 en route air traffic control instructors from the FAA Academy in Oklahoma City, OK. All had formerly been fully-certified controllers at en route centers. Two participants had controlled traffic at some of the Kansas City Center sectors included in the traffic samples, though none had worked all the sectors included in the study.

Sector training materials

Computerized training sessions were provided that described the characteristics and applicable procedures for each sector. Participants examined copies of sector maps and the sector binder (which contained additional information about the sector). Participants could also examine flight plan information (derived from recorded flight strip messages) for each aircraft controlled by the sector during the traffic sample.

Subjective workload

Participants provided a subjective workload assessment using the Air Traffic Workload Input Technique (ATWIT; Stein, 1985). The ATWIT measured mental workload in "real-time" by presenting auditory and visual cues that prompted a controller to press one of seven buttons within a specified amount of time to indicate the amount of mental workload experienced at that moment. In this study, instead of rating their own workload, the participants entered ATWIT ratings to indicate the amount of mental workload they thought the R controller experienced in reaction to the taskload that occurred during the traffic sample. The participants were prompted every four minutes during each traffic sample to provide ATWIT ratings.

Objective taskload measures

The objective taskload measures used in this study were derived from the POWER software (Mills, Pfleiderer, & Manning, 2002). The POWER measures included information about the number of controlled aircraft; handoffs made and accepted; altitude changes; controller data entries and data entry errors; variations in aircraft headings, speeds, and altitudes; and the average time aircraft were under control. In all, 23 POWER measures were utilized in this study.

Communication events

Communication events were obtained from voice tapes associated with the traffic samples. The measures analyzed in this study were the total number of communications, total time spent communicating during a traffic sample, and time required for individual communication events (for all speakers, and analyzed separately for the controller and other speakers).

The communication events were also categorized by content. The content categories were based on a set derived by Prinzo, Britton & Hendrix (1995), and consisted of 1) Address, 2) Courtesy, 3) Clearance, 4) Advisory/Remark, 5) Request, 6) Readback/Acknowledgment, and 7) Non-codable. The clearance category was then divided into two sub-categories, Instructional Clearances and Frequency Changes, to distinguish between clearances instructing an aircraft to proceed and more routine instructions for the pilot to change the radio frequency when leaving the sector. Communications were not otherwise separated into specific message types (e.g., altitude or heading clearance) or coded as errors (e.g., transposed numbers/letters) in order to retain sufficient numbers for analysis.

Procedure

Participants reviewed a description of the study, completed a consent and a biographical information form, then reviewed instructions for making the ATWIT workload assessments, as well as two other types of post-scenario subjective workload assessments not analyzed in this study. 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 concurrently as a block. The order of observing the four blocks of traffic samples (corresponding to the four sectors) was counter-balanced, as was the order of presentation of the two experimental traffic samples within each block.

During each traffic sample, participants recorded any mistakes using a behavioral observation form (see Manning et al., 2001, for more details). The ATWIT aural prompt occurred every four minutes, and participants responded by entering a number between 1 and 7 on a keypad. At the end of each traffic sample, participants completed the other subjective workload assessments, summed errors they had recorded, then completed an over-the-shoulder performance rating form (see Manning et al., 2001, for more details). Completing the training process and observing the three traffic samples for each sector required about 1½ hours.

Data processing

Communication events during each traffic sample were transcribed. Message contents of each transmission were categorized, along with the identity of the speaker (i.e., controller, pilot, other speaker) and the start and stop times. These data were used to compute the total number of communications and time spent communicating during each 4-minute period, as well as the mean time for individual communication events and their contents.

The 23 POWER measures were computed for the five 4-minute segments included in each experimental traffic sample. The other two workload assessments were not analyzed in this study because they were only obtained at the end of each traffic sample; thus, only eight observations (one for each traffic sample) were available for analysis.

The ATWIT ratings were averaged across participants for each 4-minute segment included in each experimental traffic sample, resulting in 40 observations.

Results

Subjective workload. ATWIT ratings, when averaged across 4-minute time periods within the eight traffic samples, ranged between 2.01 and 3.54. The mean ATWIT rating across the eight traffic samples was 2.76 (SD = .59). This value is significantly lower than 4 ($t(39) = -13.2$, $p < .001$), the mid-point of the 7-point workload scale, suggesting that participants thought that workload was generally low during the traffic samples.

