DOT/FAA/AM-00/2

Office of Aviation Medicine
Washington, D.C. 20591

Measuring Air Traffic Controller Performance in a High-Fidelity Simulation

Carol A. Manning, Editor

Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma 73125

January 2000

Final Report

This document is available to the public through the National Technical Information Service, Springfield, Virginia 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-00/2	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Measuring Air Traffic Controller Performance in a High-Fidelity Simulation	5. Report Date January 2000 6. Performing Organization Code
7. Author(s) Manning, C.A., Editor	8. Performing Organization Report No.
9. Performing Organization Name and Address FAA Civil Aeromedical Institute 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 Aviation 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

This research was performed under Task HRR-509.

16. Abstract

In the summer of 1997, the Air Traffic Selection and Training (AT-SAT) High Fidelity Simulation Study was conducted at the FAA Academy in Oklahoma City, OK. The purpose of the study was to test the performance of 107 operational en route controllers during 2½ days of simulations. The performance of these controllers during the high-fidelity simulations was compared with their performance on two medium-fidelity performance measures to assess the construct validity of the latter measures to serve as criteria against which to validate a set of selection tests. The reports included in this document describe the high- fidelity simulation exercise, the development of performance measures utilized during the exercise, and the interrelationships between the performance measures.

The first report describes the development of a work sample approach to capturing air traffic controller performance, and establishes that high fidelity performance measures can adequately reflect the performance of the controller. The work sample was developed in an environment that simulated as nearly as possible the actual conditions existing in the controller's job, but was conducted in a "generic" airspace. Scenario development included the most important tasks from the task-based job analysis developed for the AT-SAT project. Sufficient time was provided for participating controllers to learn the airspace and procedures and demonstrate their knowledge through 1) a multiple choice test of airspace knowledge and 2) running 8 practice scenarios. Performance was measured by 1) an over-the-shoulder (OTS) rating scale, 2) counts of mistakes, 3) counts of actions that would be required to move aircraft from the sector at the end of the scenario, and 4) statistics derived from aircraft positions and controller/pilot data entries recorded for the simulation.

The second report used measures collected during the high-fidelity simulation study to predict the overall OTS performance rating. It was found that a model that included both counts of mistakes and the computer-derived performance measures predicted the OTS rating reasonably well, while a model containing only the computer-derived measures did not. Remaining actions did not contribute to the prediction of the OTS rating in addition to the contribution provided by the other types of measures.

17. Key Words Air Traffic Control, Performance Measurement, Simulation	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 37	22. Price

Reproduction of completed page authorized

MEASURING AIR TRAFFIC CONTROLLER PERFORMANCE IN A HIGH-FIDELITY SIMULATION

TABLE OF CONTENTS

Measuring the Performance of Air Traffic Controllers Using a High-Fidelity Work Sample Approach

Kenneth T. Bruskiewicz
Jerry W. Hedge
Personnel Decisions Research Institutes, Inc.

Carol Manning
Federal Aviation Administration
Civil Aeromedical Institute

Henry Mogilka
Federal Aviation Administration
FAA Academy

Introduction

Job performance is a complex concept that can be measured with a variety of techniques. A number of researchers (e.g., Ghiselli & Brown, 1948; Guion, 1979; Robertson & Kandola, 1982) have advocated the use of work sample tests because they are direct, relevant measures of job proficiency. Work sample tests measure an individual's skill level by extracting samples of behavior under realistic job conditions. Individuals are asked to demonstrate job proficiency by performing the activities required for successful completion of the work sample.

Measuring the job performance of air traffic controllers is a unique situation where reliance on a work sample methodology may be especially applicable. Use of a computer-generated simulation can create an air traffic control environment that allows the controller to behave realistically in a realistic setting. Such a simulation approach allows the researcher to provide high levels of stimulus and response fidelity (Tucker, 1984). Simulator studies of air traffic control problems have been reported in the literature since the 1950's. Most of the early research was directed toward evaluating the effects of workload variables and changes in control procedures on overall system performance, rather than focused on individual performance assessment (Boone, Van Buskirk, and Steen, 1980).

However, there have been some research and development efforts (e.g., Buckley, O'Connor, Beebe, Adams, and MacDonald, 1969; Buckley, DeBaryshe, Hitchner, and Kohn, 1983; and Sollenberger, Stein, and Gromelski, 1997) aimed at capturing the performance of air traffic controllers. These include full-scale dynamic simulations that allow controllers to direct the activities of a sample of simulated air traffic, performing characteristic functions such as ordering changes in aircraft speed or flight path, all within a relatively standardized work sample framework.

The current high fidelity performance measures were developed for construct validating a computerized low fidelity air traffic controller situational judgment test, the Computer-Based Performance Measure (CBPM) and behavior-based rating scales (see Borman et al. [1999] for more information on each of these measures). The Borman et al. (1999) measures were used as criterion measures for a large scale selection and validation project, the Federal Aviation Administration (FAA) Air Traffic Selection and Training (AT-SAT) project.

The intention of the High Fidelity Performance Measure (HFPM) study reported here was to provide an environment that would as nearly as possible simulate actual conditions existing in the controller's job. One possibility considered was to test each controller working in his or her own facility's airspace. This approach was eventually rejected, however, because of the problem of unequal difficulty levels (i.e., traffic density, airspace layout, etc.) across facilities and even across sectors within facility (Borman, Hedge, & Hanson, 1992; Hanson, Hedge, Borman, & Nelson, 1993; Hedge, Borman, Hanson, Carter, & Nelson, 1993). Comparing the performance of controllers working in environments with unequal (and even unknown) difficulty levels is extremely problematic. Therefore, we envisioned that performance could be assessed using a "simulated" air traffic environment. This approach was feasible because of the availability at the FAA Academy of several training laboratories equipped with radar stations similar to those found in field facilities. In addition, the Academy uses a generic airspace (Aero Center) designed to allow presentation of typical air traffic scenarios that must be controlled by the trainee (or in our case, the ratee). Use of a generic airspace also allowed for standardization of assessment. See Figure 1 for a visual depiction of the Aero Center airspace.

Thus, through use of the Academy's radar training facility (RTF) equipment, in conjunction with the Aero Center generic airspace, we were able to provide a testing environment affording the potential for both high stimulus and response fidelity. Our developmental efforts focused on: 1) designing and programming specific scenarios in which the controllers would control air traffic; and 2) developing measurement tools for evaluating controller performance.

Method

Scenario Development

The air traffic scenarios used in this study were designed to incorporate performance constructs central to the controller's job, such as maintaining aircraft separation, coordinating, communicating, and maintaining situation awareness. Also, attention was paid to representing in the scenarios the most important tasks from the task-based job analysis (see Nichels, Bobko, Blair, Sands, & Tartak, 1995).

Finally, it was decided that in order to obtain variability in controller performance, scenarios should be developed with either moderate or quite busy traffic conditions. Thus, to develop our HFPM scenarios, we started with a number of pre-existing Aero Center training scenarios, and revised and reprogrammed them to the extent necessary to include relevant tasks and performance requirements with moderate to high density traffic scenarios. In all, 16 scenarios were developed, each designed to run no more than 60 minutes, inclusive of start-up, position relief briefing, active air traffic control, debrief, and performance evaluation. Consequently, active manipulation of air traffic was limited to approximately 30 minutes. Time required for aircraft manipulation with the two part-task exercises was approximately 20 minutes, not including performance evaluation. After initial preparation of the scenarios, a pretest (using Academy instructors) and a pilot test (using 6 controller ratees), were conducted to increase the efficiency of the process, and minor revisions were made to general administrative procedures.

The development of a research design that would allow sufficient time for both training and evaluation was critical to the development of scenarios and accurate evaluation of controller performance. Sufficient training time was necessary to ensure adequate familiarity with the airspace, thereby eliminating differential knowledge of the airspace as a contributing factor to controller performance. Adequate testing time was important to ensure sufficient opportunity to capture controller performance, and allow for stability of evaluation. A final consideration was the need for controllers in our sample to travel to Oklahoma City to be trained and evaluated. With these criteria in mind, we arrived at a design that called for one-and one-half days of training (using 8 of the 16 scenarios), followed by one full day of performance. This schedule allowed us to train and evaluate two groups of ratees per week.

Development of Measurement Instruments

High fidelity performance data were captured by means of behavior-based rating scales and checklists, using trainers with considerable air traffic control experience or current controllers as raters. Development and implementation of these instruments, and selection and training of the HFPM raters, are discussed below.

OTS Rating Scales. Based on past research, it was decided that controller performance should be evaluated across broad dimensions, as well as at a more detailed step-by-step level. Potential performance dimensions for a set of rating scales were identified through reviews of previous literature involving air traffic control, existing on-the-job-training forms, performance verification forms, and current project work on the development of behavior summary scales. The over-the-shoulder (OTS) nature of this evaluation process, coupled with the maximal performance focus of the high fidelity simulation environment, required the development of rating instruments designed to facilitate efficient observation and evaluation of performance.

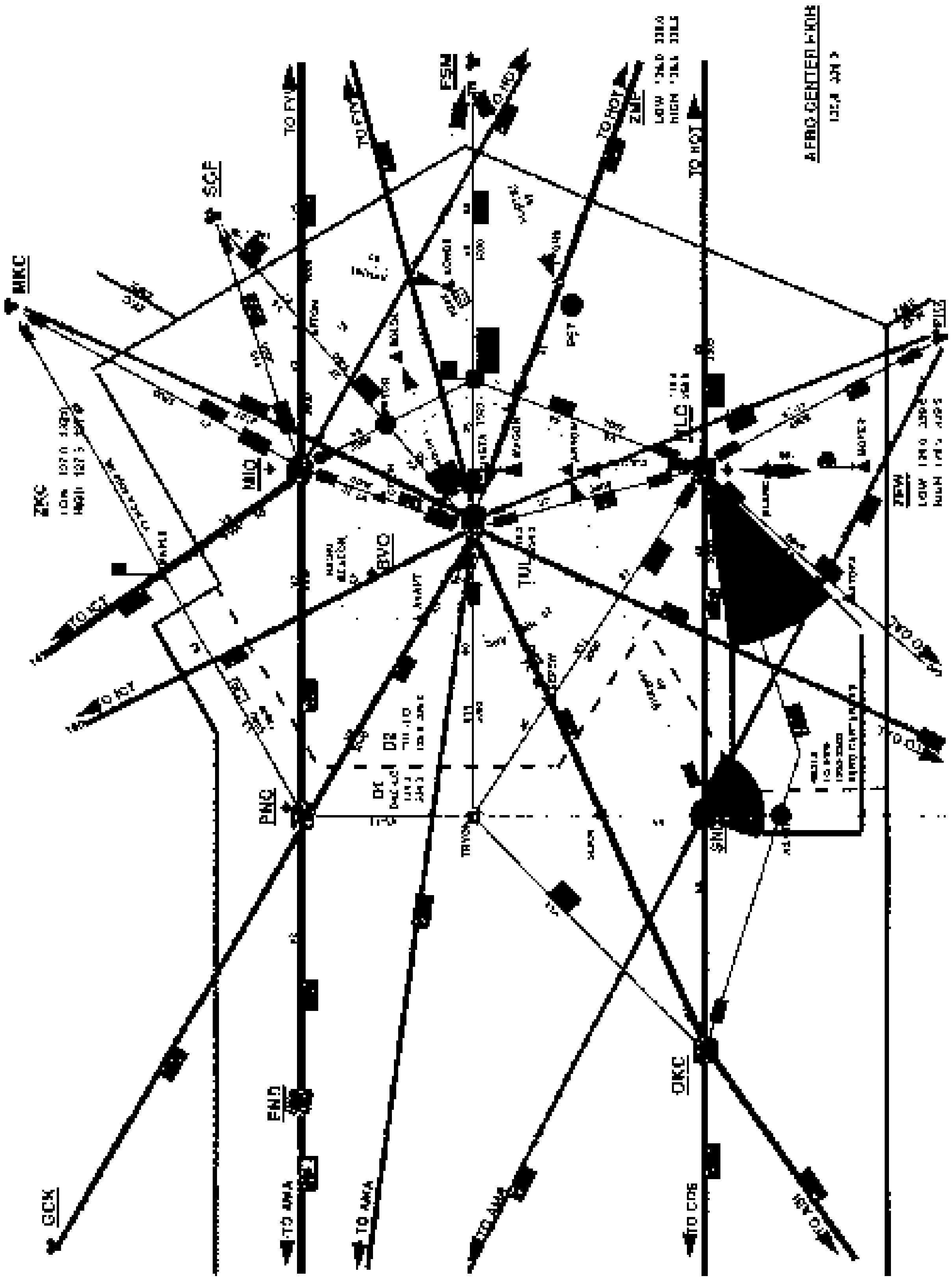


Figure 1. Aero Center Airspace.

After examining several possible scale formats, we chose a 7-point effectiveness scale for the OTS form, with the scale points clustered into three primary effectiveness levels; i.e., below average (1 or 2), fully adequate (3, 4, or 5), and exceptional (6 or 7). Through consultation with controllers currently working as Academy instructors, we tentatively identified eight performance dimensions, and developed behavioral descriptors for these dimensions to help provide a frame-of-reference for the raters. The eight dimensions were: (1) Maintaining Separation; (2) Maintaining Efficient Air Traffic Flow; (3) Maintaining Attention and Situation Awareness; (4) Communicating Clearly, Accurately, and Concisely; (5) Facilitating Information Flow; (6) Coordinating; (7) Performing Multiple Tasks; and, (8) Managing Sector Workload. We also included an "overall" performance category. As a result of rater feedback subsequent to pilot testing (described later in this chapter), "Facilitating Information Flow" was dropped from the form. This was due primarily to perceived overlap between this dimension and several others, including Dimensions 3, 4, 6, and 7.

