technical note technical note

Air Traffic Control System Baseline Methodology Guide

Kenneth R. Allendoerfer, ACT-530
Joseph Galushka, ACT-530

June 1999

DOT/FAA/CT-TN99/15

U.S. Department of Transportation
Federal Aviation Administration

William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

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 or use thereof.

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/CT-TN99/15		

4. Title and Subtitle	5. Report Date
Air Traffic Control System Baseline Methodology Guide	June 1999
	6. Performing Organization Code ACT-530
7. Author(s) Kenneth R. Allendoerfer and Joseph J. Galushka, ACT-530	8. Performing Organization Report No. DOT/FAA/CT-TN99/15
9. Performing Organization Name and Address Federal Aviation Administration William J. Hughes Technical Center Atlantic City International Airport, NJ 08405	10. Work Unit No. (TRAIS) 11. Contract or Grant No. F2202J
12. Sponsoring Agency Name and Address Federal Aviation Administration Human Factors Division 800 Independence Ave., S.W. Washington, DC 20591	13. Type of Report and Period Covered Technical Note 14. Sponsoring Agency Code AAR-100

15. Supplementary Notes

16. Abstract

The *Air Traffic Control System Baseline Methodology Guide* serves as a reference in the design and conduct of baseline studies. Engineering research psychologists are the intended audience for the Methodology Guide, which focuses primarily on techniques for studying the interaction between ATC systems and the controllers who use them. The Methodology Guide provides the following information: (a) descriptions of and references to past baselines that have successfully used the methodology, (b) detailed descriptions of the baseline operational constructs and corresponding objective and subjective measures, (c) a description of the overall baseline methodology, (d) other recommendations and lessons learned regarding the successful conduct of system baselines, and (e) a discussion of the role of system baselines in the ATC system acquisition process.

17. Key Words	18. Distribution Statement
Air Traffic Control Research Methods System Baselines	This document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	86	

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Table of Contents

Table of Contents (Cont.)

Page

Table of Contents (Cont.)

List of Illustrations

Executive Summary

The Federal Aviation Administration (FAA) has sponsored several system baseline studies since 1995. These studies used controlled human-in-the-loop simulations to collect data regarding the operational effectiveness of several major air traffic control (ATC) systems. These data allowed direct comparisons between ATC systems and helped identify deficiencies in new ATC systems. System baseline studies provide data following five operational constructs: safety, capacity, performance, workload, and usability. Each construct comprises objective and subjective measures and provides converging indicators for that construct. In addition, data are collected about the realism of the baseline simulations to ensure their external validity.

The *Air Traffic Control System Baseline Methodology Guide* serves as a reference for engineering research psychologists and others interested in conducting system baselines in the ATC domain. The Methodology Guide provides the following information: (a) descriptions of and references to past baselines that have successfully used the methodology, (b) detailed descriptions of the operational constructs and corresponding objective and subjective measures, (c) a description of the overall baseline methodology, (d) other recommendations and lessons learned regarding the successful conduct of system baselines, and (e) a discussion of the role of system baselines in the ATC system acquisition process.

1. Introduction

Since early 1995, the Federal Aviation Administration (FAA) has sponsored several system baseline studies. These studies collected data on the operational effectiveness of several major air traffic control (ATC) systems under controlled simulation conditions. These data allowed comparisons of operational effectiveness between these ATC systems and identification of areas where new systems needed improvement.

1.1 Background

System baseline studies (hereafter referred to as baselines) are an important component of the human factors evaluation process. These studies collect data in high fidelity, human-in-the-loop simulations of everyday ATC operations. Simulation conditions are tightly controlled to allow comparisons with past and future systems. Only relatively stable systems are suitable for baselines, requiring that baselines be conducted late in the acquisition process.

Baselines provide data following five operational constructs: Safety, Capacity, Performance, Workload, and Usability. Each construct comprises multiple objective and subjective measures, providing converging indicators for that construct. When examined together, the measures provide a thorough description of the system for that construct. In addition, data are collected about the realism of the baseline simulations to ensure their external validity.

The *Air Traffic Control System Baseline Methodology Guide* serves as a reference in the design and conduct of baselines. It focuses primarily on techniques for studying the interaction between ATC systems and the controllers who use them. Engineering research psychologists are the intended audience for the Methodology Guide.

The Methodology Guide provides

a. descriptions of and references to past baselines that have successfully used the methodology described here,

b. detailed descriptions of the baseline operational constructs and corresponding objective and subjective measures (details about how each measure is administered and how the corresponding data are analyzed),

c. a description of the baseline methodology (which is flexible enough to apply to a wide range of ATC systems with only a minimum of modification),

d. other recommendations and lessons learned regarding the successful conduct of system baselines, and

e. a discussion of the role of system baselines in the ATC system acquisition process.

1.1.1 Host Computer System and Plan View Display

In early 1995, the FAA sponsored the first system baseline, which collected data for the Host Computer System (HCS) and the Plan View Display (PVD), the operational equipment currently used in Air Route Traffic Control Centers (ARTCCs). This baseline was conducted at the FAA William J. Hughes Technical Center using en route controllers from Washington ARTCC. The

operational constructs, baseline measures, methodology, and reporting style described here were originally developed for this baseline. The results of this baseline are contained in the *Plan View Display Baseline Research Report* (Galushka, Frederick, Mogford, & Krois, 1995).

1.1.2 Automated Radar Terminal System IIIA and Data Entry and Display Subsystem

In late 1995, the FAA sponsored a baseline study to collect data for the Automated Radar Terminal System (ARTS) IIIA and the Data Entry and Display Subsystem (DEDS), the operational equipment currently used in many Terminal Radar Approach Control (TRACON) facilities. This baseline study was conducted at the Technical Center using terminal controllers from Boston TRACON. The ARTS IIIA Baseline used the constructs, measures, methodology, and reporting style of the PVD Baseline with some modifications for the terminal domain. The results of this baseline are contained in the *ARTS IIIA Terminal Baseline Research Report* (Mogford, Allendoerfer, & Galushka, 1999).

1.1.3 Operational Display and Input Development IV

In 1996, the FAA co-sponsored a baseline study to collect data for the Operational Display and Input Development (ODID) IV system, a Eurocontrol developmental ATC program. This baseline was conducted at the Eurocontrol Experimental Centre using en route controllers and supervisors from a variety of FAA en route facilities. The ODID IV Baseline used the constructs, measures, methodology, and reporting style of the PVD Baseline with some modifications for European ATC operations and the ODID IV hardware and software. The results of this baseline and a comparison of the ODID IV to the HCS-PVD are contained in the *FAA ODID IV: En Route Baseline Comparison Simulation Final Report* (Krois & Marsden, 1997) and the *En Route ODID-PVD Baseline Comparisons* (Skiles, Graham, Marsden, & Krois, 1997).

1.1.4 Display System Replacement

In 1997, the FAA sponsored a baseline study to collect data for the Display System Replacement (DSR). This display system will replace the PVD and its associated consoles throughout 1999 and 2000. The DSR baseline was conducted at the Technical Center using en route controllers from Washington ARTCC. Many of these individuals had participated in the PVD Baseline 2½ years earlier. The DSR Baseline used the operational constructs, suite of measures, general methodology, and reporting style used of the PVD Baseline with some modifications for the DSR hardware and software. The results of this baseline and a comparison of the DSR to the HCS-PVD are contained in the *Comparison of the Plan View Display and Display System Replacement System Baselines* (Allendoerfer, Mogford, & Galushka, 1999).

1.1.5 Standard Terminal Automation Replacement System

In the future, the FAA plans to conduct two baseline studies to collect data for the Standard Terminal Automation Replacement System (STARS), the new TRACON and tower radar display and automation equipment.

The STARS Baselines will use the operational constructs, suite of measures, general methodology, and reporting style used in the ARTS IIIA Baseline with some modifications for the new hardware and software capabilities. Other improvements will be made to the methodology based on lessons learned from earlier baselines.

2. Operational Constructs and Measures

In 1994, the Air Traffic Requirements Organization (now the Air Traffic System Requirements Service [ARS]) identified four high-level operational constructs on which to base evaluations of ATC systems. These constructs were: Safety, Capacity, Performance, and Workload. During preparations for the PVD Baseline, a fifth operational construct, Usability, was added. In addition, a non-operational construct, Simulation Fidelity, was developed to assess the realism and validity of simulation conditions. Throughout the subsequent baselines, the formal definitions of these constructs have been gradually refined. The current definitions are presented as follows:

a. Safety represents the extent to which the system allows aircraft to traverse a section of airspace without a dangerous incident such as a violation of applicable separation minima.

b. Capacity represents the amount of traffic that the system allows to safely and efficiently traverse a section of airspace during a period of time.

c. Performance represents the amount and quality of user interaction with the system.

d. Workload represents the cognitive and physical task demands of the system as experienced by its users.

e. Usability represents how easily particular aspects of the system such as controls and displays can be learned and used for their intended purpose.

f. Simulation Fidelity represents characteristics of the traffic scenarios and laboratory environment and simulation participant opinions about the realism and accuracy of the simulation.

A team composed of engineering research psychologists, ARS representatives, air traffic control specialists (ATCSs), and automation specialists developed a set of objective and subjective measures for each construct. Objective measures are based on verifiable quantities such as the number of data entries made during the simulation and are typically collected using automated sources such as System Analysis Recording (SAR) tapes. Subjective measures are based on the opinions and perceptions of individuals and are typically collected using questionnaires and rating scales.

From a scientific standpoint, objective measures are usually preferable to subjective ones. Objective measures are less likely to be biased and can be replicated and verified by others. In some cases, however, objective measures may be unavailable, impractical, or may not provide the appropriate level of detail. Subjective measures can be effective data collection tools when developed and administered carefully. For these reasons, we believe that a combination of objective and subjective measures provides the best description of the operational effectiveness of an ATC system.

The following sections describe each construct and the measures it comprises. For each measure, several pieces of information are provided. First, the Definition provides a concise, formal description of the measure. Like the constructs, the definitions of the measures have been refined during each baseline. Second, the Source describes where data for that measure can be obtained. Third, the Reporting Level describes the level of detail that we recommend be reported for that measure. Fourth, the Other Information provides any other lessons we have learned during past baselines about the successful collection and analysis of data for that measure.

2.1 Safety

2.1.1 Operational Errors

Definition: This measure represents the total number of violations of applicable separation minima.

Source: Data for this measure come from recordings made by the Target Generation Facility (TGF). If the TGF is not used for target generation, data can also be reduced from SAR tapes (but with more difficulty). In addition to the automated tools, subject matter expert (SME) observers should record the occurrence of an operational error on the Observer Log, noting the sector, the simulation time, and the aircraft involved.