Communication events. Nine hundred ninety-nine communication events (or, on the average, about 125 per traffic sample) were recorded during the eight traffic samples. Four hundred seventy-one of these (47%) were made by a controller, and 528 (53%) were made by another speaker (pilot or other controller.) The average number of communication events for a 4-minute period was 25.0 (SD = 10.0). Controllers made, on the average, 11.8 (SD = 4.8) of the communications, while other speakers made 13.2 (SD = 5.4).

On the average, the total amount of time spent communicating during a 4-minute period was 69.18 seconds (SD = 25.0), or about 29% of the 240 available seconds. Controllers spent, on the average, 38.3 seconds (SD = 15.3) speaking during each 4-minute period, while others spoke for an average of 30.9 seconds (SD = 12.1).

The average duration of a single communication event was 2.86 seconds (SD = 0.63). Single communication events for controllers lasted, on the average, 3.38 seconds (SD = 0.95), while single communication events for other speakers lasted, on the average, 2.41 seconds (SD = 0.60).

The average number of communication events by content is shown in Table 1. Because each transmission may have included more than one topic of conversation, each communication event may include

Table 1. Descriptive statistics for communication event content categories.

Content of communication events	Mean	SD
Address	15.1	5.4
Courtesy	4.4	2.6
Advisory	5.2	3.2
Request	2.6	2.3
Readback	9.9	4.6
Instructional clearances	3.8	2.1
Frequency changes	2.4	1.8

more than one content category. Thus, the number of times a content category was addressed in a 4-minute time period was greater than the number of communication events that occurred.

Addresses occurred most frequently, on the average, about 15 times in a 4-minute period. Readbacks occurred about 10 times per period. Requests, instructional clearances, and frequency changes occurred least often. Non-codable communications were not reported here and were excluded from all subsequent analyses.

Table 2 shows inter-correlations between the communication events computed for the 4-minute periods. Total times and numbers of communications, for both controllers and all other speakers, were highly correlated with each other. The average times for individual communication events were significantly correlated with each other and with the total number and timing of communication events (with a negative valence), but the correlations were not very high. The number of Addresses was significantly correlated with all other content categories, but that was not true of any other content category. Readbacks had high correlations with Addresses and Advisories, and were related to all other content categories except Frequency Changes. Frequency Changes were only significantly related to Addresses and Courtesies.

While interesting, the pattern of correlations was difficult to interpret, so two Principal Components Analyses (PCAs) with Varimax rotation were conducted to identify a smaller set of components that would describe the relationships between the communication events more concisely. The first PCA included only the variables describing the number and duration of communication events. The second PCA included only the content categories for the communication events. We decided that because counts and timing of communication events were sufficiently different from their content, separate PCAs were warranted.

The first PCA, which included variables describing the number and duration of communication events, produced two components with eigenvalues greater than 1. The two components accounted for 81.6% of the variability in the data set. The rotated component matrix is shown in Table 3. The entries in the table are correlations between each communication measure and the two components derived from the analysis. For ease of interpretation, correlations less than .3 were excluded from the table.

The number and duration of all communications that occurred during the 4-minute period had high correlations with component 1 and, thus, it was labeled *All Communications Number and Duration*. The mean time for an individual communication event, both for controllers and other speakers, had the highest correlations with component 2, although total communication time for controllers was positively correlated and number of communications by other speakers was negatively correlated. Thus, component 2 was labeled *Individual Communication Duration.*

The second PCA, which included variables describing the communication content categories, produced three components with eigenvalues greater than 1. These components accounted for 84.2% of the variability in the data set. The rotated component matrix is shown in Table 4. For ease of interpretation, correlations less than .3 were excluded from the table.

Both Requests and Advisories had the highest correlation with component 1, although Readbacks and Addresses were also correlated. Thus, component 1 was labeled *Requests andAdvisories*. Frequency Changes and Courtesies had the highest correlation with component 2, although Addresses and Readbacks were also correlated. Thus, component 2 was labeled *Frequency Changes/Courtesies.* Instructional clearances had the highest correlation with component 3, although addresses, advisories, and readbacks were also correlated. Thus, component 3 was labeled *Instructional Clearances*.