Behavioral and Event Checklist. A second instrument required the raters to focus on more detailed behaviors and activities, and note whether and how often each occurred. The "Behavioral and Event Checklist" (BEC) required raters to actively observe the ratees controlling traffic during each scenario and note behaviors such as: (1) failure to accept handoffs, issue weather information, coordinate pilot requests, etc.; (2) Letters of Agreement (LOA)/directive violations; (3) readback/hearback errors; (4) unnecessary delays; (5) incorrect information input into the computer; and, (6) late frequency changes. Raters also noted operational errors, deviations, and special use airspace (SUA) violations.

Participants

The ratee participants were experienced controllers (*N*=107) from one of the en route air traffic control facilities across the United States. They were primarily white (81%) males (76%) with an average age of 37.4 years and had been at the full performance level (FPL; journeyman status) for an average of 8.7 years. The majority of them (80%) had attended college, and 40% of the sample had obtained a college degree. Fourteen persons served as raters for the data collection. Five of these raters were FAA Academy instructors, and the remaining 9 were staff/supervisors at en route facilities. As with the ratee sample, the rater sample consisted of primarily white (93%) males (100%) with an average age of 42.2 years who had worked as FPL controllers for an average of 9.5 years. All but one had attended college.

Rater Training

Fourteen highly experienced controllers from field units, or currently working as instructors at the FAA Academy, were detailed to serve as raters for the HFPM portion of the AT-SAT project. To allow for adequate training and pilot testing, raters arrived approximately three weeks before the start of data collection. Thus, rater training occurred over an extended period of time, affording an opportunity for ensuring high levels of rater calibration.

During their first week at the Academy, raters were exposed to (1) general orientation to the AT-SAT project, its purposes and objectives, and the importance of the high fidelity component; (2) airspace training; (3) the HFPM instruments; (4) all supporting materials (such as Letters of Agreement, etc.); (5) training and evaluation scenarios; (6) part-task exercises; and, (7) rating processes and procedures. During this first week raters served as both raters and ratees, controlling traffic in each scenario multiple times, as well as serving as raters of their associates who took turns as ratees. This process allowed raters to become extremely familiar with both the scenarios and evaluation of performance in these scenarios. With multiple raters evaluating performance in each scenario, project personnel were able to provide immediate critique and feedback to raters, aimed at improving accuracy and consistency of rater observation and evaluation. In addition, prior to rater training, we "scripted" performances on several scenarios, such that deliberate errors were made at various points by the individual controlling traffic. Raters were exposed to these "scripted" scenarios early in the training so as to more easily facilitate discussion of specific types of controlling errors. Thus, the training program was an extremely hands-on, feedback-intensive process.

A standardization guide was also developed, such that rules for how observed behaviors were to be evaluated could be referred to during data collection if any questions arose (see the Appendix). All of these activities contributed to near optimal levels of rater calibration.

Pilot Test

A pilot test of the HFPM was conducted to determine whether the rigorous schedule of one-and one-half days of training and one day of evaluation was feasible administratively. Our admittedly ambitious design required completion of up to eight practice scenarios and eight graded scenarios. Start-up and shutdown of each computer-generated scenario at each radar station, setup and breakdown of associated flight strips, pre-and-post position relief briefings, and completion of OTS ratings and checklists all had to be accomplished within the allotted time, for all training and evaluation scenarios. Thus, smooth coordination and timing of activities was essential. Prior to the pilot test, preliminary "dry runs" had already convinced us to eliminate one of the eight available evaluation scenarios, due to time constraints. Table 1 provides a brief description of the design of the seven remaining evaluation scenarios.

Six experienced controllers currently employed as instructors at the Academy served as our ratees for the pilot test, and were administered the entire two-and one-half day training/evaluation process, from orientation through final evaluation scenarios. As a result of the pilot test, and in an effort to increase the efficiency of the testing and rating process, minor revisions were made to general administrative procedures. In general, procedures for administering the HFPM proved to be effective; all anticipated training and evaluation requirements were completed on time and without major problems.

Procedure

Controllers from 14 different ATC facilities throughout the United States participated in the 2 ½ day high fidelity performance measurement process. The 1 ½ days of ratee training consisted of 4 primary activities: orientation, airspace familiarization and review, airspace certification testing, and scenarios practice. In order to accelerate learning time, a hard copy and computer disk describing the airspace had been developed and sent to controllers at their home facility to review prior to arrival in Oklahoma City. After completing the orientation, and training on the first 2 scenarios, the ratees were required to take an airspace certification test. The certification consisted of 70 recall and recognition items designed to test knowledge of the airspace. Those individuals not receiving a passing grade (at least 70% correct) were required to retest on that portion of the test they did not pass. The 107 controllers scored an average of 94% on the test, with only 7 failures (6.5%) on the first try. All controllers subsequently passed the retest and were certified by the trainers to advance to the remaining day of formal evaluation.

After successful completion of the air traffic test, each ratee received training on 6 additional air traffic scenarios. During this time, the raters acted as trainers, and facilitated the ratee's learning of the airspace. While questions pertaining to knowledge of airspace and related regulations were answered by the raters, coaching ratees on how to more effectively and efficiently control traffic was prohibited. Once all training scenarios were completed, all ratees' performance was evaluated on 7 "graded" scenarios and 2 part-task exercises, that, together, required 8 hours to complete. The 7 graded scenarios consisted of 4 moderately busy and 3 extremely busy air traffic conditions, increasing in complexity from Scenario 1 to Scenario 7. During this 8 hour evaluation period, raters were randomly assigned to ratees before each scenario, with the goal that a rater should not be assigned to a ratee (1) from the rater's home facility; or (2) if he/she was the ratee's trainer during training.

While the ratee was controlling traffic in a particular scenario, the rater continually observed and noted performance using the BEC. Once the scenario had ended, each rater completed the OTS ratings. In all, 11 training/evaluation sessions were conducted within a 7 week period. During four of these sessions, each ratee was evaluated by 2 raters, while a single rater evaluated each ratee performance during the other 7 sessions.

Analyses

Means and standard deviations were computed for all criterion measures collected via the work sample methodology. Criterion variable intercorrelations were also computed. Interrater reliabilities were examined by computing intraclass correlations (Shrout & Fleiss, 1978) between rater pairs for the OTS rating scales for those 24 ratees for whom multiple rater information was available. Selected variables were subjected to principal components analyses in order to create composite scores from the individual measures. The composite scores were then correlated with the other criterion measures from the AT-SAT project.

Table 1. Descriptions of Air Traffic Scenarios Used in High Fidelity Simulation.
HFG1 — This scenario contains 14 aircraft. Two of the aircraft are “pop-ups” and request IFR clearances. There are three MLC arrivals and three TUL arrivals with two of the three TUL arrivals conflicting at WAGON intersection. There are four departing aircraft; one from MIO, one from TUL, and two from MLC. There is one pair of overflight aircraft that will lose separation if no action is taken. The only unusual situation is that ZME goes DARC at the beginning of the scenario and requires manual handoffs.
HFG2 — This scenario contains 25 aircraft. One aircraft is NORDO and one aircraft turns off course without authorization. There is one MLC arrival and four TUL arrivals with three of the four TUL arrivals conflicting at WAGON intersection. There are 10 departing aircraft; one from MIO, seven from TUL, and two from MLC. There is one pair of overflight aircraft that will lose separation if no action is taken. There are no unusual situations.
HFG3 — This scenario contains 26 aircraft. One aircraft loses Mode C, one aircraft squawks 7600, and one aircraft requests a more direct route around weather. There is one BVO arrival, two MIO arrivals, and four TUL arrivals with three of the four TUL arrivals conflicting at WAGON intersection. There are 10 departing aircraft; three from MIO, five from TUL, and two from MLC. There is one pair of overflight aircraft that will lose separation if no action is taken. One military aircraft requests a change of destination. There are no unusual situations.
HFG4 — This scenario contains 25 aircraft. One aircraft reports moderate turbulence and requests a lower altitude, one aircraft requests vectors around weather, one aircraft requests a lower altitude to get below weather, and one aircraft descends 500 feet below assigned altitude without authorization. There are two MIO arrivals, three MLC arrivals, and three TUL arrivals. There are nine departing aircraft; seven from TUL, and two from MLC. The only unusual situation is that TMU requests all ORD arrivals be re-routed (only applies to one aircraft).
HFG5 — This scenario contains 28 aircraft. One aircraft requests vectors around weather, one aircraft requests a vector to destination, one aircraft requests RNAV direct to destination, and one aircraft descends 800 feet below assigned altitude without authorization. There are two MIO arrivals, one MLC arrival, and no TUL arrivals. There are 11 departing aircraft; eight from TUL, and three from MLC. One military aircraft requests a change of destination. There are two unusual situations; ZFW goes DARC and requires manual handoffs, and one overflight aircraft declares an emergency and requests to land at TUL.
HFG6 — This scenario contains 32 aircraft. One overflight aircraft requests to change their destination to DAL and one aircraft requests RNAV direct to destination. There are two MIO arrivals, one MLC arrival, and two TUL arrivals. There are 12 departing aircraft; eight from TUL, two from MLC, and two from MIO. There are two pairs of overflight aircraft that will lose separation if no action is taken. One military aircraft requests a change of destination. There are no unusual situations.
HFG7 — This scenario contains 33 aircraft. One overflight aircraft requests a change in destination, one overflight aircraft requests RNAV direct to destination, and one aircraft requests vectors to destination. There are three MIO arrivals, two MLC arrivals, and six TUL arrivals with five of the six TUL arrivals conflicting at WAGON intersection. There are 11 departing aircraft; seven from TUL, three from MLC, and one from MIO. There is one pair of overflight aircraft that will lose separation if no action is taken. One military aircraft requests a change of destination. There are no unusual situations.

Results

Descriptive Statistics

Table 2 contains descriptive statistics for the variables included in both of the rating instruments used during the HFPM graded scenarios. For the OTS dimensions and the BEC, the scores represent averages across each of the seven graded scenarios.

The means of the individual performance dimensions from the 7-point OTS rating scale are in the first section of Table 2 (Variables 1 through 7). They range from a low of 3.66 for *Maintaining Attention and Situation Awareness* to a high of 4.61 for *Communicating Clearly, Accurately and Efficiently*. The scores from each of the performance dimensions are slightly negatively skewed, but are for the most part, normally distributed.

Variables 8 through 16 in Table 2 were collected using the BEC. To reiterate, these scores represent instances where the controllers had either made a mistake or engaged in some activity that caused a conflict, a delay, or in some other way impeded the flow of air traffic through their sector. For example, a *Letter of Agreement (LOA)/Directive Violation* was judged to have occurred if an aircraft was not established at 250 knots prior to crossing the appropriate arrival fix or if a frequency change was issued prior to completion of a handoff for the appropriate aircraft. On average, each participant had 2.42 *LOA/Directive Violations* in each scenario.

Table 2. Descriptive Statistics of High Fidelity Performance Measure Criterion Variables.

	N	Mean	SD
OTS Dimensions:			
1. Maintaining Separation	107	3.98	1.05
2. Maintaining Efficient Air Traffic Flow	107	4.22	.99
3. Maintaining Attention and Situation Awareness	107	3.66	1.02
4. Communicating Clearly, Accurately, and Efficiently	107	4.61	.96
5. Coordinating	107	4.17	.97
6. Performing Multiple Tasks	107	4.40	1.00
7. Managing Sector Workload	107	4.39	1.03
Behavior and Event Checklist:			
8. Operational Errors	107	.05	.04
9. Operational Deviations	107	.11	.07
10. Failed To Accept Handoff	107	.31	.46
11. LOA/Directive Violations	107	2.42	1.26
12. Readback/Hearback Errors	107	.46	.44
13. Fail to Accommodate Pilot Request	107	.45	.33
14. Make Late Frequency Changes	107	.44	.43
15. Unnecessary Delays	107	2.68	1.56
16. Incorrect Information in Computer	107	1.04	.66

Intercorrelations for criterion variables

Table 3 shows intercorrelations for the OTS dimensions and BEC items. The OTS dimensions were very highly correlated, with intercorrelations ranging from .80 to .97 (median r = .91). The BEC variables were negatively correlated with the OTS dimensions (higher scores on the BEC indicated more errors or procedural violations, while higher ratings on the OTS rating scales indicated better performance.) Most BEC variables had statistically significant intercorrelations, although *Operational Errors* was not significantly correlated with *Incorrect Information in Computer. Delays* had correlations of .55 or higher with *Fail to Accept Handoffs*, *LOA/Directive Violations*, and *Fail to Accommodate Pilot Request. LOA/Directive Violations* correlated .53 with *Operational Errors* and to a lesser degree with *Operational Deviations* (r =.35).

Interrater Reliabilities

Table 4 contains interrater reliabilities for the OTS ratings for the 24 ratees for whom multiple rater information was available. Overall, the interrater reliabilities were quite high for the OTS ratings, with median interrater reliabilities ranging from a low of 0.83 for *Maintaining Attention and Situation Awareness* to a high of 0.95 for *Maintaining Separation*.