Reporting Level: Overall and Sector Levels

Other Information: Because a separation violation can raise serious concerns about system safety, every reported error should be independently verified. Occasionally, events that are recorded as an error actually result from an incorrect pseudopilot action or a traffic scenario inconsistency. On other occasions, a controller may issue a visual approach clearance, but the automated tool has no way of recording this. We recommend the use of videotapes of the simulation run along with printouts of data from the TGF to review possible errors. If a more in-depth verification is needed, the Systematic Air Traffic Operations Research Initiative (SATORI) system provides excellent replay capabilities. Future ATC automation systems are planned to have playback capabilities as well.

2.1.2 Conflict Alerts

Definition: This measure represents the total number of warnings issued to controllers about imminent separation violations. These warnings are issued by the ATC automation system according to FAA algorithms.

Source: Data for this measure come from SAR tapes in the en route domain or from Continuous Data Recording (CDR) tapes in the terminal domain. SME observers should record the occurrence of a conflict alert on the Observer Log, noting the sector, the simulation time, and aircraft involved.

Reporting Level: Overall and Sector Levels

Other Information: As with operational errors, conflict alerts can raise serious concerns about system safety. We recommend that each conflict alert be independently verified by an SME to determine if the alert is genuine (i.e., occurred because of controller action or inaction). Videotapes or the SATORI system can be used to verify conflict alerts.

Researchers should ensure that the number of conflict alerts is based on the actual number of occurrences and not on the raw number of recorded conflict alert messages. A conflict alert message will be written many times for the same aircraft pair, which will produce an inaccurate count if the data reduction is not conducted carefully.

During the DSR Baseline, Air Traffic SMEs indicated that some controllers may show a high number of conflict alerts due to their controlling style. The SMEs claim that these controllers are no less safe than others but that they control by conflict alert (e.g., they allow planes to become close enough to cause the conflict alert but not close enough to cause an operational error). In this way, these ATCSs may be even more efficient than controllers who keep the aircraft farther apart. Though we still believe that this measure provides substantial information about system safety, we offer this insight to discourage others from concluding that a system is unsafe based solely on the number of conflict alerts.

2.1.3 Halo Initiations

Definition: This measure represents the total number of times a controller initiated the display of the halo (also known as the J-Ring). The halo currently exists only in the en route domain.

Source: SAR tapes

Reporting Level: Overall and Sector Levels

Other Information: Initiating the halo surrounds the aircraft target with a polygon of an adapted radius (typically 5 nm). The halo aids in visual judgment of horizontal separation and can also be used as an emphasis tool and memory aid. Increasing the halo is not a reduction in safety in and of itself. Instead, increased halo use may indicate that controllers are having difficulty judging separation or maintaining an accurate picture of the air traffic situation, or both.

Researchers should ensure that the number of halo initiations is based on the actual number of initiations and not on the raw number of times controllers made the "J" entry. In the HCS, the same command is used to turn the halo on and off. This will produce an inaccurate count if the data reduction is not conducted carefully.

2.1.4 Data Block Positioning

Definition: This measure represents the total number of times a controller changed leader-line lengths and leader-line directions to maintain data block readability.

Source: In the en route domain, controllers change leader-line length and direction using data entries that are processed by the HCS and are recorded on SAR tapes. In the terminal domain, however, controllers can also change leader length and direction using knobs on the Full Digital

ARTS Display (FDAD) or DEDS. We recommend against attempting to collect these data in the terminal domain until STARS, with its fully digital display controls, is fielded.

Reporting Level: Overall, Sector, and Interval Levels

Other Information: Controllers position data blocks to maintain the readability of critical flight data. Controllers also use data block positioning as memory aids (e.g., by placing the data blocks on the right for all northbound aircraft). As with the halo, increased data block positioning is not a reduction in safety in and of itself. Instead, it may indicate that aircraft are flying in close proximity and that the controller does not have time to keep the data blocks separated.

It is appropriate to filter out data block positioning actions that are not related to maintaining readability such as "slant zero" (/0), which is used instead to indicate the transfer of communication. The specific entry types that are filtered out must be consistent across baselines that will be compared.

2.1.5 Other Safety-Critical Issues

Definition: Final Questionnaire and Observer Log

Source: Data for this measure come from questionnaires completed by study participants and SME observers.

Reporting Level: Overall Level only

Other Information: This measure is designed to record safety issues not addressed by the other measures. Researchers should ensure that issues raised for this measure are appropriate for the Safety construct. For example, many controllers view any system deficiency as a safety issue rather than a usability or workload issue. For this reason, any issues that are identified as safety critical should be independently reviewed by supervisory, training, or quality assurance SMEs and moved to other constructs if warranted.

2.2 Capacity

2.2.1 Aircraft Under Control

Definition: This measure represents the total number of aircraft receiving ATC services from a controller.

Source: Data for this measure come from TGF recordings. If the TGF is not used for target generation, Aircraft Management Program (AMP) tapes also can provide these data.

Reporting Level: Overall, Sector, and Interval Levels

Other Information: For the purposes of data collection, an aircraft is considered under track control if (a) the controller has accepted the handoff from the previous sector and (b) the handoff to the next sector has not yet been accepted. In operational ATC, however, transfer of track control technically does not occur until the aircraft is both on a controller's frequency and in his

or her airspace. To facilitate rapid data reduction, however, we recommend using the handoff-to-handoff definition and identify this in their report.

Researchers should ensure that the sectors and times being compared are precisely measured. Any discrepancy, even a few minutes, can have a substantial effect on this measure.

2.2.2 Time in Sector

Definition: This measure represents the average time aircraft spend in a particular sector.

Source: Data for this measure come from TGF recordings. If the TGF is not used for target generation, AMP tapes also can provide these data.

Reporting Level: Sector Level only

Other Information: Care should be taken to ensure that the simulators used in a comparison employ identical aircraft performance models. Valid conclusions about capacity become difficult to draw if a particular aircraft type performs better on one simulation platform than another.

As with the Aircraft Under Control measure, an aircraft is considered in a sector if (a) the controller has accepted the handoff from the previous sector and (b) the handoff to the next sector has not yet been accepted. In operational ATC, however, transfer of track control technically does not occur until the aircraft is both on a controller's frequency and in his or her airspace. To facilitate rapid data reduction, however, we recommend using the handoff-to-handoff definition and identify this in their report. We also recommend reporting data for this measure separately for arrivals and departures in terminal baselines and in en route baselines where appropriate.

2.2.3 Spacing on Final Approach

Definition: This measure represents the distance between two arrival aircraft where the first aircraft is over the middle marker and the second is trailing behind it. This measure is appropriate only for the terminal domain.

Source: Data for this measure come from TGF recordings. If the TGF is not used for target generation, AMP tapes also can provide these data.

Reporting Level: Sector Level only

Other Information: If warranted, a similar measure of aircraft spacing could be developed for the en route domain, though one has not been used in past en route baseline studies.

2.2.4 Time Between Arrivals

Definition: This measure represents the elapsed time between consecutive arrival aircraft passing over the middle marker. This measure is appropriate only for the terminal domain

Source: Data for this measure come from TGF recordings. If the TGF is not used for target generation, AMP tapes also can provide these data.

Reporting Level: Sector Level only

Other Information: No additional information

2.3 Performance

2.3.1 Overall Data Entries

Definition: This measure represents the number of data entries made by a controller using the keyboard and/or trackball across all data entry types.

Source: SAR or CDR tapes

Reporting Level: Overall, Sector, and Interval Levels

Other Information: Data for this measure should be reported separately for each staffed position. For example, baselines in the en route domain should include separate data entry counts for the radar, data, and assistant controller positions.

This measure is particularly sensitive to shifts in workload across controller positions. For example, in the DSR Baseline, we observed that radar controllers made many more data entries than in the PVD Baseline. We observed the opposite pattern for data controllers. We believed this resulted from a lack of involvement in the simulation by the data controllers due to changed requirements for between-sector coordination. To explore this further, we combined the data entries made by the sector as a whole (i.e., radar and data controllers combined) and found that the difference between systems disappeared for some sectors.

Researchers should ensure that pilot entries are not included in this measure. In Dynamic Simulations (DYSIMs), the pilot entries are recorded on SAR tapes and may inadvertently be counted as controller entries when, in fact, they are not. In TGF simulations, the pseudopilots work on a discrete system, so this is not an issue. However, ghost sectors may also make data entries, and researchers should ensure that their entries are not counted with the controller entries.

2.3.2 Specific Data Entry Types

Definition: This measure represents the number of data entries made by a controller using the keyboard and trackball for specific data entry types.

Source: SAR or CDR tapes

Reporting Level: Sector Level only

Other Information: As with the Overall Data Entries measure, the counts for specific entry types should be reported separately for each staffed position.

There are literally dozens of data entry types in the HCS and ARTS, many of which are rarely used by controllers. Stable, reliable measurements of rare data entry types are difficult to obtain and are unlikely to show reliable differences between systems. In addition, controllers can often make equivalent data entries using different command syntaxes.

We recommend recording data for all data entry types and including them in the Overall Data Entries measure. However, to facilitate data reduction, we recommend reserving the increased detail of the Specific Data Entry Types measure for a subset of types. The subset should include all the major entry types used at the facility being simulated. The subset should be chosen in consultation with SMEs and should include all common syntactic variations.

As with the Overall Data Entries measure, researchers should ensure that the pilot entries are not included. In DYSIMs, the pilot entries are recorded on SAR tapes and may inadvertently be counted as controller entries when, in fact, they are not. In TGF simulations, the pseudopilots work on a discrete system, so this is not an issue. However, ghost sectors may also make data entries, and researchers should ensure that their entries are not counted with the controller entries.

2.3.3 Data Entry Errors

Definition: This measure represents the total number of data entry error messages returned by the automation system.

Source: SAR or CDR tapes

Reporting Level: Overall, Sector, and Interval Levels

Other Information: If a controller makes a typographical error, he or she usually notices the error and corrects it using BACKSPACE or CLEAR. Because this measure counts data entry error messages returned by the automation system, only typographical errors that remain uncorrected at the time ENTER is pressed are counted. If typographical errors are a particular concern, more sophisticated analysis methods may be necessary. For example, the NAS Human Factors Branch, ACT-530, is developing a data analysis capability called the Keyboard Data Recorder (KDR) that will capture data entries keystroke by keystroke from operational ATC keyboards. From this record, a more detailed analysis of the nature of the typographical errors will be possible.

2.3.4 Number of Altitude, Speed, and Heading Changes

Definition: This measure represents the total number of controller-initiated altitude, speed, and heading changes made by simulated aircraft.