Table 2. Intercorrelations of communication events.

	N comms - Cont	N comms - Other	Total comm time - Cont	Total comm time - Other	Mean time single comm - Cont	Mean time single comm - Other	Addr	Crtsy	Advsry	Reqst	Rdbck	Instruct clrnce	Freq change
N comms - Cont	1.00												
N comms - Other	.89**	1.00											
Total comm time - Cont	.72**	.60**	1.00										
Total comm time - Other	.83**	.83**	.67**	1.00									
Mean time single comm – Cont	- .33*	- .34*	.37**	- .18	1.00								
Mean time single comm - Other	- .14	- .29	.06	.27	.32	1.00							
Address	.84**	.78**	.71**	.78**	-.19	- .06	1.00						
Courtesy	.47**	.41**	.22	.31	-.33*	- .24	.44**	1.00					
Advisory	.78**	.82**	.62**	.67**	- .11	- .21	.61**	.11	1.00				
Request	.62**	.63**	.27	.39*	- .39*	- .38*	.38*	.15	.58**	1.00			
Readback	.89**	.83**	.70**	.82**	- .23	.00	.78**	.43**	.70**	.49**	1.00		
Instructional Clearance	.39*	.38	.50**	.43**	.12	.12	.43**	.06	.34*	.04	.55**	1.00	
Frequency change	.38*	.25	.22	.29	.25	-.04	.43**	.59**	-.04	- .10	.30	- .13	1.00

** Correlation is significant at p <.01 level

* Correlation is significant at p <.05 level

Table 3. Rotated component matrix for 2 principal components representing number and duration of communication events.

Communication number and duration measures	Comp 1: All Communications Number and Duration	Comp 2: Individual Communication Duration
Total N comms – controller	.95	
Total N comms – other speaker	.90	-.35
Total comm time - controller	.83	.38
Total comm time – other speaker	.93	
Mean time individual comm – controller		.87
Mean time individual comm – other speaker		.74

*Correlations less than .3 are not displayed.

Table 4. Rotated component matrix for 3 components representing communication content categories.

Communication content measures	Comp 1: *Requests/ Advisories*	Comp 2: *Frequency Changes/ Courtesies*	Comp 3: *Instructional Clearances*
Address	.50	.53	.53
Courtesy		.84	
Advisory	.81		.38
Request	.93		
Readback	.60	.40	.60
Instructional clearance		.92	
Frequency change			.96

*Correlations less than .3 are not displayed.

Taskload measures. Table 5 shows descriptive statistics for the 23 POWER measures averaged across the 4-minute periods in each traffic sample. Some of the POWER measures (primarily certain kinds of data entries, such as handoffs and altitude changes) occurred several times during the 4-minute periods. However, many of the other data entries (e.g., pointouts, data block offsets, Distance Reference Indicators [DRIs, also known as J-rings], track reroutes) and the conflict alerts (both displayed and suppressed) occurred very infrequently. In fact, many of the variables occurred in fewer than 30% of the time segments, resulting in near-zero means and corresponding standard deviations that were greater than the means. Subsequent analyses excluded these infrequent variables.

Moreover, two variables (R controller data entries and D controller data entries) were a compilation of all subsets of specific data entries (such as Data Block Offsets, Route Displays, R and D controller Pointouts, DRIs requested and deleted, and altitude changes). If all specific data entries were summed, they would total the values of the R and D controller data entries. It is not appropriate to analyze both individual measures and a variable that comprises their sum, so for the purpose of this study, the individual measures were excluded from further analysis. However, the average time to accept a handoff and average time until initiated HOs are accepted were retained for analysis because they are independent of the number of handoffs made and accepted.

To reduce the number of POWER measures by grouping similar variables, correlations between the measures were first computed. These are shown in Table 6. Significant correlations were observed between a number of the variables. However, visual examination of the correlations did not provide a systematic method for interpreting the relationships between variables. A PCA, with Varimax rotation, was conducted to identify a smaller set of components that would describe the relationships between the POWER measures more concisely. Four components were produced with eigenvalues greater than 1 that accounted for 71.2% of the variance in the data.