Principal Components Analysis

Relevant variables for the OTS and BEC measures were combined and subjected to an overall principal components analysis to represent a final high fidelity performance criterion space. The resulting two factor solution is presented in Table 5. The first component, *Overall Technical Proficiency*, consists of the OTS rating scales, plus *Operational Error*, *Operational Deviation*, and *LOA/Directive Violation* variables from the BEC. The second component is defined by 6 additional BEC variables, and represents a s*ector management* component of controller performance. More specifically, this factor represents *Poor Sector Management*, whereby the controllers more consistently made late frequency changes, failed to accept hand-offs, commited readback/ hearback errors, failed to accommodate pilot requests, delayed aircraft unnecessarily, and entered incorrect information in the computer. This interpretation is reinforced by the strong negative correlation (-.72) found between *Overall Technical Proficiency* and *Poor Sector Management.*

Correlations of High Fidelity Criterion Composites with CBPM and Supervisor/ Peer Rating Scale Composite

In order to provide a broader perspective within which to place the HFPM, this section provides a brief overview of relationships between the HFPM and other AT-SAT criterion measures. First, we briefly describe the content of the CBPM and the rating scale composites. Interested readers are referred to Borman et al. (1999) for a more in-depth description of the development and design of the CBPM and the supervisor/peer rating scale composite.

In the CBPM, air traffic controllers are presented with a series of air traffic scenarios, flight strips providing detailed information about flight plans for each of the aircraft in the scenario, and a status information area (e.g., containing weather information). Controllers were given one minute to review the materials for each scenario, after which they watched the scenario unfold. They were then required to answer a series of questions about each scenario. The final version of the CBPM, which was used in computing the following correlations, consisted of 38 items and had an internal consistency reliability of .61.

The other component of the criterion space for the AT-SAT validation effort was a set of behavior-based rating scales. Ten performance categories were initially included: (1) Maintaining Safe & Efficient Air Traffic Flow, (2) Maintaining Attention & Vigilance, (3) Prioritizing, (4) Communicating & Informing, (5) Coordinating, (6) Managing Multiple Tasks, (7) Reacting to Stress, (8) Adaptability & Flexibility, (9) Technical Knowledge, and (10) Teamwork. Ratings were collected from both supervisor and peer perspectives and subjected to factor analyses. The factor analyses indicated that the one-factor model was sufficient for describing the data, thus, the ratings were averaged into an overall composite.

Table 6 contains correlations between scores on the 38 item CBPM, the two HFPM factors, and the combined supervisor/peer ratings. First, the correlation between the CBPM total scores and the HFPM Component 1, arguably our purest measure of technical

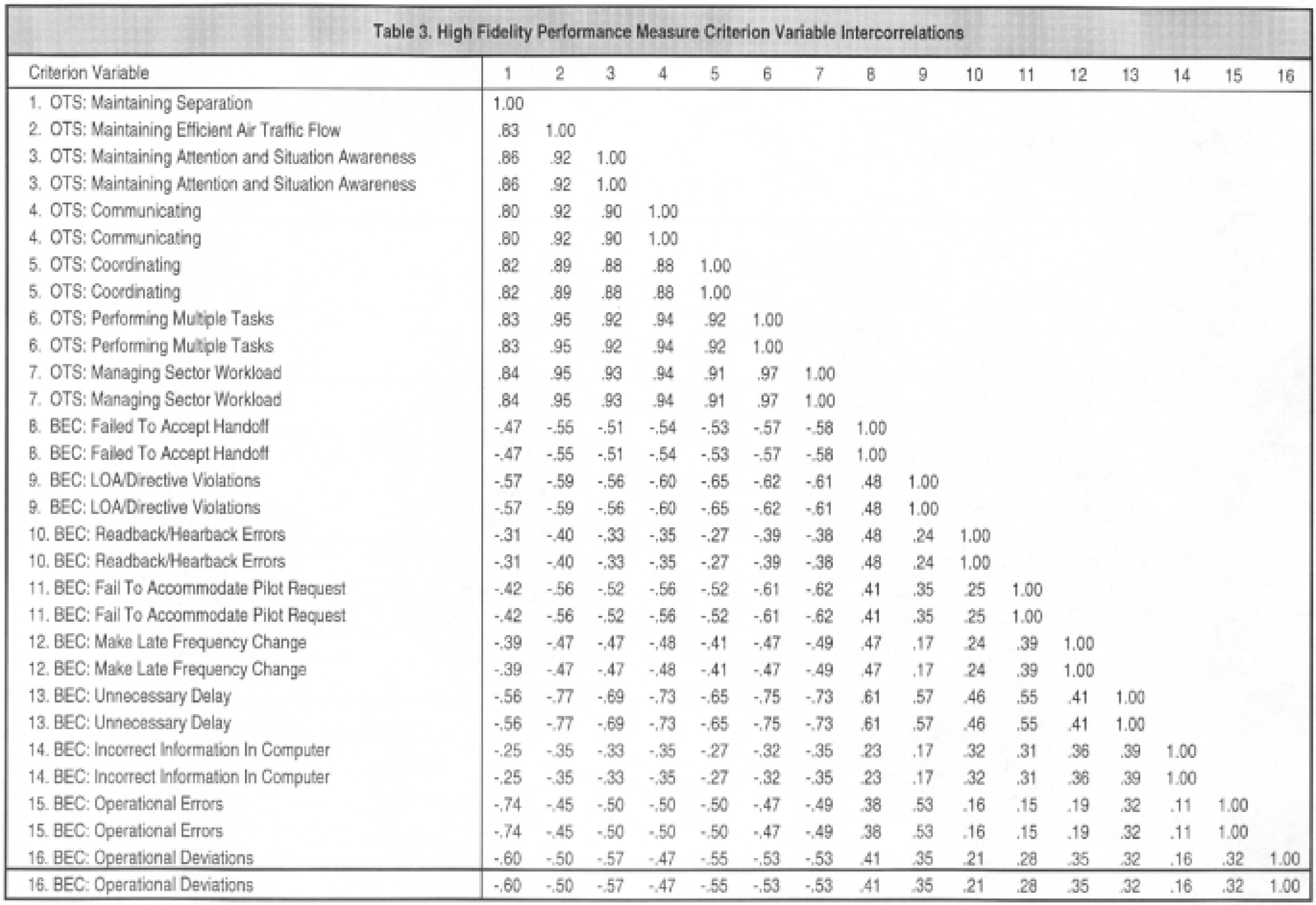


Table 3. High Fidelity Performance Measure Criterion Variable Intercorrelations

Criterion Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. OTS: Maintaining Separation	1.00															
2. OTS: Maintaining Efficient Air Traffic Flow	.83	1.00														
3. OTS: Maintaining Attention and Situation Awareness	.86	.92	1.00													
3. OTS: Maintaining Attention and Situation Awareness	.86	.92	1.00													
4. OTS: Communicating	.80	.92	.90	1.00												
4. OTS: Communicating	.80	.92	.90	1.00												
5. OTS: Coordinating	.82	.89	.88	.88	1.00											
5. OTS: Coordinating	.82	.89	.88	.88	1.00											
6. OTS: Performing Multiple Tasks	.83	.95	.92	.94	.92	1.00										
6. OTS: Performing Multiple Tasks	.83	.95	.92	.94	.92	1.00										
7. OTS: Managing Sector Workload	.84	.95	.93	.94	.91	.97	1.00									
7. OTS: Managing Sector Workload	.84	.95	.93	.94	.91	.97	1.00									
8. BEC: Failed To Accept Handoff	-.47	-.55	-.51	-.54	-.53	-.57	-.58	1.00								
8. BEC: Failed To Accept Handoff	-.47	-.55	-.51	-.54	-.53	-.57	-.58	1.00								
9. BEC: LOA/Directive Violations	-.57	-.59	-.56	-.60	-.65	-.62	-.61	.48	1.00							
9. BEC: LOA/Directive Violations	-.57	-.59	-.56	-.60	-.65	-.62	-.61	.48	1.00							
10. BEC: Readback/Hearback Errors	-.31	-.40	-.33	-.35	-.27	-.39	-.38	.48	.24	1.00						
10. BEC: Readback/Hearback Errors	-.31	-.40	-.33	-.35	-.27	-.39	-.38	.48	.24	1.00						
11. BEC: Fail To Accommodate Pilot Request	-.42	-.56	-.52	-.56	-.52	-.61	-.62	.41	.35	.25	1.00					
11. BEC: Fail To Accommodate Pilot Request	-.42	-.56	-.52	-.56	-.52	-.61	-.62	.41	.35	.25	1.00					
12. BEC: Make Late Frequency Change	-.39	-.47	-.47	-.48	-.41	-.47	-.49	.47	.17	.24	.39	1.00				
12. BEC: Make Late Frequency Change	-.39	-.47	-.47	-.48	-.41	-.47	-.49	.47	.17	.24	.39	1.00				
13. BEC: Unnecessary Delay	-.56	-.77	-.69	-.73	-.65	-.75	-.73	.61	.57	.46	.55	.41	1.00			
13. BEC: Unnecessary Delay	-.56	-.77	-.69	-.73	-.65	-.75	-.73	.61	.57	.46	.55	.41	1.00			
14. BEC: Incorrect Information In Computer	-.25	-.35	-.33	-.35	-.27	-.32	-.35	.23	.17	.32	.31	.36	.39	1.00		
14. BEC: Incorrect Information In Computer	-.25	-.35	-.33	-.35	-.27	-.32	-.35	.23	.17	.32	.31	.36	.39	1.00		
15. BEC: Operational Errors	-.74	-.45	-.50	-.50	-.50	-.47	-.49	.38	.53	.16	.15	.19	.32	.11	1.00	
15. BEC: Operational Errors	-.74	-.45	-.50	-.50	-.50	-.47	-.49	.38	.53	.16	.15	.19	.32	.11	1.00	
16. BEC: Operational Deviations	-.60	-.50	-.57	-.47	-.55	-.53	-.53	.41	.35	.21	.28	.35	.32	.16	.32	1.00
16. BEC: Operational Deviations	-.60	-.50	-.57	-.47	-.55	-.53	-.53	.41	.35	.21	.28	.35	.32	.16	.32	1.00

Note. A correlation of approximately .25 is statistically significantly different from zero at $p < .01$. *N*=107 ratees.

Table 4. Interrater Reliabilities[a] for OTS Ratings.

Dimension	Median	Range
1. Maintaining Separation	.95	.83 to .98
2. Maintaining Efficient Air Traffic Flow	.89	.71 to .94
3. Maintaining Attention and Situation Awareness	.83	.79 to .87
4. Communicating	.91	.88 to .93
5. Coordinating	.91	.86 to .96
6. Managing Multiple Tasks	.88	.82 to .93
7. Managing Sector Workload	.91	.85 to .95

[a] Reliabilities are 2-rater intraclass correlation coefficients; these coefficients reflect the reliability of the mean ratings. *N*=24 ratees.

proficiency, is .61. This provides strong evidence for the construct validity of the CBPM. Apparently, this lower fidelity measure of technical proficiency is tapping much the same technical skills as the HFPM, in which controllers worked in an environment highly similar to their actual job setting. In addition, a significant negative correlation exists between the CBPM and the second HFPM component, Poor Sector Management.

Discussion

The likelihood of accurately measuring controller performance increases with the extent to which one is able to place controllers in a standardized and realistic environment in which they must control traffic, and that affords reliable measurement of their performance. The current set of high fidelity performance measures represents a simulation that provides for reliable individual air traffic controller performance measurement. As such, the 16 individual performance scores and two component model represent a parsimonious, comprehensive, and psychometrically sound depiction of the air traffic controller criterion space.

By their very nature, rating scales and checklists that rely on human raters to provide assessment of performance involve a certain degree of subjectivity. However, we believe that the Over-the-Shoulder (OTS) rating form and the Behavioral and Event Checklist (BEC) increase the chances that the evaluations provided are relatively accurate depictions of the performance observed. This is so for at least three reasons: 1) HFPM raters received extensive training on all aspects of the simulation process, especially observation and rating accuracy training; 2) the OTS and BEC forms were developed with detailed attention to the performance requirements of the simulation scenarios and exercises, and the evaluation requirements of the raters; and, 3) the simulation scenarios provide a relatively standardized environment within which ratees can perform, and raters can evaluate that performance.

In addition to using the "rater collected" measures reported here, further research is exploring the utility of collecting measures derived directly from the computer system itself. These computer-derived measures certainly offer advantages over subjective ratings, but are not without disadvantages. Advantages

Table 5. Principal Components Analysis Results

Label	Variables	Component 1	Component 2
Overall Technical Proficiency	OTS: Maintaining Separation	.95	.05
	OTS: Coordinating	.87	-.12
	BEC: Operational Errors	-.85	-.36
	OTS: Maintaining Attention/Awareness	.83	-.20
	OTS: Performing Multiple Tasks	.81	-.27
	OTS: Managing Sector Workload	.80	-.29
	OTS: Communicating	.79	-.27
	OTS: Maintaining Efficient Air Traffic Flow	.78	-.30
	BEC: LOA/Directive Violations	-.76	-.07
	BEC: Operational Deviations	-.59	.05
Poor Sector Management	BEC: Incorrect Information in Computer	.10	.72
	BEC: Readback/Hearback Errors	-.01	.63
	BEC: Make Late Frequency Changes	-.13	.60
	BEC: Fail to Accommodate Pilot Request	-.27	.54
	BEC: Unnecessary Delays	-.45	.53
	BEC: Fail to Accept Handoffs	-.37	.45
	Percent Variance Accounted For:	59	9

Table 6. Correlations of High Fidelity Criterion Composites with CBPM and Supervisor/Peer Rating Scale Composite.