Source: TGF recordings

Reporting Level: Overall, Sector, and Interval Levels

Other Information: Consistent definitions must be applied for which pseudopilot commands are counted. Researchers should ensure that equivalent commands with different syntax (e.g., turn

right 20 degrees and fly heading 350) are counted correctly. A complete list of these commands is available from the TGF.

2.3.5 Self-Assessments of Performance

Definition: This measure represents subjective performance ratings given by a controller participant at the end of a simulation run. Ratings range from 1 (low) to 7 (high). The measure comprises two submeasures:

a. Quality of ATC services from a controller point of view
b. Quality of ATC services from a pilot point of view

Source: Post-Run Questionnaire (Appendix A)

Reporting Level: Overall and Sector Levels

Other Information: This measure has been refined to a 7-point scale so that it matches the 7-point scale used by the Air Traffic Workload Input Technique (ATWIT) Workload measure. In past studies, using different scales for similar measures has created confusion and made data analysis more difficult. Researchers who plan to compare data for this measure to studies that use an 8-point scale should use the original 8-point version.

2.3.6 Observer Assessments of Performance

Definition: This measure represents ratings of participant performance during a simulation run made by one or more SME observers. Ratings range from 1 (Least Effective) to 8 (Most Effective). The measure comprises six submeasures with three to five rating scales each. In past baselines, we have reported data for only the overall items for each submeasure. These items are as follows:

a. Maintaining Safe and Efficient Traffic Flow
b. Maintaining Attention and Situation Awareness
c. Prioritizing
d. Providing Control Information
e. Technical Knowledge
f. Communicating

Source: Data for this measure come from the Subject Matter Expert Observer Rating Form (Appendix A). Separate versions of the form are available for the terminal and en route domains.

Reporting Level: Overall and Sector Levels

Other Information: Sollenberger, Stein, and Gromelski (1997) provide detailed information on the development and administration of the Observer Rating Form. The form is based on observable controller actions and behaviors and has been widely used and validated. We recommend that researchers consult the original source for information about the successful use

of this form. In particular, we emphasize thorough training of the SMEs who will complete the form. This will improve the reliability and validity of the ratings.

The same version of the Observer Rating Form must be used in all baselines that will be compared. Researchers who plan to compare their data to older baselines should ensure that they use the same version of the form as the earlier research. The forms have undergone substantial revisions and improvements, and comparisons to data collected using earlier versions of the forms may no longer be valid.

The SME Observer Rating Form uses 8-point scales, which differs from the 7-point scales on the ATWIT, the Post-Run Questionnaire, and the Final Questionnaire. Though consistency across instruments is desirable, we believe that using the scales developed and validated by the authors of the SME Observer Rating Form adds validity and reliability to this measure.

2.4 Workload

2.4.1 ATWIT Workload

Definition: This measure represents the subjective workload ratings given by the participants during a specific time interval. To ensure stable workload ratings, the score for this measure is the average of three workload ratings made during the interval. Ratings range from 1 (low) to 7 (high).

Source: Data for this measure are collected using Workload Assessment Keypads (WAKs), one for each controller participant.

Reporting Level: Overall, Sector, and Interval Levels

Other Information: Figure 1 shows the temporal relationship between ATWIT prompts and intervals. An ATWIT probe occurs at each solid vertical line. The ATWIT Workload score for a particular interval is calculated by averaging the three ratings prompted during that interval.

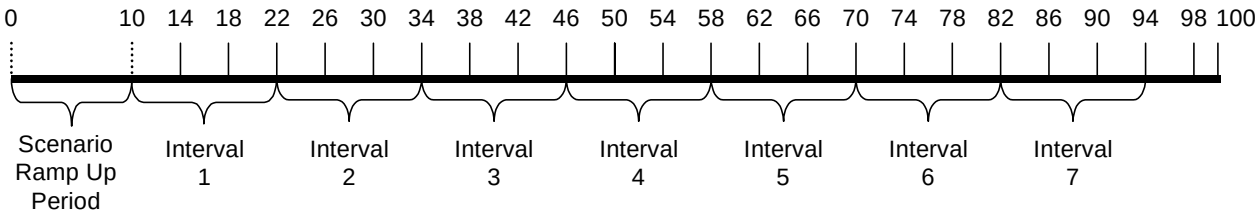

Figure 1. ATWIT probes and intervals.

For example, the first interval begins at 10:00:01 and ends at 22:00:00. The ATWIT Workload score for interval 1 is calculated by averaging the ratings given to the prompts at 14:00:00, 18:00:00, and 22:00:00. This technique provides somewhat more reliable and stable scores for each interval and allows for detailed analyses of smaller time frames if warranted.

2.4.2 Post-Run Workload

Definition: This measure represents subjective workload rating given by the controller participants at the end of the simulation run. Ratings range from 1 (low) to 7 (high).

Source: Post-Run Questionnaire (Appendix A)

Reporting Level: Sector Level only

Other Information: The scale for this measure has been adjusted from earlier studies so that it matches the scale used by the ATWIT Workload measure. Researchers planning to compare data for this measure to earlier studies that used an 8-point scale should consider returning to an 8-point scale. However, using different scales for the ATWIT Workload and Post-Run Workload measures can make comparisons more difficult.

2.4.3 Communication Taskload

Definition: This measure represents the total number of controller-initiated, push-to-talk (PTT), air-ground communications (i.e., communications between a controller and the pseudopilots working traffic in his or her sector).

Source: In earlier baselines, the data for this measure were collected manually by listening to audio recordings. However, they are now available electronically from the applicable communication system such as the Voice Switching and Control System (VSCS) or Enhanced Terminal Voice Switch (ETVS).

Reporting Level: Overall, Sector, and Interval Levels

Other Information: Reduction and analysis of air-ground PTT is extremely time-consuming because the reduction and analysis tools are not yet mature. ACT-530 has developed some techniques to make the process more efficient, but these will require modification for future baselines. Currently, the programmers in the VSCS group are working to improve their tool to facilitate future baselines.

Consistent definitions for what constitutes an air-ground PTT must be applied between studies. For example, automated tools will typically count any time the controller keys his or her microphone as a PTT regardless of whether anyone speaks or not. If data are reduced manually by reviewing audiotapes, however, this typically will not count these PTTs because no one speaks and nothing is recorded on the tape. Researchers should establish a consistent criterion for the inclusion and exclusion of PTTs before the baseline and should choose the data collection, reduction, and analysis method that best suits their criteria.

2.4.4 Coordination Taskload

Definition: This measure represents the total number of controller-initiated, PTT, ground-ground communications (i.e., communications between a controller and controllers working in other sectors or ghost sectors).

Source: Data for this measure now come from the applicable communication system such as Amecom, VSCS, or ETVS. Data for this measure can also be collected manually by listening to audio recordings.

Reporting Level: Overall, Sector, and Interval Levels.

Other Information (See comments for Communication Taskload, Section 2.4.3): The participants should be encouraged to complete their coordination actions through the voice switch rather than by talking to the controller sitting next to them. If controllers handle coordination outside the voice switch, these communications will not be counted by the automated tools and may be missed.

Controllers must follow the letters of agreement (LOAs) consistently, particularly in the cases of handoffs and point outs. In simulation conditions, some controllers are less vigilant than they would be in the field regarding coordination. This leads to unrealistically low workload and higher boredom and reduces internal validity. We strongly encourage researchers to enforce LOAs fully to add realism and to ensure that all controllers adhere to the same rules during the baseline.

Consistent definitions for what constitutes a ground-ground PTT must be applied between studies. For example, automated tools will typically count any time the controller keys his or her microphone as a PTT regardless of whether anyone speaks or not. If data are reduced manually by reviewing audiotapes, however, this typically will not count these PTTs because no one speaks and nothing is recorded on the tape. Researchers should establish a consistent criterion for the inclusion and exclusion of PTTs before the baseline and should choose the data collection, reduction, and analysis method that best suits their criteria.

2.5 Usability

Usability measures are collected from rating scales and open-ended survey questions on the Final Questionnaire (Appendix A). The Final Questionnaire should be administered after all simulation runs have been completed. In our experience, this questionnaire serves as a good starting place for an end-of-simulation briefing and discussion. Some items on the questionnaire are not appropriate for particular domains and should be eliminated from the questionnaire when appropriate. All data for this construct should be reported at the Overall Level only. The items on the Final Questionnaire address the following issues.

a. Flight Progress Strip Access
b. Flight Progress Strip Read/Mark
c. Ease of Access of Controls
d. Operation of Controls Intuitive
e. Keyboard Ease of Use
f. Radar and Map Ease of Reading
g. Radar and Map Ease of Understanding
h. Workstation Space

i. Equipment, Displays, and Controls Support Efficient ATC

j. Equipment, Displays, and Controls Impose Limitations

k. Equipment, Displays, and Controls Overall Effectiveness

l. Overall Quality of Interaction with Equipment

2.6 Simulation Fidelity

This construct is not designed to evaluate systems. Instead, it assesses whether the data for the other constructs have been collected under equivalent, realistic conditions. Data for this construct are also crucial when replicating the baseline or conducting follow-up research.

2.6.1 Traffic Scenario Characteristics

Definition: This measure represents important features of the traffic scenarios used in the simulation. It consists of several submeasures, such as

a. length of each scenario,

b. average number of aircraft entering the scenario each minute,

c. total number of arrivals,

d. total number of departures,

e. total number of overflights,

f. total number of propeller aircraft,

g. total number of jet aircraft, and

h. total number of scripted pilot deviations and requests.

Source: TGF recordings

Reporting Level: Overall and Sector Levels

Other Information: Researchers should ensure that the same algorithms and assumptions are made for what constitutes arrival, departure, or overflight aircraft. In many en route sectors, this distinction is not meaningful, and this portion of the measure should not be reported.

The TGF recordings will provide data corresponding to different aircraft types. These data must be parsed to categorize particular types such as jet or propeller. Researchers should consult with an SME if a particular aircraft type is unclear.

2.6.2 Other Simulation Characteristics

Definition: This measure represents other important features of the simulation environment outside the traffic scenarios. It consists of several submeasures such as

a. a list of standard operating procedures and LOAs used in the baseline;

b. if applicable, a list of the timing parameter for flight strips (i.e., the length of time a flight strip prints before the aircraft appears in the simulation); and

c. if applicable, a list of the Surveillance Communications Interface Processor (SCIP) settings regarding the size and offset of radar and beacon targets.

Reporting Level: Overall Level only

Other Information: These items were identified in the DSR-PVD Baseline Comparison (Allendoerfer et al., 1999) as areas that contributed to faults in internal validity. Other areas of concern certainly exist, and researchers should strive to identify and report these areas in future baseline reports.