Table 5. Descriptive statistics for POWER measures obtained at 4-minute intervals.

	Descriptive statistics	
Power Measures	Mean	*SD*
Total N aircraft controlled	7.20	2.73
Max aircraft controlled simultaneously	5.48	2.35
Average time aircraft under control	158.35	34.38
Avg Heading variation	1.06	0.86
Avg Speed variation	4.22	2.46
Avg Altitude variation	2.00	1.48
* Total N altitude changes	3.50	2.20
* Total N handoffs accepted	1.15	1.12
Avg time to accept handoff	25.91	27.58
* Total N handoffs initiated	1.98	1.29
Avg time until initiated HOs are accepted	41.00	45.45
N Radar controller data entries	11.35	5.54
N Radar controller data entry errors	0.23	0.58
N Data controller data entries	1.93	2.04
N Data controller data entry errors	0.08	0.27
* N Route displays	0.40	0.84
* N Radar controller pointouts	0.08	0.27
* N Data controller pointouts	0.08	0.47
* N data block offsets	0.15	0.43
* Total N CAs displayed	0.08	0.27
* Number of CA suppression entries	0.05	0.22
* N DRIs requested	0.05	0.22
* N DRIs deleted	0.03	0.16

Note: * indicates variables excluded from further analysis.

Table 6. Correlations between POWER measures.

	N aircraft	Max aircraft	Hdg Var	Spd Var	Alt Var	HO acc dur	HO ini dur	Avg time contr	R entries	R errors	D entries	D errors
N aircraft	1.00											
Max aircraft	.88**	1.00										
Hdg Variation	.05	.03	1.00									
Spd Variation	- .23	- .13	.73**	1.00								
Alt Variation	- .15	- .03	.37**	.63**	1 .00							
HO acc dur	.38**	.33*	-.18	- .30*	-.14	1.00						
HO ini dur	- .13	- .08	.16	.26	.04	- .13	1.00					
Avg time controlled	.26	.58**	.21	.40**	.41**	.00	.12	1.00				
R entries	.65**	.57**	-.01	- .32*	-.11	.21	- .16	.08	1.00			
R errors	.00	.05	-.09	- .04	-.04	-.20	.55**	.17	.09	1.00		
D entries	.09	- .09	.07	- .04	-.08	.00	- .27*	-.18	- .04	- .31*	1.00	
D errors	- .02	- .14	-.12	- .14	-.27*	.01	- .23	-.29*	-.07	- .11	.39**	1.00

** Correlation is significant at *p* <.01 level
* Correlation is significant at *p* <.05 level

The results of this analysis should be interpreted with some caution because 1) only 4-minute time segments were analyzed, and 2) only 40 observations from 4 sectors were available for analysis. Subsequent analyses using larger data sets should be conducted to obtain more stable results. However, the primary purpose of conducting this analysis was to derive a smaller number of variables to be used in later analyses. Table 7 contains the rotated component matrix for the 4 components. For ease of interpretation, correlations less than .3 were excluded.

Component 1 was primarily related to the number of aircraft controlled, those controlled simultaneously, and R controller data entries. To a lesser extent, the component was also related to the time to accept handoffs and control duration. Component 1 was labeled *Activity* because higher values for these measures were associated with the presence of more aircraft in a sector that required more controller activity.

Component 2 was related to variation in heading, speed, and altitude, and, to a lesser extent, control duration. Component 2 was labeled *Low Altitude Maneuvers* because these measures were related to aircraft maneuvering consistent with arrivals at and departures from low altitude sectors surrounding the St. Louis Lambert Airport. This interpretation is supported by a comparison of average heading, speed, and altitude variability, which were all significantly higher in low altitude sectors than in high altitude sectors ($t(38) = 2.82$, 5.75, and 3.49, respectively; $p < .01$ in all three cases).

Component 3 was primarily related to R controller data entry errors and the time required to accept initiated handoffs. To a lesser extent, it was also negatively related to time to accept handoffs. These conditions were consistent with busy R controllers making more data entry errors, having to attend to whether the next controller had accepted handoffs that he/she had initiated, and taking longer to accept aircraft handed off to his/her sector. Thus, Component 3 was called *Overload.*

Component 4 was primarily related to D controller data entries and D controller data entry errors. To a lesser extent, the component was also related to lower altitude variation and lower control duration. While the number of D controller data entries and errors were not related to the number of aircraft in the sector, the presence of aircraft that changed altitude less frequently and were in the sector for a shorter period of time was related to a higher number of D errors. Thus, Component 4 was called *D Activity*.