	High Fidelity Criterion Composites	
CBPM and Supervisor/Peer Rating Scale Composite.	**Overall Technical Proficiency**	**Poor Sector Management**
CBPM (38 Items)	.61**	-.42**
Rating Scale Composite	.40**	-.28**

Note. Sample sizes range from 106 to 107. ***p* < .01

include: 1) they are objective and thus, possibly more reliable; 2) they can provide more precise and accurate measurement of some variables; and 3) they can provide performance assessment in terms of system outcome effects. Potential disadvantages include: 1) some important job measures may be inaccessible by system measures; 2) they provide no direct information about techniques and procedures used; 3) they allow for considerable possibility of contamination (built-in bias) due to system parameters that have nothing to do with individual controller proficiency; and 4) they may be extremely sensitive to variations in experience, job assignment, and non-standardization of work samples being recorded, which are factors that contribute to both bias and unreliability.

Additional research is currently underway further exploring the manner in which measures collected directly from the computer system fit into the model described here. Specifically, we are examining what sorts of information additional computer-derived measures add to measures currently collected using raters. We expect that the computer-derived measures should be able to add significantly to the criterion space, especially to the extent that they are able to more accurately and precisely measure operational errors and deviations, and provide for more exact measurement of an air traffic controller's efficiency (e.g., the number of altitude/heading changes required to guide aircraft through the sector).

While a work sample approach to criterion measurement is not appropriate for all jobs and situations, the air traffic control environment offers a unique opportunity to design and develop a work sample with high stimulus and response fidelity characteristics. We believe the present research describes one example of the successful application of the work sample methodology, as well as the development and evaluation of several useful techniques for measuring performance within the work sample framework.

References

Boone, J., Van Buskirk, L., & Steen, J. (1980). *The Federal Aviation Administration's radar training facility and employee selection and training* (FAA-AM-80-15). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Borman, W. C., Hedge, J. W., & Hanson, M. A. (1992, June). *Criterion development in the SACHA project: Toward accurate measurement of air traffic control specialist performance* (Institute Report #222). Minneapolis: Personnel Decisions Research Institutes.

Borman, W. C., Hedge, J. W., Hanson, M. A., Bruskiewicz, K. T., Mogilka, H. J., Manning, C., Bunch, L. B., & Horgen, K. E. (1999). Development of criterion measures of Air Traffic Controller performance. In B. Ramos (Ed.) *Air Traffic Selection and Training (AT-SAT) final report.* Alexandria, VA: HumRRO.

Buckley, E. P., DeBaryshe, B. D., Hitchner, N., & Kohn, P. (1983). *Methods and measurements in real-time air traffic control system simulation* (DOT/FAA/CT-83/26). Atlantic City, NJ: DOT/FAA Technical Center.

Buckley, E. P., O'Connor, W. F., Beebe, T., Adams, W., & MacDonald, G. (1969). *A comparative analysis of individual and system performance indices for the air traffic control system* (NA-69-40). Atlantic City, NJ: DOT/FAA Technical Center.

Ghiselli, E., & Brown, C. (1948). *Personnel and industrial psychology.* New York: McGraw-Hill.

Guion, R. M. (1979*). Principles of work sample testing: I. A non-empirical taxonomy of test users* (ARI-TR-79-A8). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Hanson, M. A., Hedge, J. W., Borman, W. C., & Nelson, L. C. (1993). Plans for developing a set of simulation job performance measures for air traffic control specialists in the Federal Aviation Administration (Institute Report #236). Minneapolis: Personnel Decisions Research Institutes.

Hedge, J. W., Borman, W. C., Hanson, M. A., Carter, G. W., & Nelson, L. C. (1993). *Progress toward development of air traffic control specialist performance criterion measures* (Institute Report #235). Minneapolis: Personnel Decisions Research Institutes.

Nickels, B. J., Bobko, P., Blair, M. D., Sands, W. A., & Tartak, E. L. (1995*). Separation and Control Hiring Assessment (SACHA) Final Job Analysis Report.* Bethesda, MD: University Research Corporation.

Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact, and applicant reaction. *Journal of Occupational Psychology, 55*, 171-183.

Sollenberger, R. L., Stein, E. S., & Gromelski, S. (1997). *The development and evaluation of a behaviorally based rating form for assessing air traffic controller performance* (DOT/FAA/CT-TN96-16). Atlantic City, NJ: DOT/FAA Technical Center.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428.

Tucker, J. A. (1984). Development of dynamic paper-and-pencil simulations for measurement of air traffic controller proficiency (pp. 215-241). In S. B. Sells, J. T. Dailey & E. W. Pickrel (Eds.), *Selection of air traffic controllers* (FAA-AM-84-2). Washington, DC: DOT/FAA/OAM.

Appendix

AT-SAT High Fidelity Standardization Guide

The following rules and interpretations of rules have been agreed to and will be used in evaluations by all AT-SAT Raters in addition to rules set forth in FAA Handbook 7110.65, Aero ARTCC and Tulsa ATCT Letter of Agreement, Aero ARTCC and McAlester ATCT Letter of Agreement, and Aero ARTCC, Memphis ARTCC, Kansas City ARTCC, Fort Worth ARTCC Letter of Agreement.

General

All aircraft have to be vectored for straight-in ILS approach to MLC.

If aircraft goes into TUL airspace then back out, just rate performance for the first time the aircraft is in your airspace.

If you make a mistake when filling out any of the forms, either erase the mark or draw a squiggly line through the incorrect mark.

If participant fails to say “Radar service terminated,” don’t mark any Remaining Actions, but consider when making OTS ratings.

If the pilot makes a mistake that results in an OE or OD, mark on behavioral checklist, put an asterisk next to indicator, and explain circumstance. If pilot causes OE or OD, the 1/2 rule does not apply (1 OE = OTS rating of 2 in Category A, 2 OES = OTS rating of 1).

Behavioral Checklist

Operational Errors

An Operational Error is considered to occur if a non-radar clearance **does not** provide for positive separation, regardless if controller corrects error prior to loss of radar separation.

If the participant makes one Operational Error, the rater shall assign a rating no higher than 2 in the Maintaining Separation (A) category on the OTS rating form. If the participant makes two Operational Errors, the rater shall assign a rating no higher than 1 in the Maintaining Separation (A) category on the OTS rating form. If participant makes no OEs, rater may assign any number for category A. Making an operational error will not necessarily affect ratings for other categories except that if a participant is rated low on A (Maintaining Separation) on the OTS form, they will also probably be rated low on C (Maintaining Attention and Situation Awareness).

If an aircraft is cleared off an airport, is auto-acquired off the departure list, but the participant is not yet talking to the aircraft, it is **NOT** an OE if another aircraft is cleared for approach into that same airport.

If an aircraft is cleared below the MIA, it is an OE.

It an aircraft is cleared for approach without telling the pilot to maintain a specific altitude, it is an OE.

If an aircraft without Mode C doesn’t report level, the participant doesn’t determine a reported altitude, and the aircraft flies over another aircraft, it shall be scored as an OE. Also, if the participant doesn’t enter a reported altitude in the computer, it shall also be scored as Incorrect Information in Computer.

Operational Deviations

An Operational Deviation is considered to occur if there is a violation of published MEAs.

An Operational Deviation is considered to occur if an aircraft comes within 2.5 miles of the airspace of another facility without being handed off. If the scenario freezes before the aircraft gets within 2.5 miles of another facility’s airspace and it hasn’t yet been handed off, count as Make Handoff under Remaining Actions.

An Operational Deviation occurred if the participant failed to point out an aircraft to the appropriate sector or if the participant issued a clearance to an aircraft while it is within 2.5 miles of the airspace boundary. Raters should check the location of the aircraft when a clearance is issued to see if it is within 2.5 miles of the boundary. If it is, an OD should be counted.

Special Use Airspace Violation

A Special Use Airspace violation is considered to occur if an aircraft does not remain clear of P57 or if an aircraft does not clear Restricted Area R931A by either 3 NM or 500 feet of altitude.

Accepted Handoff/Pointout Late

Acceptance of a Handoff/Pointout will be considered late if the radar target is within 2.5 NM of 1) Tulsa Approach boundary if the aircraft is exiting Tulsa Approach airspace or 2) crossing the Aero Center boundary if the aircraft is transiting En-Route airspace.

LOA/Directive Violation

A violation of the Tulsa Letter of Agreement is considered to occur if a jet aircraft is not established at 250 knots prior to crossing the appropriate arrival fix, if an aircraft is not level at prescribed arrival altitudes at appropriate arrival fix, even if a different altitude, etc., was coordinated, or if aircraft are not appropriately spaced.

There will be no blanket coordination of altitude or speed restrictions different than those specified in the LOA. For specific circumstances when pilots aren't going to meet crossing restrictions, if that is coordinated, it won't be counted as an LOA violation.

Count as LOA/Directive Violation if a frequency change is issued prior to completion of a handoff for the appropriate aircraft, if the participant changes frequency but did not terminate radar, or if the participant flashed the aircraft too early.

Count as LOA/Directive Violation if the participant failed to forward a military change of destination to FSS.

Count as LOA/Directive Violation if the participant makes a handoff to and switches the frequency to the incorrect facility. Don't include in Remaining Actions.

Count as LOA/Directive Violation if the participant drops a data block while the aircraft is still inside the airspace.

Count as LOA/Directive Violation if the participant fails to inform the pilot of radar contact.

If participant has an LOA/Directive Violation, also mark as Coordination error. If mark several violations, consider marking down Coordination and overall categories.

Failed to Accommodate Pilot Request

Participants shall be rated as failing to accommodate a pilot request if the controller never takes appropriate action to accommodate the request, if the controller says unable when he/she could have accommodated the request, or if the controller says stand by and never gets back to the pilot. This situation applies if the rater determines that the controller could have accommodated the request without interfering with other activities. Rater must balance failing to accommodate pilot requests or other delays against factors involved in Managing Sector Workload.

If another facility calls for a clearance and the participant fails to issue it unnecessarily, counts as Delay, not as Failure to Accommodate Pilot Request.

Unnecessary Delay

An unnecessary delay is considered to occur if a pilot request can be accommodated and the controller delays in doing so, if the participant levels any departure at an altitude below the requested altitude and there was no traffic, or if an aircraft previously in holding due to approaches or departures at MIO and MLC airports is not expeditiously cleared for approach.

If the participant leaves an aircraft high on the localizer it is considered a delay if the pilot/computer says unable. If the pilot/computer does not say unable but the participant could have descended the aircraft sooner, count down on category C (Maintaining Attention and Situation Awareness).

If another facility calls for a clearance and the participant fails to issue it unnecessarily, counts as Delay, not as Failure to Accommodate Pilot Request.

Incorrect Information in Computer

If an aircraft does not have Mode C, the participant shall enter the reported altitude 1) when the pilot reports it, 2) prior to handoff, or 3) by the end of the scenario. If this does not happen, count as Incorrect Information in Computer, Also, see OE.

Incorrect Information in Data Block

Altitude information in data blocks shall be considered incorrect if and when reported altitude differs by 1000 feet or more from assigned altitude displayed in same data block.

OTS Rating Form

Coordinating

In the event any information needs to be passed to a supervisor, the AT-SAT Rater shall be considered acting as same supervisor. Coordination of climbing aircraft shall NOT be required as long as the aircraft's data block/flight plan correctly displays the aircraft's assigned altitude.

If participant doesn't enter computer information (for example, change in route), enters incomplete information, or enters information in the computer for the wrong aircraft, rate them down under OTS Category E (Coordination). Don't mark the Behavioral Checklist or use the Remaining Actions form. This is not to be rated as an OD.

If participant didn't coordinate a WAFDOF for aircraft within 2.5 miles of sector boundary, it counts as a coordination error (Category E on OTS). If scenario freezes before coordination occurred but there was still time to accomplish coordination within 2.5 miles of sector boundary, doesn't count against Coordinating category (E) on the OTS. Instead count as Required Coordination on Remaining Actions form.

For specific circumstances when pilots aren't going to meet crossing restrictions, if that is not coordinated, it will be counted as an LOA violation and coordination error.

If participant has an LOA/Directive Violation, also mark as coordination error. If mark several violations, consider marking down Coordination and overall categories.

Managing Sector Workload

If participant doesn't meet TMU in-trail restriction, count under G (Managing Sector Workload).

Prediction of Subjective Ratings of Air Traffic Controller Performance by Computer-Derived Measures and Behavioral Observations

Carol A. Manning[1]
Scott H. Mills[1]
Henry J. Mogilka[2]
Jerry W. Hedge[3]
Kenneth W. Bruskiewicz[3]
Elaine M. Pfleiderer[1]

[1]Federal Aviation Administration
Civil Aeromedical Institute

[2]Federal Aviation Administration
FAA Academy

[3]Personal Decisions Research Institutes, Inc.

Introduction

Performance measures of various types have been developed for air traffic control specialists (ATCSs) during the past 30 years (Buckley, DeBaryshe, Hitchner, & Kohn, 1983; Broach & Manning, 1998; Manning & Heil, 1998). These measures were used for different purposes, including evaluating performance in simulation or on-the-job training, assessing performance in experimental simulations, providing criterion measures against which to validate selection procedures, and comparing baseline ATCS performance with performance resulting from the use of new air traffic control (ATC) procedures or technologies (Albright, Truitt, Barile, Vortac, & Manning, 1995; Vortac, Barile, Albright, Truitt, Manning, & Bain, 1996; Borman et al., 1999; Galushka, Frederick, Mogford, & Krois, 1995; "Flight Strip Reduction Task Force Report," 1998).