2.6.3 Realism Rating

Definition: This measure represents the perceived realism and fidelity of the simulation run as rated by a controller participant. Ratings range from 1 (Not Very Realistic) to 7 (Extremely Realistic).

Source: Post-Run Questionnaire (Appendix A).

Reporting Level: Overall and Sector Levels

Other Information: We recommend analyzing data for this measure during the baseline so that it can be discussed with the participants. If the participants do not view the simulation as being realistic and credible, researchers should take steps to improve the simulation environment even if this requires discounting some data. Researchers should address the low realism ratings in their report.

2.6.4 Impact of Technical Problems Rating

Definition: This measure represents the perceived impact of technical problems on the participants' ability to control traffic during the simulation run. Ratings range from 1 (Not Very Much) to 7 (A Great Deal).

Source: Post-Run Questionnaire (Appendix A)

Reporting Level: Overall and Sector Levels

Other Information: See comments for the Realism Rating measure, Section 2.6.3.

2.6.5 Impact of Pseudopilots Rating

Definition: This measure represents the perceived impact of the pseudopilots on the participants' ability to control traffic during the simulation run. Ratings range from 1 (Not Very Much) to 7 (A Great Deal).

Reporting Level: Overall and Sector Levels

Source: Post-Run Questionnaire (Appendix A)

Other Information: Just like controllers and pilots, the pseudopilots differ in ability. Some pseudopilots have real pilot experience and can provide very realistic pilot communications and behavior. Others are less experienced and may provide less realistic communications. Procedures at the TGF rotate pseudopilots among roles and positions between runs, and some combinations may work better than others. Problem situations typically surface during simulation shakedown and should be addressed by TGF personnel. If this measure shows low ratings, researchers should coordinate with the TGF to ensure that the problem is rectified.

2.6.6 Scenario Difficulty Rating

Definition: This measure represents the perceived difficulty of the traffic scenario as rated by participants. Ratings range from 1 (Not Very Difficult) to 7 (Extremely Difficult).

Source: Post-Run Questionnaire (Appendix A)

Reporting Level: Overall and Sector Levels

Other Information: Data for this measure are intended as a check on the scenario development. Did the aircraft in the scenario perform normally? Was the traffic complexity too difficult or too easy?

2.7 Other Metrics

In addition to the baseline metrics described previously, a variety of other metrics has been used in baselines to examine specific questions. We recommend that researchers review these metrics to determine their applicability to their specific baseline and to include them if desired. Other metrics that focus on particular topics or tasks of interest can also be included to collect data not covered here or in the baseline metrics.

1. The PVD Baseline used a metric of strip bay management wherein a participant's use of flight progress strips was recorded and measured. This technique may be useful in future baselines where the frequency and characteristics of strip-related activities is of interest.

2. The PVD Baseline reported entry times for various data entry types. This technique may be useful in future baselines where the speed of data entries is of interest such as in the evaluation of a new keyboard or data entry syntax.

3. The DSR Baseline tested the KDR, which automatically records each keystroke made by the controller. These data may be useful in future baselines for comparing typographical errors or for analyzing the usability of a particular keyboard layout or design.

4. Items 9-11 on the Background Questionnaire have never been formally used in a baseline comparison. These items deal with controller level of familiarity with computers, satisfaction with current equipment, and level of training with a new system. These items

may be useful in future baselines to examine differences on the metrics attributable to differences in the participant sample.

5. Sections B, C, and D of the Final Questionnaire were not reported in the PVD Baseline or the ARTS IIIA Baseline, but data for these sections were collected. The ODID IV Baseline successfully used these data to compare systems. These sections contain additional information about the usability of ATC systems and are appropriate for the Usability construct.

6. The individual items on the SME Observer Rating Forms have never been formally used in a baseline comparison. Only the 6 overall metrics described in Section 2.3.6 have been used in baselines, though the detailed items have been extensively researched and validated (Sollenberger et al., 1997).

7. The NASA-TLX instrument was used in the ODID IV Baseline at the end of each run to measure workload. The NASA-TLX is a widely used measure of workload, and it could be used in future baselines in place of the Post-Run Workload measure or as a supplement to ATWIT. For more information on NASA-TLX, we recommend Hart and Staveland's article (1988). ACT-530 owns tools to electronically administer and score the NASA-TLX.

3. Baseline Methodology

3.1 Consistent Simulation Conditions

Tightly controlled simulation procedures and laboratories provide the foundation for a successful system baseline. However, the facilities and equipment associated with ATC system baselines are extremely complex, making tight control over all aspects of the simulation very difficult. The Test Director, typically an engineering research psychologist, is responsible for ensuring that consistent conditions are maintained across all baselines that will be directly compared.

Re-creating conditions from studies conducted years earlier is impossible without proper documentation and configuration management. The laboratories at the Technical Center are used constantly by many organizations. Therefore, the precise configuration of a laboratory or facility is difficult to determine after the fact. Researchers have a responsibility to document as many procedures, parameter settings, and configurations as possible and to provide this information to future studies. This should be done during the baseline.

All past baselines have been conducted using only one ATC system at a time. As such, comparisons between systems were made using data collected from separate simulation activities sometimes conducted years apart. This method has some advantages in terms of scheduling, but it makes internal validity and configuration management especially difficult.

We recommend that future baselines collect data for each system that will be compared as part of a single, large baseline. For example, the participants could run the same scenarios using both systems and alternate between systems on subsequent runs or days. This would reduce or eliminate many internal validity problems and provide much tighter simulation control. All scenarios, operating procedures, the participants, auxiliary equipment, pseudopilots, SME

observers, and questionnaires would be identical for both data sets. With a within-subjects design, the variance due to differences between individuals is reduced.

A single, side-by-side comparison is likely to be long and costly. Overall, however, we believe that a side-by-side comparison will save time and money by reducing the need to organize, prepare, run, and analyze separate simulations for each system. More importantly, a side-by-side comparison provides the highest level of internal validity.

3.2 Simulation Realism

In baseline simulations, researchers should strive for a very high level of simulation realism. The SMEs involved with scenario testing and shakedown are the best source for feedback about realism. We recommend that researchers consult with these individuals after each shakedown run. Researchers should examine the following areas.

a. Pseudopilots need adequate training during shakedown. In particular, pseudopilots need to learn the fixes associated with the sectors and when and where actions are typically taken. If they do not receive adequate training during shakedown, their communications and pilot actions may not be made in the most realistic or timely fashion.

b. Personnel staffing the ghost sectors also need adequate training during shakedown. In particular, these personnel need to learn when to accept and reject handoffs and point outs. If they do not receive adequate training, they may not provide realistic between-sector communications.

c. Researchers should ensure that the operating procedures and LOAs used in the simulation are accurate with regard to those used at the facility.

3.3 Test Plan

As part of the formal preparations for a baseline, the Test Director should develop a formal test plan. The plan should contain the following sections.

1. Introduction: This section should provide a historical context and rationale for the baseline.
2. Method: This section should describe how the baseline will be conducted. It should contain the following subsections.
   a. Facilities: This subsection should describe which laboratories and other Technical Center facilities (e.g., the TGF) are needed during the planning and conduct of the baseline.
   b. Equipment: This subsection should describe what other equipment is needed (e.g., the WAKs).
   c. Personnel: This subsection should describe the study participants and the simulation support personnel needed.
   d. Procedure: This subsection should describe the general data collection method including the sectors and scenarios to be used, the data collection tools and techniques, and the simulation schedule.
3. Data Reduction and Analysis: This section should describe how the data from the baseline will be reduced and analyzed. It should contain the following subsections:

a. Equipment: This subsection should describe what equipment is needed during data reduction and analysis (e.g., the Data Reduction and Analysis Tool [DRAT]).

b. Personnel: This subsection should describe what support personnel and facilities are needed.

c. Procedure: This subsection should describe the general data reduction and analysis method, detailing which measures will be calculated.

4. References: This section should include references to related literature, particularly regarding any tools and techniques used in the study.

5. Appendix: This section should contain copies of all the questionnaires, schedules, and briefing packages that will be given to the participants.

The National Air Traffic Controllers Association (NATCA) is involved with most FAA research and acquisition activities. NATCA will assign a representative to the program, and coordination involving the controller participants must be conducted through this individual. The Test Director should provide the NATCA representative with a copy of the test plan before any baseline data are collected.

3.4 Schedules and Rotation

In our experience, about 12 controllers is the maximum that can be made available to participate in a simulation due to staffing requirements at their home facilities. If the participants are drawn from multiple facilities, as they were in the ODID IV Baseline, a larger number can be used. In addition, the Technical Center laboratories are scheduled continuously. In our experience, 3 weeks is the maximum that can be made available for a baseline. Even less time will be available during the formal engineering test period.

Researchers must not develop a schedule that violates the labor agreement between the FAA and the NATCA. That is, bargaining unit controllers must not be required to staff a position for more than 2 consecutive hours without a break. The agreement also requires a 30-minute meal break, no more than 8 hours per day (including breaks), and no more than 5 days a week.

Other practical considerations set further limits on the schedule. Controllers, pseudopilots, simulation support staff, SME observers, and researchers all should be given short breaks (15-20 minutes each) between simulations and meal breaks (1 hr each). Fewer or shorter breaks will lead to fatigue and poor relations among the research team. Remember that participating in human factors research is voluntary and if participants feel ill-treated or overworked, they are unlikely to volunteer again (and are likely to tell their friends). In addition, the laboratory and simulation equipment requires reconfiguration time. We recommend scheduling a minimum of 20 minutes between runs. In our experience, 5 hours of actual simulation time a day is about the maximum that can be supported.

We also recommend against running scenarios longer than about 100 minutes without a position relief. Some controllers may become fatigued, bored, or unresponsive if required to staff a position longer than this. We also strongly recommend using at least two traffic scenarios. If participants work the same scenario multiple times, they quickly learn to “beat” it and to anticipate occurrences. This can lead to bored participants and unreliable data. Rotating

participants through two scenarios and several sectors or positions usually is adequate to keep controllers' interest through a 1-week simulation. If the simulation covers multiple weeks with the same participant sample, we recommend using more than two traffic scenarios.

Researchers should design the schedule so that every participant serves in every position, sector, and scenario once during the simulation. The schedule should also allow each SME observer to evaluate each participant at least once. In en route baselines, we recommend that SME observers evaluate the participants while they staff the radar position. In terminal baselines, we recommend that the SME observers evaluate the participants while they staff a challenging sector, such as Final. If additional SME observers are available, more sectors or positions can be evaluated.