Prediction of mental workload

Correlations. Table 8 shows correlations between the ATWIT subjective workload rating, the four objective workload components, the two components related to number and duration of communication events, and the three communication content components. By definition, the principal components are unrelated, so their inter-correlations are 0. The mean ATWIT rating had a correlation of .80 *($p < .01$)* with *Activity*, a correlation of .62 ($p < .01$) with *All*

Table 7. Rotated component matrix for 4 components representing reduced set of POWER measures.

Power Measures	Comp 1: Activity	Comp 2: Low Alt Maneuvers	Comp 3: Overload	Comp 4: D activity
Max aircraft controlled simultaneously	.94			
Total N aircraft controlled	.94			
Avg Heading variation		.81		
Avg Speed variation		.92		
Avg Altitude variation		.73		-.33
Avg time to accept handoff	.40		-.40	
Avg time until initiated HOs are accepted			.80	
Avg time aircraft under control	.44	.54		-.39
N Radar controller data entries	.76			
N Radar controller data entry errors			.87	
N Data controller data entries				.76
N Data controller data entry errors				.77

*Correlations less than .3 are not displayed.

Table 8. Correlations of subjective workload rating, taskload components, communication time and duration components, and communication content components.

	ATWIT	Activity	Low Alt Man.	Overload	D Activity	All Comm N, Duration	Individual Comm. Duration	Requests/ Advisories	Frequency Changes/ Courtesies	Instructional Clearances
Subjective workload										
ATWIT	1.00									
Taskload components										
Activity	.80**	1.00								
Low Alt Man.	.15	.00	1.00							
Overload	.02	.00	.00	1.00						
D Activity	-.02	.00	.00	.00	1.00					
Communication N and duration components										
All Comm N, Duration	.62**	.63**	-.02	.27	-.23	1.00				
Individual Comm Duration	.36*	.21	.20	-.09	.30	.00	1.00			
Communication content components										
Requests/ Advisories	.13	.24	-.02	.30	-.27	.65**	-.38	1.00		
Frequency Changes/ Courtesies	.20	.36*	-.28	.01	-.07	.39*	-.22	.00	1.00	
Instructional Clearances	.65**	.52**	.23	.23	-.12	.54**	.28	.00	.00	1.00

** Correlation is significant at the $p < 0.01$ level. * Correlation is significant at the $p < 0.05$ level.

Communications Number and Duration, a correlation of .65 (p < .01) with *Instructional Clearances*, and a correlation of .36 (p < .05) with *Individual Communication Duration*.

Activity was significantly correlated with *All Communications Number and Duration* (r = .63, p < .01), *Clearances* (r = .52, p < .01), and with *Frequency Changes/Courtesies* (r = .36, p < .05). The *Overload, Low Altitude Maneuvers*, and *D Activity* components were not significantly correlated with any of the other variables. *All Communications Number and Duration* was significantly correlated with all three content components: For *Requests/Advisories*, r = .65, p < .01; for *Frequency Changes/Courtesies*, r = .39, p < .05; and for *Instructional Clearances*, r = .54, p < .01. *Individual Communication Duration* was not significantly correlated with any of the content components.

Regression. A set of analyses was performed to assess the effectiveness of alternative multiple regression models in predicting the subjective ATWIT ratings. Table 9 shows the results of these analyses. Row 1 shows the multiple correlation of the full model containing the 4 taskload, 2 communication number and duration, and 3 communication context components as predictors. The multiple correlation of the full model with the ATWIT ratings was R = .88, accounting for about 77% of the variance in the subjective workload ratings. Succeeding lines show multiple correlations between alternative (reduced) regression models containing fewer than the total number of predictors. The column containing F for the test of R^2 change compares the relative effectiveness of a reduced model with the effectiveness of the full model in predicting the ATWIT rating. If the probability is greater than .05 that the change in R^2 between the two models is significantly different from 0, then the reduced model is considered to be as effective as the full model in predicting subjective workload. On the other hand, if the probability is less than or equal to .05 that the change in R^2 between the

Table 9. Results of analyses comparing alternative multiple regression models predicting ATWIT ratings.