The dynamic, and yet cognitive, nature of ATC makes it difficult to measure ATCS performance. The continuing movement of individual aircraft and the constant change of the overall traffic situation make it inappropriate to evaluate discrete activity snapshots. Instead, ATC performance measures should take into account relevant activities that occur during a segment of time.

It is also difficult to measure ATCS performance because controllers often use different approaches to resolving air traffic problems. For example, some controllers may prevent two aircraft from being in conflict by changing the speed of one or both aircraft, whereas others may change their altitude or heading. Furthermore, some controllers may take no immediate action until the situation progresses further. Individual controllers may also sequence aircraft using different orderings. The use of such different approaches may occur because controllers are encouraged to utilize their own techniques to accomplish objectives. In research studies where all controller participants begin a simulated ATC scenario with a set of aircraft in the same configuration, the consequence of using distinct approaches to control traffic is that aircraft will end up in very dissimilar locations when the scenario is finished. In addition, because separation between aircraft is almost always maintained in ATC simulation studies, it is often difficult for observers to evaluate the relative effectiveness of different actions taken by controllers to accomplish the task of moving a set of aircraft through a sector.

A third reason it is difficult to measure ATCS performance is that the observable actions made by a controller reflect only part of the activity that occurs. Considerable cognitive processing also takes place that cannot be directly observed or measured. For example, controllers constantly review aircraft positions, directions, and speeds but take an observable action only when they need to make a change in the traffic flow. The cognitive effort involved in both evaluating aircraft separation and maintaining effective and efficient air traffic flow is difficult to measure directly. Although activities such as keyboard entries, which are made to update, obtain, or highlight information, and use of flight progress strips (placing in strip bay, sorting, marking, removing) can be counted, measured, or otherwise evaluated, they give little indication of actual cognitive effort.

The occurrence of certain outcomes (e.g., loss of separation) can also be determined, though those outcomes typically occur infrequently.

One method often used to evaluate controller performance is having Full Performance Level (FPL) Subject Matter Expert (SME) controllers rate the performance effectiveness of other controllers who control traffic in a live or simulated environment. These SMEs provide over-the-shoulder (OTS) ratings to evaluate performance in training, on the job, or in experiments involving simulations. OTS ratings are often recorded using rating forms that record frequency of specific events as well as subjective ratings of control judgment, situation awareness, and the effectiveness of activities performed.

Several problems are encountered when using OTS ratings to evaluate ATCS performance effectiveness. First, extensive rater training is required to establish and ensure the reliability and accuracy of OTS SME ratings, as well as to reduce their bias and subjectivity (Borman et al., 1999; Sollenberger, Stein, & Gromelski, 1997). Second, a substantial time commitment is required from the raters to familiarize themselves with the scenarios and the rating process, and also to provide them with sufficient practice in making OTS ratings. Third, it is difficult to ensure that sufficient numbers of SMEs will be available to provide OTS ratings for research studies, especially when a considerable amount of time is required to train them to make ratings.

Moreover, if SMEs making the observations and ratings do not constantly attend to the process, they can easily fail to observe the occurrence of events that would influence their ratings. For example, Manning, Mills, Mogilka, & Pfleiderer (1998) found that raters failed to identify some operational errors that were determined by a computer analysis of recorded simulation data. Once a simulation is completed, because of limitations in the available hardware and software it is usually not possible to re-create the scenario in sufficient detail to allow an SME to review available recordings and re-evaluate performance. Thus, the only way to measure the reliability of OTS ratings is to have two raters observe a controller's performance at the same time. Obtaining reliability data, therefore, necessarily requires running either half the number of participants or using twice as many raters as would be needed for a typical session.

Because of the subjective and time-consuming nature of obtaining OTS performance ratings, it would be desirable to measure ATCS performance using other methods that are less difficult. Several types of ATCS performance measures (other than OTS ratings) have been developed. Buckley et al. (1983) identified a set of computer-derived measures that described system functioning during ATC simulations. These measures were grouped into 4 factors: conflict, occupancy, communications, and delay. Galushka, Frederick, Mogford, & Krois (1995) used both counts of controller activities and OTS ratings to assess baseline performance of en route air traffic controllers during a simulation study. Human Technology Inc. (1993) assessed the use of computer-based performance measures in simulation-based training. Computer-derived measures of controller performance and taskload, based on routinely recorded air traffic control data, are being developed by the FAA as part of an ongoing project (Manning, Albright, Mills, Rester, Rodgers, & Vardaman, 1996; Manning, Mills, Albright, Rester, & Pfleiderer, 1997; Mills, 1998).

While some computer-based performance measures have been developed and tested in the ATC environment, their effectiveness, as compared with SME ratings, has not yet been evaluated. The argument may also be made that, because computerized measures are based on observable output alone, they cannot sufficiently describe the cognitive aspects of ATC or the complexity of the traffic situation. In spite of some of the drawbacks encountered when using OTS ratings, if SME raters have been sufficiently trained to rate ATCS performance accurately and reliably, their OTS ratings should be considered the "ultimate" criteria, in the absence of any other ATCS performance measures whose characteristics are better understood.

The purpose of this study was to determine whether alternative methods for measuring ATCS performance could be as effective as OTS ratings. The methods used in the study were a set of computer-derived measures and two types of checklists measuring different aspects of ATCS behavior. Replacing OTS ratings with computer-derived measures would be advantageous because the computerized measures are objective and their collection does not require rater participation. If the computerized measures are found to be as effective as OTS ratings, then it would not be necessary to have SME raters present to rate ATCS performance during simulations. If, however, the computer-derived measures are not sufficient to describe ATCS performance, then they might be supplemented by having SMEs complete behavioral checklists. Asking SME raters to complete behavioral checklists may be more effective than asking them to use OTS ratings because 1) SMEs should require less rater training for using behavioral checklists than for making OTS ratings, 2) making a judgment about whether or not an event occurred may be more accurate and reliable than making a subjective judgment about controller performance effectiveness, and 3) making a judgment about whether or not an event occurred may be easier for an SME to accomplish than providing a subjective rating.

Also considered in this study was whether the two types of behavioral checklists were redundant. If the checklists were redundant, then the more reliable and/or more efficient checklist could be retained for use and the other eliminated.

Method

This study was conducted using data collected in support of the Air Traffic Selection and Training (AT-SAT) High-Fidelity Validation Study (Borman et al., 1999). The study collected performance data from a limited subset of ATCS participants drawn from a much larger sample who participated in the AT-SAT Concurrent Validation Study. The Concurrent Validation Study was conducted to assess the validity of a new ATCS selection procedure by correlating scores from a set of cognitive predictor tests with performance on two "medium-fidelity" criterion measures. The High-Fidelity Validation Study was conducted to assess the validity of the medium-fidelity criterion measures by correlating scores from those measures with a set of other performance measures derived from 7 ATC "graded" simulations. The study was conducted from June through July 1997 at the FAA Academy's En Route Radar Training Facility in Oklahoma City, OK.

Participants

One hundred seven controllers participated in the High-Fidelity Validation Study. All were FPL controllers from 14 FAA en route facilities that provided volunteers for the AT-SAT Concurrent Validation Study. The participants were either currently active controllers or worked in supervisory or staff positions that required them to maintain their job currency (work traffic operationally for a specific number of hours per month). All had previously taken the AT-SAT predictor battery and had completed the 2 medium-fidelity criterion measures.

Raters

Fourteen SME raters were trained to observe the performance of participants during the scenarios. The raters were either operational controllers, controllers assigned to staff, supervisory, or management duties at en route field facilities, or FAA or contract instructors assigned to work at the FAA Academy. The raters averaged 42.2 years of age (*SD* = 7.8) and had been FAA controllers for an average of 16.8 years (*SD* = 6.5).

Raters observed the ATCS participants as they controlled traffic during 8 practice, 7 graded, and 2 part-task scenarios in a high-fidelity simulation environment. After each of the graded scenarios, raters completed 3 forms: The OTS Rating Form, the Behavioral and Event Checklist (BEC) and the Remaining Actions Form (RAF). These measures will be described below. The raters developed and used a Standardization Guide to document their agreement on a set of rules for assessing controller performance when using each of the forms. They also went through a rigorous 3-week training process to ensure the reliability of their responses.

Simulation capability

This study was conducted at the Radar Training Facility (RTF) En Route Simulation Laboratories located at the FAA Academy. Each laboratory contains 10 radar positions controlled by a Digital Equipment Corporation mainframe computer. The radar positions include the same equipment used operationally at en route facilities.

Scenarios

The practice and graded scenarios used in the study were customized from Academy training scenarios. Events were built into the scenarios that included, to the extent possible, the 40 most critical activities performed by en route controllers. For this study, only data for the 7th graded scenario were used, because this scenario was designed to be the most difficult and resulted in the highest frequency of operational errors.

Airspace

The airspace used for this study was the Academy's Tulsa Sector, a low altitude, fictitous sector. Tulsa Sector was created to allow trainees to control traffic in a variety of air traffic situations, including arrivals, departures, and overflights. It contains one Terminal Radar Approach Control (TRACON), one controlled airport that is not a TRACON (McAlester), and one uncontrolled airport. Tulsa Sector was used in the En Route Radar Training Course, which some participants took between 5 and 15 years previously. Prior to coming to Oklahoma City for the study, participants were sent materials describing the airspace. On the first day of the study, they were also provided with a briefing on the airspace. Before beginning the scenarios, they had to pass a test containing multiple choice and completion questions to demonstrate their familiarity with the airspace. Those who did not pass the airspace test the first time repeated it before beginning the graded scenarios. No participant failed the airspace test more than once.

Performance Measures

Over-the-Shoulder (OTS) Rating Form

Each rater (or pair of raters) observed the performance of one participant as he or she ran a graded scenario. Afterwards, raters completed the OTS Rating Form (shown in Figure 1) to evaluate observed performance. The OTS rating form contained 7 specific performance categories (*Maintaining separation; Maintaining efficient air traffic flow; Maintaining attention and situation awareness; Communicating clearly, accurately, and efficiently; Coordinating; Performing multiple tasks; Managing sector workload*) and one *Overall performance* category. The development of the OTS Rating Form is described in Borman et al. (1999).

Computer-Derived Measures

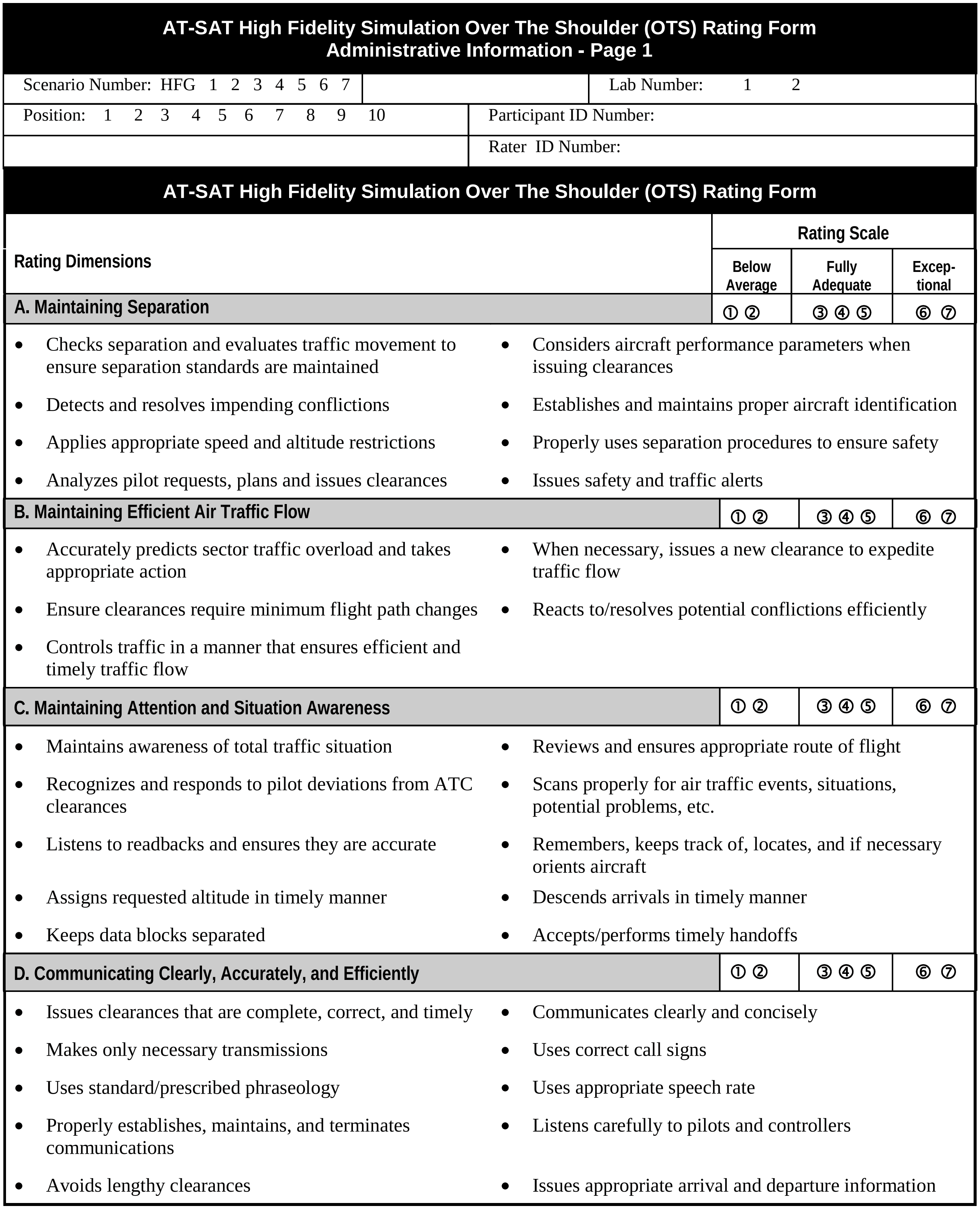
The computer-derived measures for the ATC simulator were based on measures developed for the POWER project, which derived measures of controller performance and taskload from routinely recorded air traffic control data (Manning, Albright, Mills, Rester, Rodgers, & Vardaman, 1996; Manning, Mills, Albright, Rester, & Pfleiderer, 1997; Mills, 1997.) While the POWER measures were originally derived from operational data, a set of comparable measures can be computed from simulation data.