Researchers should ensure that schedules do not over-sample a particular participant, observer, or scenario because this may bias the data. If technical problems force the cancellation of a run, researchers should assess any potential biases that may be introduced and discount data to provide a balanced data set if necessary. We also recommend that researchers schedule several make-up runs that can be used in case of technical problems.

A sample baseline schedule is provided in Table 1. In the sample schedule, eight controller participants staff two sectors with two positions. The participants work two scenarios, one using sectors 26 and 38 and the other using sectors 27 and 35. Each participant staffs each sector twice, once as the radar controller and once as the data controller. Two make-up runs are scheduled for the last day of the simulation to be used if needed. Two SME observers evaluate the participants while they staff the radar positions. Each SME observer evaluates each participant twice. We encourage researchers to adapt this schedule to the design of their baseline.

3.4.1 Runs per Scenario

In traditional experimental design, increasing the number of trials increases confidence in the conclusions that can be drawn from the experiment. This also applies to ATC system baselines in that more simulation runs will lead to more stable data and more reliable comparisons between systems.

However, the desire for stable data must be balanced against practical considerations such as the availability of participants and facilities. In many cases, it is simply not practical to conduct a baseline with as many runs as traditional experimental design requires. Accounting for all the practical constraints described previously, we recommend scheduling 8-10 simulation runs a scenario. The PVD Baseline scheduled fewer runs per scenario and some of the data reported there have been found unreliable (Allendoerfer et al., 1999). The ARTS IIIA, ODID IV, and DSR Baselines each scheduled eight or more runs a scenario. There is also a good chance that at least some data will be lost or unusable due to technical problems or unforeseen occurrences. We strongly recommend scheduling at least two makeup runs.

Table 1. Sample Baseline Schedule

Information for participants:

The simulation will begin each day promptly at 1600 hrs and will end at approximately 2310 hrs. Please be in the lab and ready to run at 1600. When you are not running, you may leave the Technical Center, though you are expected to be in the lab and ready to run when your next run begins. We will try to stick to this schedule as closely as possible but technical problems may force us to reschedule runs. We will complete 4 full runs every night. Please note the briefings on Monday and Friday afternoons. If no makeup runs are necessary, the closing briefing will be rescheduled for Friday morning.

		Participant							
Date	Time	1	2	3	4	5	6	7	8
Monday, June 9	1500 hrs	Pre-Simulation Briefing: Human Factors Lab Briefing Room							
	1600 – 1710	26-R, SME1	26-D	38-D	38-R, SME2				
	1730-1910					35-R, SME1	35-D	27-D	27-R, SME2
	1910-2000	Break							
	2000-2110			26-R, SME1	26-D	38-D	38-R, SME2		
	2130-2310	27-D	27-R, SME2					35-R, SME1	35-D
Tuesday, June 10	1600-1740			35-D	35-R, SME1	27-R, SME2	27-D		
	1800-1910					26-D	26-R, SME1	38-D	38-R, SME2
	1910-2000	Break							
	2000-2140	35-R, SME1	35-D	27-R, SME2	27-D				
	2200-2310	38-D	38-R, SME2					26-R, SME1	26-D

Table 1. Sample Baseline Schedule (continued)

		Participant							
Date	Time	1	2	3	4	5	6	7	8
Wednesday, June 11	1600 – 1710					26-R, SME1	26-D	38-R, SME2	38-D
	1730-1910	35-D	35-R, SME1	27-D	27-R, SME2				
	1910-2000	Break							
	2000-2110	38-R, SME2	38-D					26-D	26-R, SME1
	2130-2310			35-R, SME1	35-D	27-D	27-R, SME2		
Thursday, June 12	1600-1740	27-R, SME2	27-D					35-D	35-R, SME1
	1800-1910	26-D	26-R, SME1	38-R, SME2	38-D				
	1910-2000	Break							
	2000-2140					35-D	35-R, SME1	27-R, SME2	27-D
	2200-2310			26-D	26-R, SME1	38-R, SME2	38-D		
Friday, June 13	0900-1010	Makeup Run 1 (if necessary)							
	1030-1210	Makeup Run 2 (if necessary)							
	1210-1300	Break							
	1300-1500	Post-Simulation Briefing and Discussion: Human Factors Lab Briefing Room							

3.4.2 Repeated-Measures Design

Furthermore, we propose to improve stability and reliability by using a true repeated-measures experimental design. In this design, a participant's data for one system can be compared directly to his or her data for the other system. While past baseline comparisons did use many of the same participants, there was never adequate control over the experimental conditions or the participants to use a true repeated-measures design. The side-by-side comparison proposed in Section 3.1 will allow this and should increase statistical power and reliability.

3.5 Laboratory Platforms

The primary laboratories that support system baseline activities are located in Building 300 of the Technical Center. The laboratories for all current ATC systems are located in this building. Laboratories for many new ATC systems are located in Building 316.

The Test Director must schedule laboratory time through the Facility Control Office (FACO). FACO creates their schedules on a priority basis. The Test Director and the Program Office should work with FACO to establish the proper priority for the system baseline. Requests should be made well in advance. FACO releases the schedules for each week on the preceding Thursday. We recommend that researchers inform the participants and technical staff that night shifts may be the only hours available. Most controllers are accustomed to working night shifts at their home facilities if these hours are the only times the laboratories are available.

3.5.1 En Route Simulation Support Facility

The En Route Simulation Support Facility (ESSF) in Building 300 houses 22 PVD consoles connected to the Technical Center HCS. The PVDs in the ESSF are arranged in two configurations as used in the operational environment. The PVDs have the full complement of hardware used in the field including flight strip bays, flight strip printers, and communication equipment. Simulations in the ESSF can be driven by the TGF or the DYSIM.

3.5.2 Display System Replacement Laboratory

The DSR will eventually replace the PVD in the field. At present, the DSR Laboratory in Building 316 is used primarily for engineering tests of hardware and software. In the future, this laboratory will become the primary laboratory for highest fidelity, human-in-the-loop simulations in the en route domain. It has already served as the platform for the DSR Baseline. Simulations in the DSR Laboratory are driven by the TGF.

3.5.3 Integration and Interoperability Facility

The Integration and Interoperability Facility (I^2F) is directed and funded by the En Route Integrated Product Team and is located in Building 27. The primary function of the I^2F is prototype integration and operational tests of new en route technology. It contains a fully functional ARTCC Laboratory with DSR controller and supervisor workstations. The laboratory

is suitable for testing hardware, software, and operator integration. It has not been used to support system baselines in the past but may provide an alternative to the DSR Laboratory in the future.

3.5.4 Terminal Simulation Support Facility

The Terminal Simulation Support Facility (TSSF) is housed in Building 300. It consists of several laboratories that simulate the different configurations used in TRACONs. These laboratories include the ARTS IIA, ARTS IIIA, ARTS IIIE, and En Route Automated Radar Tracking System (EARTS) Laboratories. The TSSF also supports simulations in the Technical Center Tower Cab Laboratory. Simulations in the TSSF Laboratories are driven by the TGF or by the Enhanced Target Generator (ETG).

3.5.5 Standard Terminal Automation Replacement System Laboratory

The ARTS computers and FDAD/DEDS displays will be replaced by the STARS. At present, the STARS Laboratory in Building 316 is used primarily for engineering hardware and software tests but will eventually be available for use in system baseline simulations.

3.5.6 Transition Laboratory

The Transition Laboratory provides a capability for researchers to explore the issues involved when an original TRACON system and its replacement are in place simultaneously at one facility. This laboratory contains FDADs and STARS displays. Simulations in this laboratory are driven by the TGF.

3.5.7 Oceanic Laboratory

The Oceanic Laboratory is located in Building 300. It includes PVDs, strip bays, Oceanic Data Link (ODL) systems, and a simulated Airline Operations Center (AOC) workstation. Simulations in this laboratory are driven by an internal target generation system rather than the TGF.

In oceanic ATC, a controller does not communicate directly with the pilots but works through an Aeronautical Radio, Incorporated (ARINC) radio operator. The radio operator establishes short-wave radio contact with each flight to relay ATC clearances. Aircraft contact the ARINC radio operator to relay position reports every 10 degrees of longitude. Therefore, in a simulation, it is only necessary to provide a pseudo-ARINC radio operator and, if an airline presence is required, a pseudo-AOC operator. A suitable traffic scenario must still be developed that includes such events as position report messages and pilot requests from each aircraft at the correct intervals.

3.6 Simulators

The TGF, operated by the System Simulation Support Branch (ACT-510), is the primary simulator for the laboratories in Buildings 300 and 316. The TGF provides simulated air traffic (up to 3,000 flight plans simultaneously). TGF pseudopilot workstations display aircraft

information and accept commands to change aircraft speeds, headings, altitudes, and so on. The TGF is also an important source of automated data. The Test Director should schedule TGF time with ACT-510.

The MicroTGF software, a version of the full TGF software that runs on standalone workstations, is also available. This version of the software can be ported to laboratories that do not receive direct TGF feeds, either within the Technical Center or at other facilities. The MicroTGF uses the same scenario definitions as the main TGF and provides the same pseudopilot and data collection tools. However, researchers should remember that the MicroTGF is not a display system simulator. It provides scenario generation and aircraft behavior, not emulation of controller hardware or software.

An alternate simulator for en route is DYSIM. DYSIM is part of the ESSF and allows the laboratory to operate in a stand-alone mode. In this case, controllers working at PVDs in the laboratory serve as simulation pilots and maneuver the simulated traffic. The DYSIM cannot use TGF scenario definitions. The Test Director should schedule DYSIM time with FACO and the ESSF. In some cases, the DYSIM Laboratories at field facilities may also be available. These facilities must be coordinated through the field training departments.

An alternate simulator for the terminal domain is the ETG. The ETG is contained in the ARTS and allows the TSSF to operate in a stand-alone mode. When using the ETG, several of the FDAD/DEDS workstations are used as simulation pilot stations. The ETG cannot use TGF scenario definitions. The Test Director should schedule ETG time with FACO and the TSSF. The ETG can be used with the STARS EDC configuration but is not available in the STARS ISC or later configurations. In some cases, the ETG Laboratories at field facilities may also be available. These facilities must be coordinated through the field training departments.

The ATCoach simulator also provides target generation for simulations in the STARS Laboratory. This software package runs on UNIX workstations. Scenario definitions that have been created for use by the TGF, DYSIM, or the ETG are not compatible with ATCoach. At present, the ATCoach software in the STARS Laboratory has not been used for baseline simulations. However, ATCoach has been used extensively in non-baseline simulations by ACT-510 and ACT-530, so some local expertise is available with this simulator.