Regression model	R	R^2	R^2 change	F for test of R^2 change	df	p
1. Full model containing all taskload, communication number and duration, and communication context components	0.88	0.77	N/A	N/A		
Reduced models based on Taskload components						
2. Model containing all taskload components	0.82	0.67	0.11	2.80	5, 30	.034
3. Model containing only the *Activity* component	0.80	0.65	0.13	2.13	8, 30	.064*
4. Model containing all taskload components except *Activity*	0.15	0.02	0.75	16.66	6, 30	.000
Reduced models based on communications components						
5. Model containing five communication components	0.78	0.60	0.17	5.67	4, 30	.002
6. Model containing two communication number and duration components	0.72	0.52	0.25	4.78	7, 30	.001
7. Model containing three communication context components	0.69	0.48	0.41	6.49	6, 30	.000
8. Model containing *All Communications Number and Duration* and *Instructional Clearances* components	0.73	0.53	0.25	4.67	7, 30	.001
9. Model containing *Instructional Clearances* component	0.65	0.43	0.35	5.81	8, 30	.000
Reduced model combining taskload and communications components						
10. Model containing *Activity*, *All Communications Number and Duration,* and *Instructional Clearances* components	0.85	0.72	0.05	1.11	6, 30	.378*
11. Model containing *Activity* and *Instructional Clearances* components	0.85	0.72	0.05	1.02	7, 30	.439*

* Indicates reduced models that predicted ATWIT ratings as well as the full model.

two models is significantly different from 0, then the reduced model is not considered to be as effective as the full model in predicting subjective workload. The goal is to identify a reduced model that contains as few predictors as possible, but accounts for a high enough percentage of the variance in the dependent variable to be considered equivalent to the full model.

The analysis of 10 reduced models is shown in Table 9 (see rows 2-11). The first group of analyses (rows 2-4) compared reduced models containing different combinations of taskload components with the full model. The second group of analyses (rows 5-9) compared reduced models containing different combinations of communication components with the full model. The final group of analyses (rows 9-11) compared reduced models containing combinations of both taskload and communications components with the full model.

As an example, row 2 compared a reduced model containing all the taskload components with the full model. The model containing all the taskload components had an R^2 of .67, compared with the full model's R^2 of .77. The F computed to assess the R^2 change of .11 had a value of 2.80, and the probability was .034 that the change in R^2 was greater than 0. Thus, the reduced model containing all the taskload components was significantly different than the full model in predicting ATWIT ratings and, thus, was not as effective as the full model.

A second example is shown on line 3, which compared a reduced model containing only the *Activity* taskload component with the full model. The model containing only the *Activity* component had an R^2 of .65, compared with the full model's R^2 of .77. The F computed to assess the R^2 change of .13 had a value of 2.13, and the probability was .064 that the change in R^2 was greater than 0. Thus, using an alpha level of .05, the reduced model containing only the *Activity* component predicted ATWIT ratings as well as the full model.

A third example is shown on line 5. The reduced model containing all the taskload components *except* the *Activity* component had an R^2 of .02, compared with the full model's R^2 of .77. The F computed to compare the R^2 change of .75 had a value of 16.66, and the probability was less than .0001 that the change in R^2 was greater than 0. Thus, the reduced model containing all the Taskload components except the *Activity* component did not predict the ATWIT ratings as well as the full model.

Four of the reduced models shown in Table 9 (one containing the *Activity* component alone, one containing the *Activity*, *All Communications Number and Duration*, and *Instructional Clearances* components, and one containing the *Activity* and *Instructional Clearances* components) predicted ATWIT ratings as well as the full model. Thus, for a reduced model to be equivalent to the full model, it must contain the *Activity* component. None of the reduced models containing any combination of the communications components were equivalent to the full model unless they contained the *Activity* component.

Discussion and Conclusions

We formed several hypotheses about the relationships between the communications variables, objective taskload variables, and subjective workload. These were:

1. Total number and duration of communication events will have a significant and positive correlation with workload and taskload.
2. Average time for an individual communication event should be negatively related to workload and taskload.
3. The content of communication events may be related to sector activity.
4. Communication events will not provide a unique contribution to the prediction of subjective workload, over and above the prediction contributed by the taskload measures.