Processing of recorded data. As each scenario was run in the RTF laboratory, simulation software recorded the positions of all aircraft. Aircraft positions were updated on the display at 12-second intervals but were recorded by the computer only at 1-minute intervals to maximize system performance. The range and bearing of each aircraft were recorded in relation to the navigational aid in the Tulsa sector closest to the aircraft. Each aircraft's altitude was also recorded. All computer entries made by controllers, ghost pilots, or remotes (personnel who simulated the activities of other controllers during the scenario) were also recorded and time-stamped. After the complete set of scenarios was finished, summary files containing data for participants who ran scenarios concurrently were transferred to a computer at the Civil Aeromedical Institute (CAMI), checked for accuracy, and then separated by participant and scenario.

SIMSTAT software. Additional processing was conducted on the simulation data using a software package called SIMSTAT (Mills, unpublished manuscript). SIMSTAT converted aircraft positions, originally recorded in reference to multiple navigational aids located in Tulsa Sector, to locations plotted in a single x-, y-coordinate system. The conversion allowed computation of the following statistics for each aircraft: number of heading, speed, and altitude changes, distance between aircraft pairs, number of actions taken by controllers to highlight or obtain information, and others (see Appendix A for a complete list). For each scenario, a matrix of performance measures was computed for all aircraft combined and for subsets of similar aircraft. A separate matrix was computed for each participant in each graded scenario.

Behavioral & Event Checklist (BEC)

The Behavioral & Event Checklist (shown in Figure 2) was used by the raters to record specific types of errors made by a participant during a scenario. Raters marked each error as it occurred and then summed the number of errors when the scenario was finished. When a controller committed an operational error (OE, which occurs when a controller allows an aircraft to come too close to another aircraft) or operational deviation (OD, which occurs when a controller allows an aircraft to enter another controller's airspace without prior authorization), raters recorded additional information such as the identity of the aircraft involved and a brief description of the circumstances. Special Use Airspace (SUA) violations (i.e., violations of military or other special use areas) were counted as ODs. Other mistakes recorded on the BEC were a controller's failure to accept a



AT-SAT High Fidelity Simulation Over The Shoulder (OTS) Rating Form
Administrative Information - Page 1

Scenario Number: HFG 1 2 3 4 5 6 7		Lab Number: 1 2
Position: 1 2 3 4 5 6 7 8 9 10	Participant ID Number:	
	Rater ID Number:	

AT-SAT High Fidelity Simulation Over The Shoulder (OTS) Rating Form

Rating Dimensions	Rating Scale		
	Below Average	Fully Adequate	Excep-tional
A. Maintaining Separation	① ②	③ ④ ⑤	⑥ ⑦

- Checks separation and evaluates traffic movement to ensure separation standards are maintained
- Detects and resolves impending conflictions
- Applies appropriate speed and altitude restrictions
- Analyzes pilot requests, plans and issues clearances
- Considers aircraft performance parameters when issuing clearances
- Establishes and maintains proper aircraft identification
- Properly uses separation procedures to ensure safety
- Issues safety and traffic alerts

B. Maintaining Efficient Air Traffic Flow	① ②	③ ④ ⑤	⑥ ⑦

- Accurately predicts sector traffic overload and takes appropriate action
- Ensure clearances require minimum flight path changes
- Controls traffic in a manner that ensures efficient and timely traffic flow
- When necessary, issues a new clearance to expedite traffic flow
- Reacts to/resolves potential conflictions efficiently

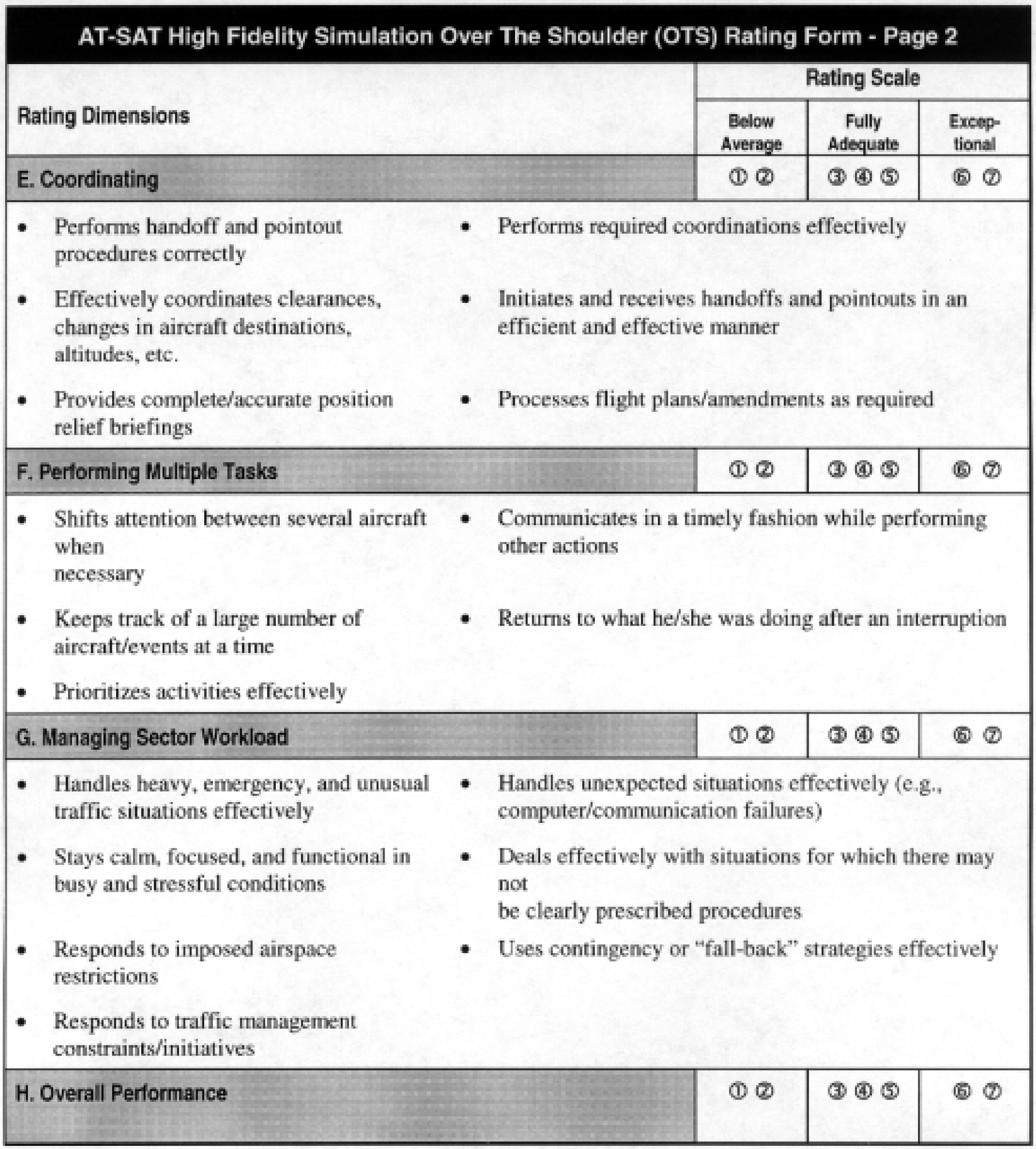
C. Maintaining Attention and Situation Awareness	① ②	③ ④ ⑤	⑥ ⑦

- Maintains awareness of total traffic situation
- Recognizes and responds to pilot deviations from ATC clearances
- Listens to readbacks and ensures they are accurate
- Assigns requested altitude in timely manner
- Keeps data blocks separated
- Reviews and ensures appropriate route of flight
- Scans properly for air traffic events, situations, potential problems, etc.
- Remembers, keeps track of, locates, and if necessary orients aircraft
- Descends arrivals in timely manner
- Accepts/performs timely handoffs

D. Communicating Clearly, Accurately, and Efficiently	① ②	③ ④ ⑤	⑥ ⑦

- Issues clearances that are complete, correct, and timely
- Makes only necessary transmissions
- Uses standard/prescribed phraseology
- Properly establishes, maintains, and terminates communications
- Avoids lengthy clearances
- Communicates clearly and concisely
- Uses correct call signs
- Uses appropriate speech rate
- Listens carefully to pilots and controllers
- Issues appropriate arrival and departure information

Figure 1. Over-the-shoulder rating form used in AT-SAT High-Fidelity Validation Study.



AT-SAT High Fidelity Simulation Over The Shoulder (OTS) Rating Form - Page 2

Rating Dimensions	Rating Scale		
	Below Average	Fully Adequate	Excep-tional
E. Coordinating	① ②	③ ④ ⑤	⑥ ⑦

- Performs handoff and pointout procedures correctly
- Effectively coordinates clearances, changes in aircraft destinations, altitudes, etc.
- Provides complete/accurate position relief briefings
- Performs required coordinations effectively
- Initiates and receives handoffs and pointouts in an efficient and effective manner
- Processes flight plans/amendments as required

F. Performing Multiple Tasks	① ②	③ ④ ⑤	⑥ ⑦

- Shifts attention between several aircraft when necessary
- Keeps track of a large number of aircraft/events at a time
- Prioritizes activities effectively
- Communicates in a timely fashion while performing other actions
- Returns to what he/she was doing after an interruption

G. Managing Sector Workload	① ②	③ ④ ⑤	⑥ ⑦

- Handles heavy, emergency, and unusual traffic situations effectively
- Stays calm, focused, and functional in busy and stressful conditions
- Responds to imposed airspace restrictions
- Responds to traffic management constraints/initiatives
- Handles unexpected situations effectively (e.g., computer/communication failures)
- Deals effectively with situations for which there may not be clearly prescribed procedures
- Uses contingency or “fall-back” strategies effectively

H. Overall Performance	① ②	③ ④ ⑤	⑥ ⑦

Figure 1. Over-the-shoulder rating form used in AT-SAT High-Fidelity Validation Study.

HFG7

Behavioral and Event Checklist

Event	Aircraft identity	Totals
Operational Errors (Write both call signs in one box)	5.	
1.	6.	
2.	7.	
3.	8.	
4.	9.	
Operational Deviations/SUA violations (Write call sign in each box)	5.	
1.	6.	
2.	7.	
3.	8.	
4.	9.	
Behavior	**Number of events**	**Totals**
Failed to accept handoff		
LOA/Directive Violations		
Readback/Hearback errors		
Failed to accommodate pilot request		
Made late frequency change		
Unnecessary delays		
Incorrect information in computer		
Fail to issue weather information		

Participant ID Number:	Rater ID Number:
Lab Number:	Position Number:

Rev. Date: 5/29/97

Figure 2. Behavioral and Event Checklist for the 7th graded scenario in AT-SAT High-Fidelity Validation Study.

handoff, violations of Letters of Agreement (LOAs) or other directives, readback and hearback errors (failure to repeat accurate information to a pilot or failure to hear that a pilot has not accurately read back information in a clearance), failure to accommodate a pilot request, making a late frequency change, unnecessarily delaying an aircraft, entering incorrect information in the computer, and failing to issue weather information to a pilot arriving at an uncontrolled airport. The development of the BEC is described in more detail in Borman et al. (1999).

Remaining Actions Form (RAF)

Figure 3 shows the Remaining Actions Form. The RAF is used to measure the number of control actions left to be completed for each aircraft at the end of the scenario. Because all controllers started with the same number of aircraft in the same configuration and ended the scenario at the same time, the number of actions remaining to be performed can be considered an indicator of the efficiency of the controllers' actions, with number of actions remaining inversely related to efficiency (Vortac, Edwards, Fuller, & Manning, 1993). The RAF has been used in several studies (Vortac, Edwards, Fuller, & Manning, 1993; Albright, Truitt, Barile, Vortac, & Manning, 1995; Vortac, Barile, Albright, Truitt, Manning, & Bain, 1996; Durso, Hackworth, Truitt, Crutchfield, Nikolic, & Manning, 1998). The version of the RAF used in this study was modified to include several additional remaining actions. The actions evaluated were: ***take handoff/pointout, make handoff/pointout, change frequency, perform required coordination, assign requested altitude, issue speed restriction, issue additional required routing, issue approach clearance, issue departure clearance,*** and ***issue weather information***. If all control actions had been completed for an aircraft at the end of a scenario, the rater indicated that no remaining actions were required.