3.6.1 Pseudopilots

ATC simulations require that someone play the role of the pilots of the simulated aircraft. These can be pseudopilots or ATCSs, depending upon the target generator. When using the TGF, pseudopilots play the role of simulated aircraft and are responsible for communicating and executing clearances associated with those aircraft. They make air-ground communications with the controller participants and make adjustments to aircraft speed, heading, altitude, and flight plan as directed by controllers. Pseudopilots are trained in aviation phraseology, simulated airspace, and aircraft behavior but most are neither controllers nor pilots. As such, they can provide realistic communications and aircraft behavior under most conditions but perform less well when asked to make impromptu communications or flight plan changes to fit changing air traffic situations. ACT-510 coordinates the Technical Center pseudopilots.

When using the DYSIM or the ETG, ATCSs serve as simulation pilots. Controllers are trained on DYSIM or ETG in the field, and most are accustomed to serving as pilots. In the PVD Baseline, which used the DYSIM, the participants alternated between the controller and pilot roles. Though this was an efficient use of controller resources, we do not recommend this method for future baselines. The variance in skill among different controllers serving as pilots can be great. This results in some participants receiving less realistic pilot communications and aircraft behavior than others and creates internal validity problems. In addition, some controllers have used this as an opportunity to play jokes on their friends. For example, some controllers serving as pilots have changed headings, speeds, and altitudes without authorization from the controller actually working traffic. For a valid system baseline using the DYSIM or ETG, we strongly recommend that a cadre of controllers be assigned to the pilot role, and another cadre be assigned to the controller role, and that they do not alternate. The controllers selected to serve as pilots should be chosen because they take the assignment seriously and are aware of the need for consistency across conditions.

3.6.2 Ghost Sectors

In addition to aircraft, ATC simulations also must simulate the other sectors and facilities (ghost sectors) with whom the controllers interact. This interaction includes approving and rejecting handoffs, point outs, and any other ground-ground communications. In past baselines, one individual from the TGF or simulation laboratory has staffed all the ghost sectors. We recommend that future baselines carefully review the workload of this individual to ensure that he or she can handle all the traffic in the simulation while still doing a credible and realistic job. If additional staffing is warranted, researchers should request it. We also recommend that the individuals staffing the ghost sector be very familiar with the operating procedures and the LOAs that apply to the sectors being simulated. The ghost sector should only accept handoffs that are made in a realistic fashion. For example, in the DSR Baseline, the participants sometimes handed aircraft off at an altitude that violated an LOA, which would have been rejected in the field. Unfortunately, the person staffing the ghost sector did not know about the LOA, accepted the handoffs, and created an unrealistic simulation condition. We strongly recommend that a controller or other SME from the facility staff the ghost sector or that the support personnel receive substantial training on the simulated sectors and operations.

3.7 Airspace

3.7.1 Simulated Airspace

The choice of airspace will affect most aspects of the baseline. The Program Office will probably choose the baseline airspace based on availability, cost, and schedule considerations. In most cases, the baseline airspace will be from one of the early facilities on the deployment schedule. Because the TGF provides target generation for Operational Test & Evaluation (OT&E), the Program Office will probably choose the OT&E facility for the baseline also.

If there is some latitude in choosing airspace, researchers should consider external validity (i.e., how easily the baseline data can be generalized to the rest of the ATC system) when choosing an airspace to simulate. Because baselines are meant to characterize the system under typical conditions, we recommend that researchers choose airspace that does not have many unusual

characteristics. Some characteristics to consider include the presence of military warning areas and other special use airspace, interaction with international airspace, the mix of aircraft types, areas of limited radar coverage, and areas of unusual weather patterns. Consultation with SMEs from the chosen facilities should reveal any unusual characteristics.

3.7.2 Generic and Unfamiliar Airspace

Generic airspace is airspace that does not exist in the field but has been developed for various testing purposes. The so-called ZCY generic airspace was developed for formal engineering tests purposes but is generally not appropriate for human factors studies. It is difficult to learn and does not "feel" like real airspace to controllers. However, a second form of generic airspace, known as Genera, has been developed by ACT-530 expressly to be easily learned by participants and to have the features of typical terminal or en route airspace (Guttman, Stein, & Gromelski, 1995). Genera Airspace allows participants to be drawn from diverse facilities thereby improving external validity and reducing staffing problems. Currently, the Genera Airspace is available only for the ATCoach simulation platform, but versions are under development for the TGF. Genera Airspace is not yet available for the oceanic domain.

Some of the benefits of Genera Airspace can also be gained by using unfamiliar airspace. In this case, airspace from one facility is used, but the participants are drawn from other facilities. This requires less development than Genera Airspace because scenarios and airspace definitions are already available but allows the participants to be drawn from multiple facilities. The ODID IV Baseline used this technique, using controllers from several facilities working Washington ARTCC airspace.

Genera and unfamiliar airspace require substantial training for the participants. In past baselines, this training has taken the form of classroom training on fixes, frequencies, routes, and procedures followed by several training runs. LOAs and operating procedures take longer to learn, depending upon their number and complexity. An SME from the home facility should identify the most important and most difficult procedures associated with the airspace, and the training should focus on those. The ODID IV baseline trained non-Washington ARTCC (ZDC) controllers on ZDC airspace for 1 week prior to beginning formal ODID training. Using the Genera Airspace, 2-3 hours of training are typically required before controllers have completely learned the airspace. We recommend that future baselines use Genera Airspace when available for the appropriate domain.

Training on Genera or unfamiliar airspace can be especially time consuming when the participants are also learning new equipment and procedures. Researchers should consider using a performance-based criterion or an over-the-shoulder rating procedure to ensure that all the participants are sufficiently trained before beginning the baseline runs. Research has shown that, with training, controller performance using Genera Airspace is equivalent to performance using home airspace (Guttman et al., 1995).

3.8 Traffic Scenarios

A scenario is a set of simulated air traffic and environmental conditions that provides input to the simulator. A scenario specifies the aircraft call signs, flight plans, types, altitudes, beacon codes,

start times, and so on. Baseline traffic scenarios should provide a moderate-to-heavy level of complexity. We have found that this level is sufficient to keep the participants engaged in the simulation but is not so complex as to overwhelm them. This complexity level is also more likely to show between-controller variability than a lower level where all controllers usually perform equally well.

Past baselines have created a moderate-to-heavy complexity by simulating a 90th percentile day for traffic volume. In each case, TGF personnel obtained traffic data from the chosen facility and converted those data to the appropriate simulator format. The resulting scenarios were refined by SMEs from the facility during shakedown. During the DSR Baseline, however, our participants remarked that the scenarios were not complex enough to keep their interest or challenge their abilities. We believe this discrepancy resulted, in large part, from reduced requirements for between-sector coordination and from unrealistic aspects of the simulation such as inconsistent flight strip printer intervals.

We recommend that researchers carefully evaluate scenarios to ensure that they contain the intended complexity level. For example, even during a 90th percentile day in the field, there are periods of high volume and periods of lower volume. If the selected time falls during a low-volume period, the resulting traffic scenarios will not contain the intended complexity level. We recommend that researchers construct the baseline scenarios so that traffic can be easily added to increase complexity. Personnel at the TGF are familiar with this technique and can program their scenarios appropriately.

Because flight data processing systems like the ARTS IIIA are designed for operational use, beginning and ending scenarios can create special technical problems. For example, aircraft cannot simply appear at altitude without the system generating serious errors. To prevent these errors, simulated aircraft usually must enter the airspace at a rate similar to the real world. As such, most simulators will require a "ramp up" period where the traffic volume is low and increases to the desired level over time. Past baselines have used a relatively short ramp up period, approximately 10 minutes. For data analysis purposes, we discounted the first 10 minutes of data to prevent biasing the data toward operations with unrealistically low traffic volumes.

3.9 Controller Participants

The controller participants for baseline simulations should be Full Performance Level (FPL). Unless Genera Airspace is being used, we recommend using only participants who are certified on the sectors that will be simulated. The ODID IV Baseline used participants who were not certified on the simulated sectors. Therefore, despite the extensive training provided in that study, it is unlikely that these controllers performed as well on the simulated airspace as on their home airspace.

Developmental controllers vary widely in their skill level and, in general, should only participate if training and transition are the focus of the program. However, because the recommended experimental design is within-subjects, the training requirement may be relaxed for appropriate reasons without biasing the results. For example, a future baseline might choose to include 10% developmental controllers to better represent the controller population in the field. If the

simulation schedule design is appropriately counterbalanced (i.e., developmentals work both systems the same number of times), the effect of the developmental controllers should be equal for both systems.

Researchers should recruit the controller participants as far in advance as possible. The union contract requires 60 days notice to distribute recruiting announcements and allow controllers to make arrangements. The controller participants receive their normal wages for the duration of the baseline plus travel costs and per diem.

Researchers must respect participant rights during a baseline simulation. They are responsible for ensuring that all the participants know that the data they provide during the baseline are anonymous and confidential. We recommend that researchers adapt the Statement of Confidentiality and Informed Consent (Appendix B) to their baseline and distribute it to the participants during the pre-simulation briefing. Researchers should also assign participant codes at this briefing. All research conducted by the FAA using human participants is subject to approval by the Institutional Review Board.

3.10 Subject Matter Expert Observers

In past baselines, SME observers were supervisors from the simulated facility. The SME observers were responsible for observing each simulation run and completing the SME Observer Rating Form. If supervisors are not available, quality assurance and training personnel are also suitable to serve in the SME observer role. However, we recommend against using field controllers who do not have this type of experience in the SME observer role. Controllers who are not accustomed to evaluating their peers may feel awkward doing so and may not provide valid results.

3.11 Briefings

Researchers should schedule at least two briefings, one before the simulation runs begin and a second after all simulation runs are complete. For the initial briefing, researchers should provide a briefing package containing copies of the baseline schedule and any appropriate reference materials about the airspace. This is especially important if the participants are being tested on airspace other than their home airspace. The participants should also complete the Background Questionnaire during this briefing. In the initial briefing, researchers should discuss the following topics:

a. Why is the research being conducted? Researchers should discuss the history of system baselines and the ATC system under evaluation.

b. How will the results of the research be used? Researchers should discuss how the baseline will be used by the Program Office.

c. How will the participants' confidentiality and anonymity be guaranteed? Researchers should assign participant codes at the initial briefing and explain that no names should be used on any materials. Researchers should also distribute the Statement of Confidentiality and Informed Consent (Appendix B).

d. What is the participants' role in the research? Researchers should discuss what is expected from the participants, emphasizing the simulation and the actions they are expected to perform.

e. How will the data be collected? Researchers should discuss each data source and describe what is expected from the participants regarding that source, emphasizing the WAK and the questionnaires.

f. How does the simulator differ from the field? Researchers should discuss hardware and software differences such as unavailable functions or entries. They should also discuss the pseudopilots and their abilities. Researchers should describe any differences in procedures and how to coordinate with ghost sectors.

g. What is the schedule for runs, breaks, and briefings? Researchers should describe when and where each activity will occur, emphasizing the importance of starting and ending each simulation run on time.