Before conducting the analyses, we derived sets of independent principal components to reduce the number of variables analyzed to a manageable set, given the number of observations in the data set. Thus, the analyses that tested the hypotheses were based on components consisting of a weighted combination of communication and taskload variables instead of the individual variables. Examination of Table 8 shows that four components, *Activity*, *All Communications Number and Duration*, *Individual Communications Duration*, and *Instructional Clearances*, had significant correlations with the ATWIT ratings. Thus, certain aspects of taskload, the number and duration of communication events, and the content of communications are all related to subjective workload.

Table 8 also shows that the *All Communications Number and Duration, Frequency Changes/Courtesies, and Instructional Clearances* components were significantly correlated with the *Activity* component. Thus, communication activity is related to taskload, especially clearances involving instructions to proceed.

Our prediction about the duration of individual communications was found to be only partially accurate. While the *Individual Communications Duration* component was significantly related to the ATWIT rating, it was not significantly correlated with any of the taskload components. Moreover, the principal components analysis did not produce different components for different speakers, suggesting that the identity of the speaker who generated the communication events was not important.

We did, however, find that the content of communications was significantly related to certain types of sector activity. *Instructional Clearances* and *Frequency Changes/ Courtesies* were significantly correlated with *Activity*.

Because we expected communications variables to overlap extensively with the taskload variables, we hypothesized that variables measuring communication events would not contribute uniquely to the prediction of subjective workload, over and above the prediction contributed by the taskload measures. Table 9 compared the effectiveness of a number of reduced regression models containing different combinations of taskload and communications variables in predicting subjective workload, as compared with the effectiveness of the full model containing all the variables. The full model accounted for 77% of the variance in subjective workload while the reduced models accounted for anywhere from 2% to 72% of the variance. Three reduced models were as effective (statistically) as the full model in predicting subjective workload. The reduced model containing *Activity* accounted for 65% of the variance in subjective workload. The reduced model containing *Activity, All Communications Number and Duration* and *Instructional Clearances* accounted for 72% of the variance in subjective workload. The reduced model containing *Activity*, *All Communications Number and Duration*, and *Instructional Clearances* also accounted for 72% of the variance in subjective workload.

A model containing all the taskload components except *Activity* predicted subjective workload very poorly, as compared with the effectiveness of the full model. Furthermore, none of the models containing only a combination of communications variables predicted subjective workload as well as the full model. For example, a reduced model containing all communications components accounted for only 60% of the variance in subjective workload, a model containing the two Communications number and duration components accounted for only 52% of the variance, and a model containing the three Communications context components accounted for only 48% of the variance.

An interesting finding from this analysis was that *Activity* must be present in order for a reduced regression model to predict ATWIT ratings as well as the full model. This result suggests that variables whose values as a function of increased air traffic activity (such as the number of aircraft, data entries, control duration, time to accept handoffs, etc.) have an important effect on the controllers' perception of workload.

The question that must be answered is whether the inclusion of any communications measures added a unique component to the prediction of subjective workload over and above the contribution of taskload. According to the analysis, *Activity* alone was statistically equivalent to the full model accounting for 65% of the variance in subjective workload,. However, adding the *Instructional Clearances* component produced a reduced model that contained only two variables and accounted for 72% of the variance in subjective workload. While *Activity* alone seems to be a good predictor of subjective workload, the combination of *Activity* and *Instructional Clearances* is slightly better.

Thus, these data suggest that those who only have access to SAR files will be able to derive a very good estimate of subjective workload using controller activity data. However, those who have access to both SAR files and recordings of communication events and want to invest the time required to analyze the transcripts may be able to obtain a better estimate of subjective workload. The question is whether the information gained is worth the additional time investment. And while it appears that it is not necessary to analyze voice communications data to assess controller workload adequately, the analysis of communications data is often valuable for other reasons.