Procedure

Simulation testing lasted for 2 ½ days for each participant. Groups of either 6 or 12 controllers participated concurrently in each simulation test. In advance of their arrival in Oklahoma City, participants were provided with a map of the fictional airspace, Tulsa Sector, used for the scenarios. Upon arrival, each received a briefing on the airspace structure (airways, navigational aids, airports, SUAs) and procedures used in Tulsa Sector. After the briefing, participants ran 8 practice scenarios, 7 graded scenarios, and 2 part-task scenarios. Each scenario lasted 30 minutes. Participants ran all scenarios as single-person sectors, rather than operating as members of a controller team.

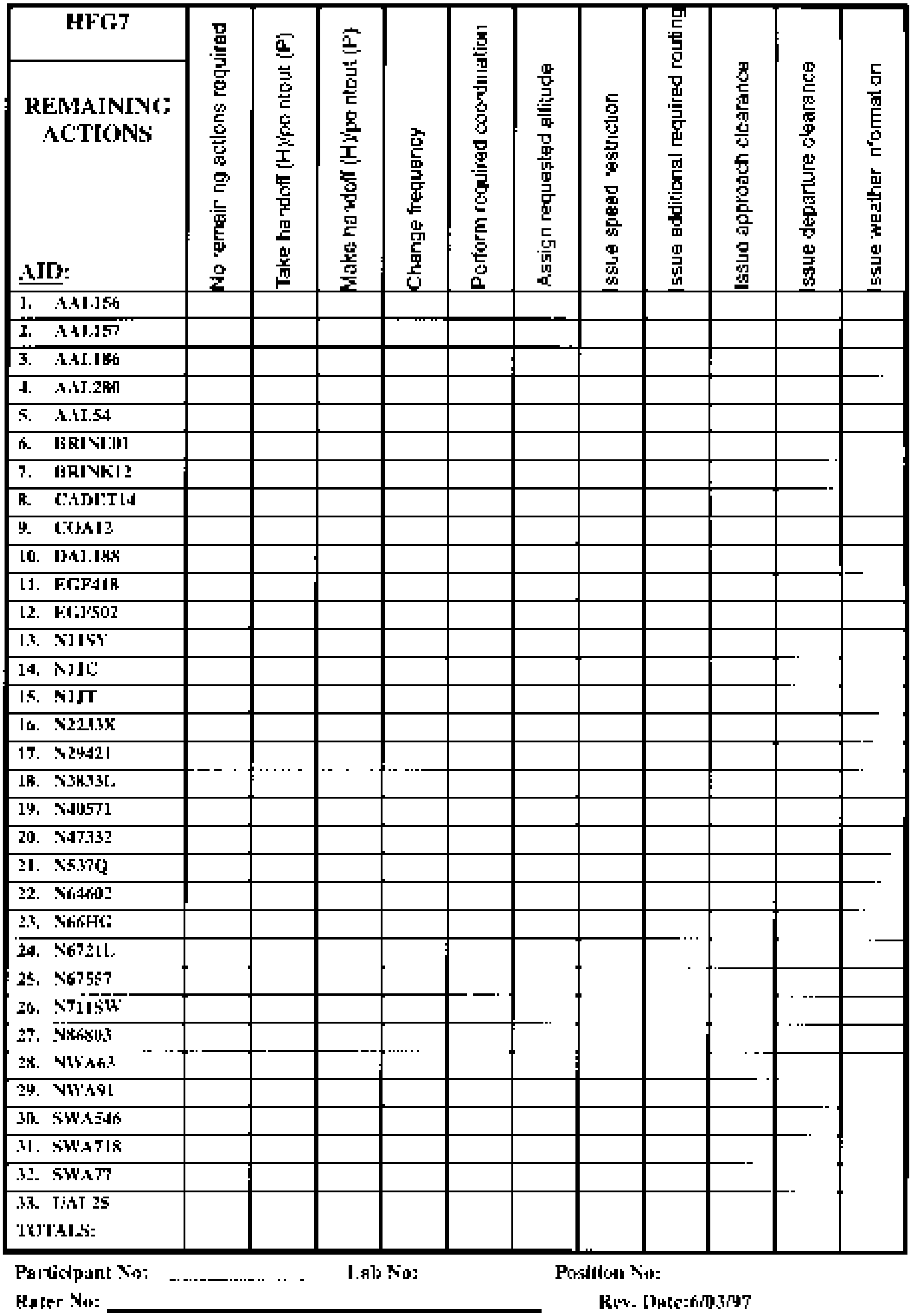
Choice of measures for analysis

Not all available ATCS performance measures were analyzed. The number analyzed was reduced for several reasons. First, the number of possible measures far exceeded the number of participants, thus producing a meaningless solution. Second, some measures appeared to duplicate others. For example, the OTS rating scales, when averaged across all scenarios, were highly correlated, with intercorrelations ranging from .80 to .97 (Manning & Heil, 1999). The Overall Performance rating had an intraclass correlation of .95, and thus, was considered a reasonable criterion measure against which to validate other ATCS performance measures.

Third, it appeared that some measures did not accurately describe the activity they were designed to measure. For example, some of the categories included in the behavioral checklist included tasks that some controllers rarely perform (e.g., issuing weather information for arrivals at uncontrolled airports, issuing arrival or departure clearances for aircraft at uncontrolled airports, issuing additional required routing, vectoring for the Instrument Landing System). It was determined that, to compare participants fairly, only the tasks they regularly perform should be included in the evaluation.

Other performance measures were considered artificial because the way the corresponding tasks were performed during the simulation was different than would have occurred in reality (e.g., accepting handoffs and pointouts, performing required coordination, issuing frequency changes, response to violation of LOAs or directives). Still other performance measures were eliminated because the raters could not reliably identify them (e.g., occurrence of unnecessary delays).

The number of computer-derived measures retained for analysis was also reduced. Tulsa arrivals were excluded from the OE count because it was previously determined that the software incorrectly identified some OEs that occurred at the boundary between the en route Tulsa Sector and Tulsa approach



HFG7 REMAINING ACTIONS AID:	No remaining actions required	Take handoff (H)/pointout (P)	Make handoff (H)/pointout (P)	Change frequency	Perform required coordination	Assign requested altitude	Issue speed restriction	Issue additional required routing	Issue approach clearance	Issue departure clearance	Issue weather information
1. AAL356											
2. AAL357											
3. AAL186											
4. AAL280											
5. AAL54											
6. BRINE01											
7. BRINK12											
8. CADET14											
9. COA12											
10. DAL188											
11. EGF418											
12. EGF502											
13. N11SY											
14. N11C											
15. N1JT											
16. N2233X											
17. N29421											
18. N3833L											
19. N40571											
20. N47332											
21. N537Q											
22. N64602											
23. N66HG											
24. N6721L											
25. N67557											
26. N711SW											
27. N86803											
28. NWA63											
29. NWA91											
30. SWA546											
31. SWA718											
32. SWA77											
33. UAL25											
TOTALS:											

Participant No: __________ Lab No: Position No:

Rater No: ______________________________ Rev. Date:6/03/97

Figure 3. Remaining Actions Form for 7th graded scenario in AT-SAT High-Fidelity Validation Study.

control (Manning, Mills, Mogilka, & Pfleiderer, 1998). Other computer-derived measures (e.g., aircraft in hold) were excluded because they occurred infrequently.

Results

Descriptive statistics

Descriptive statistics for the set of ATCS performance measures selected for analysis are shown in Table 1. These measures were computed for 104 of the 107 participants who had complete data. The mean Overall Performance rating was fairly low, falling on the low end of the Fully Adequate rating category (on the OTS rating form shown in Figure 3). Mean numbers of mistakes recorded on the BEC were also fairly low (none exceeded 2.0). Standard deviations for most of the behavioral checklist variables were typically about as high or higher than the means.

Table 2 shows the intercorrelation matrix for the ATCS performance measures. Several correlations were statistically significant. The correlations between the Rater OE and OD counts and the Overall Performance rating were negative and statistically significant, and accounted for between 25 and 15%, respectively, of the variance in the rating. Correlations between the other BEC items and the Overall Performance rating were also all negative and statistically significant. This result is in the expected direction (i.e., fewer BEC errors = a higher overall rating) and because the BEC items were considered by the raters when they completed the OTS rating form.

Table 1

Descriptive statistics for ATCS performance measures (N=104).

Name of measure	**Mean**	**Standard Deviation**
OTS Rating Scales		
Overall performance	3.07	1.32
Behavioral & Event Checklist		
Rater count of OEs	0.78	0.84
Rater count of ODs	1.30	1.40
N readback/hearback errors	0.76	1.34
Failed to accommodate pilot requests	1.08	1.43
Made late frequency change	1.00	1.26
Entered incorrect information in computer	1.60	1.55
Remaining actions		
N aircraft with no remaining actions	20.27	2.98
N requested altitude assignments remaining	1.62	1.33
N handoffs, pointouts to be made	8.46	2.10
N speed restrictions remaining	0.20	0.94
Computer-derived measures		
N entries – all aircraft	195.27	45.62
N entry errors – all aircraft	9.86	8.68
N heading changes – Tulsa arrivals	10.28	5.03
N altitude changes – Tulsa arrivals	12.65	3.66
All aircraft OEs excluding Tulsa arrivals	0.12	0.38

Table 2

Intercorrelation matrix for ATCS performance measures.

	OPR	ROE	ROD	RHE	FAPR	MLFC	EIIC	NRA	RAAR	MHPR	SSR	Ent	Err	HdC	AlC
High fidelity simulation performance measures															
OTS Rating Scales															
Overall performance rating (OPR)															
BEC measures															
Rater count of OEs (ROE)	-.49**														
Rater count of ODs (ROD)	-.39**	.04													
Number readback/hearback errors (RHE)	-.31**	.17	.35**												
Fail to accommodate pilot requests (FAPR)	-.33**	.07	.25**	.12											
Made late frequency change (MLFC)	-.35**	.23*	.34**	.40**	.20*										
Entered incorrect information in computer (EIIC)	-.23**	.16	.32**	.23*	.34**	.34**									
RAF measures															
Number aircraft with no remaining actions (NRA)	.21*	-.08	-.33**	-.09	-.18	-.26**	-.26**								
N requested altitude assignments remaining (RAAR)	-.19*	.11	.38**	.27**	.25*	.26**	.23*	-.22*							
N handoffs, pointouts to be made (MHPR)	-.08	-.11	.23*	.13	.15	.17	.16	-.72**	.26**						
N speed restrictions remaining (SRR)	-.05	.03	.03	-.02	.05	.02	.05	-.11	.05	.11					
Computer-derived measures															
N entries – all aircraft (Ent)	.27**	-.03	-.21*	-.08	-.16	-.15	-.14	.25*	-.18	-.30**	-.06				
N errors – all aircraft (Err)	.00	-.03	.04	.37**	.12	.21*	.20*	.00	.05	.08	-.03	.30**			
N heading changes – Tulsa arrivals (HdC)	-.11	-.19	.14	.07	.05	-.07	.08	-.07	.04	.10	-.04	.06	-.06		
N altitude changes – Tulsa arrivals (AlC)	.08	-.02	.04	.14	-.04	.05	-.01	.05	.04	-.09	-.03	.36**	.15	.37**	
N OEs all aircraft except Tulsa arrivals (COE)	.02	.14	-.05	-.02	-.22*	-.00	-.05	.12	-.15	-.23*	-.07	.12	-.14	-.13	.02

* indicates correlation significant at $p < .05$.

** indicates correlation significant at $p < .01$.

The number of aircraft with no remaining actions had a positive, significant correlation with the Overall Performance rating (.21). Fewer remaining actions should be related to efficiency of control, and efficiency should be a component of the overall rating. The number of computer entries made also had a positive, significant correlation with the Overall Performance rating (.27).

The correlation between the number of aircraft with no remaining actions and the number of handoffs and pointouts to be made was -.72. This finding is not unexpected because most of the actions remaining at the end of the scenario involved making handoffs and pointouts (see Table 1.)

Of all the performance measures, only the Made Late Frequency Change count was significantly correlated (r = .23) with the rater count of OEs. However, a number of the BEC and Remaining Actions measures were significantly correlated with the rater count of ODs. While it might be expected that the computer count of OEs would be significantly correlated with the Rater count of OEs, the raters' OE count included those that occurred in a non-radar environment, which could not be identified by the computer (Manning et al., 1998).

Regression Analyses

The regression analyses examined questions concerning whether the computer-derived performance measures and 2 behavioral checklists could sufficiently account for the variance in the subjective Overall Performance rating made by a set of trained raters. An initial multiple regression analysis was conducted to determine whether the multiple correlation between the set of all predictor variables and the dependent rating variable was significantly different from zero, and whether the complete set of predictor variables appeared to account sufficiently for the variance in the dependent variable. The results of this analysis are shown in the first line of Table 3. The multiple correlation between a model containing all predictor variables and the criterion measure was .71, which was significantly different from 0 ($F(15, 88) = 5.98$, $p < .001$.) The complete set of predictor variables accounted for just over 50% of the variance in the Overall Performance rating.

It was determined that the multiple correlation between the predictors and the criterion measure was sufficiently large to consider further the possibility of replacing the Overall Performance rating with other

Table 3
Results of model comparison regression analyses.