Researchers should also conduct a final briefing after all simulation runs have been completed. In this briefing, researchers should guide the discussion about the system under evaluation and about the baseline process itself. In particular, researchers should focus their discussion around the constructs so that adequate information is provided for each one. The participants should complete the Final Questionnaire during this briefing. We recommend that researchers discuss the following topics with the participants.

a. Was there a difference between the systems? Researchers should discuss each operational construct in general terms and solicit comments. They should also seek to understand how the participants compensated for any differences.

b. Which aspects of the new system need to be evaluated more closely or improved in the future?

c. Which aspects of the new system are an improvement over the existing system?

d. How realistic was the simulation relative to operations in the field? Researchers should discuss areas where the simulation was less than perfect such as pseudopilots, ghost sectors, and procedures and try to understand how these may have affected participant performance. They should also seek to understand if the WAK, video cameras, or SME observers were intrusive or distracting.

3.12 Training

Training for baselines can be a difficult issue. If the baseline uses fielded systems and the participants work their home airspace, as was the case in the PVD and ARTS IIIA Baselines, the training requirements should be minimal. In these studies, the participants required training with the WAK and the questionnaires but little else. On the other hand, if the participants are using new equipment or working unfamiliar airspace, they will require substantial training. In the ODID VI Baseline, the controller participants required a week of training on the Washington ARTCC airspace using the HCS-PVD and a week of training using the ODID equipment before data collection. In the DSR Baseline, the participants received 2 weeks of training on the DSR

and had completed 2 weeks of other OT&E activities before the baseline. At a minimum, researchers should provide training in new equipment, unfamiliar airspace, unfamiliar procedures, the WAK, and the questionnaires.

4. Data Collection Techniques and Tools

4.1 Target Generation Facility Recordings

All simulations using the TGF for target generation can record a variety of information about aircraft positions, flight plans, separation, pseudopilot actions, and so on. The TGF records data to disk and to 8mm data tape. The TGF does not record any data about controller interactions with the display or automation systems such as data entries. The Test Director should arrange with the TGF personnel to create and archive TGF tapes for each simulation run.

4.2 System Analysis Recording Tapes

The ESSF, the DSR Laboratory, and the Oceanic Laboratory can record SAR tapes. SAR tapes record a variety of information about controller interaction with the HCS. The Test Director should arrange with laboratory personnel to create and archive SAR tapes for each simulation run. The SAR tapes can be made in a variety of modes, depending on what data are needed. The Test Director should consult with the laboratory personnel and provide them with a list of the measures that will be reduced from the SAR tapes to ensure that the proper modes are activated.

The DSR Laboratory can also record a special version of SAR tapes called DSR SAR. These tapes contain mostly redundant information with the HCS SAR tapes. However, as data analysis tools are developed, DSR SAR tapes may eventually provide more detailed information than is currently available.

4.3 Aircraft Management Program Tapes

The ESSF, DSR Laboratory, and Oceanic Laboratory can also record AMP tapes. These tapes provide information about aircraft movement and flight data such as the number of aircraft in the sector and the duration of each flight. Most of the data recorded on AMP tapes can also be obtained from TGF recordings, but AMP tapes can be useful as backups.

4.4 Continuous Data Recording

The TSSF Laboratories can record CDR tapes that contain information about controller interaction with the ARTS. The Test Director should arrange with laboratory personnel to create and archive the CDR tapes from each simulation run. During the ARTS IIIA Baseline, the CDR tape drive was not reliable and introduced gaps and errors into the data. We recommend that future terminal baselines record data using the CDR disk drive rather than the CDR tape recorder.

4.5 Communications Data

The laboratories at the Technical Center vary in the specifics of their voice switch capabilities. In each case, the voice switch can provide automated data about the number of PTT

communications between the participants and the pseudopilots (air-ground) and between the participants and the other sectors and ghost sectors (ground-ground). The Test Director should coordinate with personnel from the laboratory to configure the voice switch to record this information.

With the development of the VSCS, more options are available for recording and analyzing communication data in the ESSF and DSR Laboratory. The VSCS can record voice communications on a system called the Legal Recorder. The VSCS can also provide data about the number of air-ground and ground-ground communications using the VSCS Log Recorder. The Log Recorder provides output of VSCS messages in 5-minute intervals. ACT-530 has developed reduction and analysis techniques to transform VSCS Log Recorder output into more useful counts of air-ground and ground-ground communications. Improved reduction and analysis tools for VSCS data are being developed by the communications specialists at the Technical Center and may be available for future baselines.

The ETVS provides a similar capability for the ARTS Laboratories but has not yet been used in a baseline simulation. As it becomes more widely used, we expect that the ETVS will become an important data collection tool.

4.6 Audiotapes and Videotapes

We recommend that researchers collect audiotapes and videotapes during each simulation run. The main purpose of these tapes is to provide backup information in case a technical problem corrupts other data sources and to allow SMEs to review critical incidents such as operational errors. Recordings are also useful for verifying start and stop times.

Controller and pseudopilot voice communications are handled by the applicable voice switch system such as the VSCS or the ETVS. The Test Director should coordinate with communications platform personnel to ensure that the voice switch is configured to record the required data. Controller ambient communications (i.e., communications with the controllers sitting next to them) are recorded using wireless microphones worn by the participants. These recordings are made to capture any ground-ground communications that are not accomplished through the voice switch.

Video recordings can be made in the laboratories using equipment in the Mobile Experimental Recording Rack (MERR) available from ACT-510. The MERR provides a complete suite of video recording equipment including low-illumination cameras, a time code generator, and multitrack recording. The MERR can receive audio input from the applicable voice switch and from wireless microphones worn by the participants. The MERR can be transported to any laboratory at the Technical Center. We recommend that video cameras be positioned above and behind the controller participants so that the radar screen, controls, and flight strip bays are visible.

We also recommend that researchers videotape a radar screen directly. This record can prove invaluable when verifying data and for reviewing operational errors. However, analog radar displays do not show up well on videotape because of their poor contrast. Digital displays show up better, but data blocks can still be difficult to read.

Taping a simulation raises some confidentiality concerns. The participants must be informed that recordings are being made, and they must give consent for these recordings. Controllers are accustomed to having their voice communications recorded, but they are less accustomed to having their physical actions and ambient discussions recorded. Researchers should explain what information will be recorded and how it will be used.

4.7 Workload Assessment Keypad

The ATWIT has been widely used in the FAA (Stein, 1985), and a similar method is in use at Eurocontrol (Hering & Coatleven, 1996). It has been administered using a variety of techniques, but we recommend collecting data for the ATWIT Workload measure using WAKs. A WAK consists of numbered and lighted keys and a tone generator. At a predetermined rate during the simulation run (e.g., every 4 minutes), the WAK emits a beep and illuminates its lights. At this time, each participant presses the key corresponding to his or her subjective workload at that moment. If the participant does not make a rating during a predetermined duration (e.g., 20 seconds), the lights extinguish and no rating is recorded for that prompt.

Up to four WAKs can be connected to one routing device that is then connected to a Windows-compatible laptop computer. ACT-510 has developed software that controls the WAKs and automatically records data on the laptop hard disk.

We believe that using the WAKs is preferable to other methods that have been used to collect ATWIT data. In the PVD Baseline, the "cuckoo" alarm in the control room sounded, and controllers made a special entry on their PVD keyboards. This required that the ATWIT data be reduced from SAR tapes, which added delay and expense. In some other non-baseline studies conducted by the Technical Center, the ATWIT was administered manually—that is, by an experimenter with a stopwatch and paper and pen. This method is undesirable due to the potential for timing and recording errors. WAKs provide an efficient and accurate way to administer the ATWIT and require no hardware or software changes to the ATC systems being evaluated.

Researchers should provide verbal and written instructions on the proper use of the WAKs. Sample instructions are provided in Appendix C (Stein, 1985). The Final Questionnaire also contains an item that serves as a check on the participants to ensure that they used the WAKs as intended.

4.8 Questionnaires and Ratings

Appendix A provides the current versions of the baseline questionnaires. If researchers plan to compare their data to data from earlier baselines, they should consult the appropriate report to ensure that they use the proper versions. There are five baseline questionnaires.

- Background Questionnaire. The controller participants complete this questionnaire as part of the initial briefing, before any simulation runs begin. It contains items about controller experience and training.

- Post-Run Questionnaire. The controller participants complete this questionnaire after each simulation run. This questionnaire contains seven items addressing the just-completed simulation run. Researchers should ensure that the participants complete every item on the questionnaire and that all coding information is complete.
- Observer Log. SME observers complete this questionnaire during each simulation run. They should record any unusual events (e.g., operational errors), noting the time and any details about the event. They should also record any technical problems.
- Final Questionnaire. The controller participants complete this questionnaire as part of the final briefing, after all simulation runs have been completed. Researchers should ensure that the participants complete every item and that they have sufficient opportunity to write comments. This questionnaire is also suitable for other interested parties such as the SME observers as long as their data are not included with the participant data. This questionnaire now contains the item formerly known as the ATWIT Questionnaire.
- SME Observer Rating Form. SME observers complete this questionnaire during and after each simulation run. Because proper completion of the form requires substantial attention, observers should only evaluate one controller at a time, typically the controller staffing a challenging sector like Final. Appendix A contains two rating forms, one for the en route environment and one for the terminal environment.

4.9 Keyboard Data Recorder

The KDR consists of a specially equipped PC and cables that attach it to the DSR or STARS controller keyboards and captures each controller keystroke and trackball action. These data can then be analyzed to determine which keys were pressed and which typographical errors were made. Currently, the data reduction and analysis routines for the KDR are not mature, but the KDR may provide valuable information about controller keyboard and trackball use in the future.

4.10 Verifying and Archiving Data

The raw data from each simulation run are irreplaceable. To prevent loss or corruption of data, researchers must verify and archive data throughout the simulation. Before each run, they should ensure that

a. all clocks are synchronized;

b. all recording media are in place, have enough available space for the entire run, and are properly labeled; and

c. enough blank copies of all questionnaires are available and labeled.