The constraints associated with this study should be considered when interpreting these results. First, the limited selection of sectors (only four, all at one center, and all surrounding a busy airport) and traffic samples (two per sector) limit the ability to generalize these results to other sector types, traffic situations, and facilities. Second, the number of observations included in the analysis limited our confidence in the results. Third, SMEs provided subjective ratings of the workload they thought other controllers were experiencing instead of rating their own workload. If those who worked the traffic had rated their own workload, the results might have been different. Fourth, we assumed that the ATWIT was the most appropriate method for measuring subjective workload. If other workload measures were obtained, such as the NASA TLX (Hart & Staveland, 1988) or other physiological methods, the results might have been different. These other methods may measure different aspects of workload (because the TLX is a post-hoc method obtained only once, after the traffic sample, and physiological measures may be influenced by factors other than subjective workload.)

Fifth, the workload experienced during all the traffic samples was fairly low, as assessed by the SMEs. Perhaps the effectiveness of communications measures would have been more pronounced if a higher workload had been experienced.

Even if the controller/pilot communications variables had been found to provide a larger contribution to the prediction of subjective workload, this relationship might be expected to change soon. Controller/Pilot Data Link Communications (CPDLC), will transmit some pilot/controller communications via a digital channel, thus increasing the visual processing and reducing the auditory processing of communications. It has been proposed that using CPDLC will reduce controller workload, but more likely it will only change the distribution of workload from both visual and aural to a primarily visual modality. Adding visual tasks to an already extensive set of tasks currently performed by controllers might increase overall workload more than would be compensated for by reducing the number of voice communications to which the controller must attend. However, regardless of the effect on workload of increasing the visual component of a controller's activity, the workload associated with verbal communications should be significantly reduced when most are transferred to another information source.

References

Bruce, D.S. (1993). An explanatory model for influences of air traffic control task parameters on controller work pressure. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*, pp. 108-112.

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.

Cardosi, K. (1993). Time required for transmission of time-critical air traffic control messages in an en route environment. *International Journal of Aviation Psychology*, *7*, 171-182.

Corker, K.M., Gore, B.F., Fleming, K., & Lane, J. (2000, June). *Free flight and the context of control: Experiments and modeling to determine the impact of distributed air-ground air traffic management on safety and procedures*. In Proceedings of 3rd USA/Europe Air Traffic Management R&D Seminar, Napoli, Italy.

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.

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.

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.

Landry, F.J. (1989). *Psychology of work behavior* (pp. 131-133). Pacific Grove, CA: Brooks/Cole Publishing Company.

Manning, C.A., Mills, S.H., Fox, C., Pfleiderer, E., & Mogilka, H.J. (2001). *Investigating the validity of Performance and Objective Workload Evaluation Research (POWER).* (Report No. DOT/FAA/AM-01/10). Washington, DC: FAA Office of Aerospace Medicine.

Mills, S.H., Pfleiderer, E.M., & Manning, C.A. (2002). *POWER: Objective Activity and Taskload Assessment in En Route Air Traffic Control.* (Report No. DOT/FAA/AM-02/2). Washington, DC: FAA Office of Aerospace Medicine

Morrow, D. & Rodvold, M. (1998). Communication issues in air traffic control. In M. W. Smolensky & E. S. Stein (Eds.) *Human Factors in Air Traffic Control* (pp. 421-456). San Diego, CA: Academic Press.

Porterfield, D.H. (1997). Evaluating controller communication time as a measure of workload. *International Journal of Aviation Psychology*, *7*, 171-182.

Prinzo, O.V., Britton, T.W., & Hendrix, A.M. (1995). *Development of a coding form for approach control/pilot voice communications.* (Report No. DOT/FAA/AM-95/15). Washington, DC: FAA Office of Aviation Medicine.

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

Stager, P., Ho, G.W., & Garbutt, J.M. (2001, March). *An on-line measure of controller workload*. Paper presented at Eleventh International Symposium on Aviation Psychology, Columbus, OH.

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.

Stein, E.S. (1998). Human operator workload in air traffic control. In M. W. Smolensky & E. S. Stein (Eds.) *Human factors in air traffic control* (pp. 155-184). San Diego, CA: Academic Press.

Wickens, C.D., Mavor, A.S., & McGee, J.P. (Eds). (1997). *Flight to the future: Human factors in air traffic control* (p. 116). Washington, DC: National Academy Press.