Regression Model Tested	*R*	R^2	*F* for model comparison	*df*	*p*
Full model containing all performance measures	.71	.51	5.98	15, 88	< .001
Reduced models					
Computer-derived measures only	.32	.10	7.16	10, 88	< .001
Computer-derived measures and BEC measures	.71	.50	0.31	4, 88	.67
Computer-derived measures and RAF measures	.41	.17	9.96	6, 88	< .001
BEC measures only	.66	.43	1.38	9, 88	.45
OEs, ODs only	.61	.37	1.79	13, 88	.17
OEs, ODs, and Computer entries	.64	.41	1.45	12, 88	.41
OEs, ODs, and N aircraft w/ no remaining actions	.61	.38	1.90	12, 88	.13

types of performance measures. Another set of analyses was conducted to determine whether any other regression models containing fewer predictor variables could be identified that were as effective as the full regression model in predicting the Overall Performance rating. These analyses assessed the effectiveness of several "reduced" regression models, containing fewer than the complete set of predictor variables, in predicting the criterion or dependent measure, as compared with the "full" model containing all predictor variables. Each comparison produces an *F* statistic. A statistically significant *F* statistic reflects a significant difference in the predictability of the two regression models, indicating that the reduced model does not predict the dependent variable as well as does the full model. If, on the other hand, the *F* statistic is not statistically significant, then there is no significant difference in the predictability of the 2 regression models, indicating that, in a statistical sense, the reduced model predicts the dependent variable as well as does the full model.

The first analysis considered whether a regression model containing only the computer-derived measures would predict the Overall Performance rating as well as a regression model containing all 3 types of performance measures. The results of this analysis are shown in the second line of Table 3. The full regression model containing the complete set of computer-derived, BEC, and remaining actions performance measures predicted the Overall Performance rating (R=.71) better than did a reduced regression model containing only the computer-derived variables (R=.32; $F(10, 88)$= 7.16, $p < .001$). This result suggests that the computer-derived measures alone cannot predict the Overall Performance rating as well as the full model.

A second analysis was conducted to assess whether having raters complete the BEC, in addition to using the computer-derived measures, would be sufficient to predict the Overall Performance rating. As shown in Table 3, the full model containing all the variables predicted the Overall Performance rating no better than did the reduced model containing only the computer-derived measures and the BEC (R=.71; $F(4, 88)$ = .31, p = .67.) This result suggests that using both the BEC and computer-derived measures can predict the Overall Performance rating as well as does the full model containing all the predictor variables.

A third analysis examined whether having raters complete the RAF, in addition to the computer-derived measures (but instead of the BEC), would be sufficient to predict the Overall Performance rating. As shown in Table 3, the full model predicted the Overall Performance rating significantly better than did the reduced model containing only the computer-derived measures and the RAF measures (R=.41; $F(6, 88)$ = 9.96, $p < .001$.) This result suggests that using only the RAF and computer-derived measures cannot predict the Overall Performance rating as well as the full model containing all the predictor variables.

Because the computer-derived measures were insufficient alone, and in combination with the RAF measures to predict the Overall Performance rating, the influence of the BEC measures was investigated next. The fourth analysis investigated whether the BEC alone would be a sufficient replacement for the Overall Performance rating. Table 3 shows that the full model predicted the Overall Performance rating no better than did a reduced model containing only the BEC (R=.66; $F(9, 88)$ = 1.38, p = .45). This result suggests that the BEC alone can predict the Overall Performance rating as well as the full model containing all the predictor variables.

The next set of analyses investigated whether subsets of the predictor variables would be as effective as entire sets of measures in predicting the Overall Performance rating. If the BEC alone were as effective as the full model, then perhaps the rater OE and OD counts alone would also be effective predictors. A model containing only the rater OE and OD counts was considered first. As shown in Table 3, the full model predicted the Overall Performance rating no better than did the reduced model containing only OE and OD counts (R = .61; $F(13,88)$ = 1.79, p = .17.) This result suggests that OE and OD counts alone were sufficient to predict the Overall Performance rating.

Although OEs and ODs were sufficient to predict the dependent variable, they accounted for less than 40% of the variance in the dependent variable. Perhaps the addition of a different type of measure to the OE and OD counts would predict a higher percentage of the variance in the Overall Performance rating. A model containing rater OE and OD counts, along with the number of computer entries (1 of the computer-derived measures) was considered next. Table 3 shows that the full model predicted the Overall Performance rating no better than did the reduced model containing OEs, ODs, and computer entries (R = .64; $F(12,88)$ = 1.45, p = .41.) This result suggests that a model including OEs, ODs, and number of computer entries is statistically equivalent to the full model in predicting the Overall Performance rating. Moreover, adding computer entries to the model containing OEs and ODs increased to about 41% the percentage of variance accounted for in the dependent variable.

An alternative model, containing rater OE and OD counts, along with the number of aircraft with no remaining actions (from the RAF) instead of the number of computer entries, was considered. Table 3 shows that the full model predicted the Overall Performance rating no better than did the reduced model containing OEs, ODs, and number of aircraft with no remaining actions ($R = .61; F(12,88) = 1.90, p = .13.$) This suggests that a model including OEs, ODs, and number of aircraft with no remaining actions required is statistically equivalent to the full model in predicting the Overall Performance rating. However, adding the number of aircraft with no remaining actions to the model did not account for a higher percentage of the variance in the dependent variable.

Discussion & Conclusions

It was determined that a full regression model containing three types of performance measures had a multiple correlation of more than .70, accounting for greater than 50% of the variance in the Overall Performance rating. This set of measures was considered sufficient to replace subjective performance ratings as ATC performance measures. A series of model comparison analyses was then conducted that yielded a set of regression models, accounting for the variance in the Overall Performance rating, as well as the full model containing all the predictor variables.

The results of the model comparison analyses showed that using the BEC measures alone produced a model equivalent to the full model containing all the predictor variables in predicting the Overall Performance rating. In addition, regression models containing rater OE and OD counts alone, OEs and ODs along with the number of computer entries made, and OEs, ODs, and number of aircraft with no remaining actions were sufficient to predict the Overall Performance rating.

Clearly, the BEC produces values that are the most similar to the Overall Performance rating. This is understandable because the same raters completed both the BEC and the OTS rating form, from which the Overall Performance rating was taken. The rater Standardization Guide used to determine how to make certain ratings specified that the value of the Overall Performance rating would depend on the OE count. Furthermore, the OE and OD counts alone seem to be about as effective as the complete set of BEC measures in predicting the Overall Performance rating.

In comparison, the RAF and its components were not very effective in predicting overall performance. The RAF did not add to the predictability of the computer-derived measures, and also did not add to the predictability of rater OE and OD counts. This result may have occurred because the count of aircraft with no remaining actions was significantly correlated with a number of other performance measures and may add nothing unique to the prediction of overall performance. On the other hand, it may be that, in its present form, the remaining actions measures seem to be primarily defined by the number of handoffs and pointouts left to be made. Perhaps this set of measures needs to be reconsidered in order to measure controller efficiency more effectively.

Although several models were identified that predicted the Overall Performance rating as well as did the full model containing all the measures, some models appeared better than others (as measured by the percentage of variance in the dependent variable accounted for by the model.) The full model accounted for just over 50% of the variance in the Overall Performance rating. The combination of the computer-derived measures and the BEC accounted for about 50% of the variance. The BEC alone accounted for about 44% of the variance, while OE & OD counts, along with the number of computer entries, accounted for about 41% of the variance. Although the latter 2 models were statistically equivalent to the full model, it would be preferable to account for as much of the variance in the dependent measure as possible, while minimizing the amount of data that must be collected. Thus, for the type of ATC simulation used here, it is suggested that a combination of the BEC and computer-derived measures could be used in place of the subjective ratings. Using the BEC would require using trained SME raters, but training them to identify errors would be less complex and time-consuming than training them to assign subjective ratings systematically. The computer-derived measures are easily collected and computed and do not require the participation of trained SMEs.

It is also worth mentioning here, as noted in the "Simulation Capability" section, that we chose to analyze data from Scenario 7, because it was the most difficult scenario and resulted in the highest frequency of operational errors. Consequently, if researchers plan to relay on OE and OD counts, then scenarios must be chosen or developed of a complexity high enough to produce sufficient controller errors.

More research needs to be done on the development of computer-derived measures of performance and workload/taskload. For a variety of reasons discussed elsewhere (Manning, Mills, Mogilka, & Pfleiderer, 1998), the restricted capabilities of the simulator used for this study limited the amount of data that could be collected. An ATC research simulator currently under development will record more variables, with a higher degree of accuracy, than those analyzed here. Other research to develop taskload and performance measures in operational settings may produce different results. Measures derived from operational ATC data do not have the same limitations as measures derived from the simulator and so may be more useful for predicting operational ATC performance.

References

Albright, C. A., Truitt, T. R., Barile, A. L., Vortac, O. U., & Manning, C. A. (1995). Controlling traffic without flight progress strips: Compensation, workload, performance, and opinion. *Air Traffic Control Quarterly*, *2*, 229-248.

Borman, W. C., Hedge, J. W., Hanson, M. A., Bruskiewicz, K. T., Mogilka, H., Manning, C., Bunch, L. B., & Horgen, K. E. (1999). Development of criterion measures of air traffic controller performance. In R. Ramos (Ed.), *AT-SAT Final Report*. Alexandria, VA: HumRRO.

Broach, D. & Manning, C. A. (1998). Issues in the selection of air traffic controllers. In M. W. Smolensky & E. W. Stein (Eds). *Human Factors in Air Traffic Control*. Orlando, FL: Academic Press, pp. 237-271.

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: Federal Aviation Administration Technical Center.

Durso, F.T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly*, *6*, 1-20.

Flight Strip Reduction Task Force Report. (1998, July). Unpublished manuscript.

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.

Human Technology, Inc. (1993, April). *Analysis of human and machine-based assessment strategies for the terminal radar training facility (RTF): Simulator data and performance measures*. (Contract OPM-91-2958). McLean, VA: Author.

Manning, C. A., Albright, C. A., Mills, S. H., Rester, D., Rodgers, M. D., and Vardaman, J. J. (1996, May). Setting parameters for POWER (Performance and Objective Workload Evaluation Research). Poster presented at the 67th Annual Scientific Meeting of the Aerospace Medical Association, Atlanta, Georgia.

Manning, C. A., & Heil, M. C. (1999). The Relationship of FAA archival data to AT-SAT predictor and criterion measures. In R. Ramos (Ed.), *AT-SAT Final Report*. Alexandria, VA: HumRRO.

Manning, C. A., Mills, S. H., Albright, C. A., Rester, D., & Pfleiderer, E. M. (1997, May). Evaluating alternative sources for POWER (Performance and Objective Workload Evaluation Research). Poster presented at the 68th Annual Scientific Meeting of the Aerospace Medical Association, Chicago, IL.

Manning, C. A., Mills, S. H., Mogilka, H. J., & Pfleiderer, E. M. (1998, May). A comparison of rater counts and computer calculations of operational errors made in en route air traffic control simulations. Paper presented at the 69th Annual Scientific Meeting of the Aerospace Medical Association, Seattle, WA.

Mills, S. H. (1997). *SIMSTAT: Data analysis software for the FAA Academy En Route Radar Training Facility Simulator*. Unpublished manuscript.

Mills, S. H. (1998). *The combination of flight count and control time as a new metric of air traffic control activity*. (Report No. DOT/FAA/AM-98/15). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Sollenberger, R. L., Stein, E. S., & Gromelski, S. (1997). *The development and evaluation of a behaviorally-based rating form for assessing air traffic controller performance*. (Report No. DOT/FAA/CT-TN96-16). Atlantic City, NJ: Federal Aviation Administration Technical Center.

Vortac, O.U., Edwards, M.B., Fuller, D. K., & Manning, C. A. (1993). Automation and cognition in air traffic control: An empirical investigation. *Applied Cognitive Psychology*, *7*, 631-651.

Vortac, O.U., Barile, A. B., Albright, C. A., Truitt, T. R., Manning, C. A. & Bain, D. (1996). Automation of flight data in air traffic control. In D. Hermann, C. McEvoy, C. Hertzog, P. Hertel, & M. K. Johnson, (Eds.) *Basic and Applied Memory Research, Volume 2*. Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix A

Line	Group	Aircraft Group	Dependent Measures									
1	0	TUL Arr. Group 1	Time of 1st to Location	Time of Last to Location	Arrival Count	Hdg Change Count	Hdg Change Amount	Alt Change Count	Alt Change Amount	OE Count	OE Duration1	OE Duration2
2	1	TUL Arr. Group 2	Time of 1st to Location	Time of Last to Location	Arrival Count	Hdg Change Count	Hdg Change Amount	Alt Change Count	Alt Change Amount	OE Count	OE Duration1	OE Duration2
3	2	MLC Arrivals	Time of 1st to Location	Time of Last to Location	Arrival Count	Hdg Change Count	Hdg Change Amount	Alt Change Count	Alt Change Amount	OE Count	OE Duration1	OE Duration2
4	3	Specific Aircraft 1			.	Hdg Change Count	Hdg Change Amount	Alt Change Count	Alt Change Amount	OE Count	OE Duration1	OE Duration2
5	4	Overflights	.		.	Hdg Change Count	Hdg Change Amount	Alt Change Count	Alt Change Amount	OE Count	OE Duration1	OE Duration2
6	5	Weather Deviations	.	.	.	Hdg Change Count	Hdg Change Amount	Alt Change Count	Alt Change Amount	OE Count	OE Duration1	OE Duration2
7	6	TUL Departures	Number of Delays		.	.	.			OE Count	OE Duration1	OE Duration2
8	7	All Controlled A/C			.	Hdg Change Count	Hdg Change Amount	Alt Change Count	Alt Change Amount	OE Count	OE Duration1	OE Duration2
9	7	All Controlled A/C	HOLD Count	HOLD Max for any A/C	FPLAN Count	FPLAN Max for A/C	Route-Display Count	Route-Display Max for A/C	.		.	.