Once the simulation run has been successfully started, researchers should verify that each automated data source is recording data by ensuring that

a. the sound level indicators on video recorders are moving and the tape counters are increasing,

b. the data tapes are turning on every automated data source and that any indicators are responding,

c. the WAKs are prompting at the appropriate interval and the ratings are being added to the database, and

d. the SME observers are making notes on their Observer Rating Forms.

After each run, we recommend that researchers conduct a more thorough verification of the data. The simulation schedule will often preclude examining every data source, but researchers should conduct spot checks. Researchers should

a. check the labels on all the data, audiotapes, videotapes, and questionnaires;

b. spot check the quality of the video and audio recordings by playing back a minute of one tape;

c. reduce one data tape to ensure the recorders are operating correctly; and

d. spot check the participant questionnaire answers to ascertain that they are completing all the questions.

At the end of each day, researchers should backup and archive all data. When using a data source that records to tape, it may not be feasible to immediately make a copy of every tape. Researchers should

a. check that the tapes are labeled and stored in a safe place;

b. make a backup of data tapes;

c. change the permissions on backup files to prevent overwrites, if possible; and

d. make a photocopy of the completed questionnaires.

5. Data Analysis Techniques and Tools

5.1 Automated Tools

Data from automated tools usually require at least one round of reduction before they can be analyzed. The Test Director should coordinate with data specialists from the TGF, the simulation laboratory, and the communications platform to run the appropriate reduction routines. The data specialists should note the routines they used and provide a list of all parameters and configurations to the Test Director so these can be archived and managed.

TGF tapes are reduced using the DRAT, which is available at the TGF. The output of TGF reductions can be provided in hard copy or electronic format. In most cases, the electronic format is desirable because some reports may require a second round of reduction. These second-round reduction routines are typically written in a specialized language such as REXX or Perl. Researchers should consult with the DRAT specialist and specify their requirements before the data reduction begins. Once reduced using the DRAT, data are generally imported into spreadsheet software and a statistical software package for analysis and testing.

SAR tapes are reduced using the Data Analysis and Reduction Tool (DART), which is available throughout the Technical Center. Using the DART requires specialized training and should be undertaken only by trained personnel. The DART produces large output reports that can be provided in hard copy or electronic format. The electronic format is desirable because most

DART reports require a second round of reduction. Researchers should consult with the DART data analyst and specify their requirements before the data reduction begins. Once reduced using the DART, data are generally imported into spreadsheet software and a statistical software package for analysis and testing.

The CDR tapes are reduced using the ARTS computer maintained by AOS-400. The Test Director should coordinate with this organization to arrange for the reduction and analysis of these data. As with SAR tapes, the output of CDR reductions can be provided in hard copy or electronic format. The reports usually must be reduced further using custom-written software. Researchers should consult with the CDR data analyst and specify their requirements before the data reduction begins. Once reduced using the DART, data are generally imported into spreadsheet software and a statistical software package for analysis and testing.

Data from the VSCS are provided in a relatively raw electronic format from the Log Recorder. ACT-530 has developed techniques for reducing these data into a more useful format. As of this writing, the VSCS data specialists are in the process of improving the data analysis capability, and we expect more capabilities in the future.

Data from the WAKs are recorded in a spreadsheet file on the laptop hard disk. Data are organized by position and by prompt (though this can be modified if required). This file can be easily imported into spreadsheet software and requires no second-level reductions.

5.2 Manual Techniques

Data from paper questionnaires must be entered manually into a spreadsheet or statistical analysis package. The entered data must then be thoroughly checked for accuracy. It is advantageous for several people to enter data, each checking the others' work for errors and wrong assumptions.

In addition, some manual reduction of videotape data may be necessary. In the ARTS IIIA Baseline, ACT-530 prepared a videotape containing clips of the 10 minutes before and after every event that was counted as an operational error by the TGF. An SME from Boston TRACON viewed these clips and determined which were true operational errors and which resulted from the simulation environment or the data analysis.

5.3 Quality Assurance

Quality assurance is an essential element of a successful baseline. Without it, the data, the analyses, and the conclusions drawn from them are called into question. Regardless of the experience or ability of a researcher or data analyst, small errors can still be introduced into the data. Researchers should take all necessary steps to ensure the integrity of the baseline data and of any analyses performed.

Because the amount of data generated by a system baseline is enormous, we cannot recommend an audit of every data point. Instead, we recommend that Researchers conduct a spot check for each baseline measure. An engineering research psychologist who was not closely involved with the original data reduction or analysis should conduct the audit. The original data analyst should provide the auditor with the definitions of each measure, the assumptions made in the analysis of

each measure, and the files from which each measure was originally calculated. The auditor should select one data point for each measure and attempt to re-create that data point. If the auditor cannot re-create a data point, the original analyst and auditor should examine the data files, calculations, and assumptions to determine the cause of the discrepancy.

5.4 Archiving

Baseline data should be carefully archived to ensure that it is available for use in the future. Researchers should follow the requirements of the Project Configuration Management Guidelines (FAA, 1996). Researchers should archive copies of all questionnaires, raw electronic data (SAR tapes, CDR tapes, etc.), reduced electronic data (spreadsheet files, statistical package routines, etc.) and videotapes. Researchers should also write a short document that is archived along with the data, explaining what is contained on each tape and disk. The Test Director should obtain a list of applicable configuration parameters from the TGF, simulation laboratory, communications platform, and data reduction and analysis personnel. These information lists should be archived along with the data. These explanations will be invaluable to future researchers trying to re-create analyses or use data from previous studies in new comparisons.

6. Methodology for Comparing Systems

6.1 Operational Review Team

In future comparisons between systems, we recommend that researchers convene an Operational Review Team that will meet at the Technical Center for a period of several weeks. The Review Team should consist of

a. The engineering research psychologists who designed and conducted the baseline;
b. The Air Traffic SMEs from the field, typically the union representatives to the program;
c. two to four controller participants from the baseline;
d. technical SMEs for the data reduction and analysis tools;
e. technical SMEs for the simulator and laboratory platform; and
f. technical SMEs for the systems being compared.

The purposes of the Operational Review Team are

a. to ensure that the data and the analyses are accurate and complete,
b. to provide operational rationales for any differences found between systems, and
c. to assist in detailed data analysis such as reviewing videotapes of operational errors to determine their cause.

In the PVD-DSR Comparison, the team first reviewed a slide presentation showing comparisons between the two systems for every baseline measure. The definitions and analyses of each measure were provided, and the team members were encouraged to ask questions about how each measure was collected and analyzed. The team then reviewed each measure in detail, discussing the propriety of each analysis and requesting additional analyses if needed. In some

cases, the team determined that particular comparisons were invalid and that the baseline measure should not be reported.

The team also provided valuable data analysis expertise by reviewing videotapes of operational errors. They then determined if the error was truly caused by controller performance or was an artifact of the simulator environment.

Finally, the team provided operational rationales for any differences observed between the DSR and HCS-PVD. For example, this analysis revealed that controllers in the DSR Baseline made many more data block positioning actions than in the PVD Baseline. Because team members had participated in both baselines, they were able to explain that the data blocks in the DSR created more obscuration than the PVD, and they needed to move the data blocks more frequently. These sorts of operational rationales are invaluable to researchers when trying to account for differences between systems.

6.2 Reporting style

In general, we recommend that researchers report data from baselines at three levels of detail: overall, sector, and interval. The level or levels at which researchers should report baseline measures are listed in Section 2. The Overall Level provides data for the entire study collapsed across runs, sectors, positions, and intervals. It also provides data that are not collected every run such as from the Background and Final Questionnaires. The Sector Level provides data for each sector collapsed across runs and intervals. The Interval Level provides data for each 12-minute interval for each section.

We recommend that comparisons between systems be reported in both tabular and graphical forms (see Table 2 and Figure 2). Personnel with limited statistics backgrounds often do not understand detailed analyses, and graphics provide them with the information that they need. Tabular data provides readers with more statistics experience with additional details and allows them to conduct analyses on their own.

All participant written comments should be included as an appendix to the report. No identifying information such as the participant names should be included in this appendix. Researchers should report participant comments in an appendix with editing only for spelling and accuracy. Researchers should try to summarize comments in the text and can use direct quotes to illustrate points.

Table 2. Averages for Sectors

Construct	Baseline Measure	DSR 26	PVD 26	DSR 27	PVD 27	DSR 35	PVD 35	DSR 38	PVD 38	Comment
Safety	Data Block Positioning	76.0	42.7	111.0	57.8	123.6	85.0	64.0	32.3	See tables 8-11 for time interval data

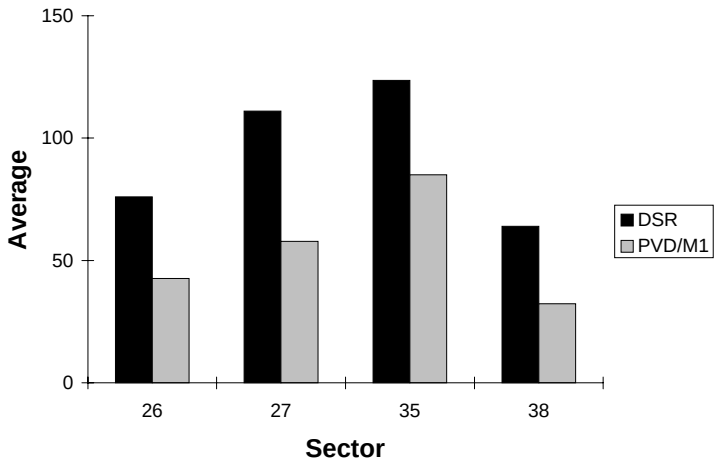

Figure 2. Average data block positioning actions per sector.

7. Using System Baseline Data

System baselines are one part of a larger process of human factors evaluations conducted throughout the system lifecycle. Baselines should not be the first or only human factors evaluation of a system nor should they be relied upon to identify all human factors problems. Baselines are not well suited to support task analyses or system specification development. Baselines are also not well suited to address detailed design issues such as how a control operates or which colors should be used. These issues are better examined in small-scale activities such as structured walkthroughs and part-task evaluations that allow researchers to focus on specific issues and allow run-offs between alternatives. These should be completed early in the acquisition process so that problems can be corrected while the impact to cost and schedule is still low.

System baseline data allow researchers to compare the system first to the system it replaces and then to subsequent modifications to system hardware, software, procedures, or adaptation. Comparisons between baseline data help ensure that the system provides a benefit over the system it replaces and continues to improve as modifications are made to it. Comparisons may also identify aspects of the system that would benefit from future modifications. Figure 3 shows a process of human factors evaluations, including the baselines, that can be conducted once a fairly mature system engineering baseline is available. This process specifically does not describe human factors activities such as task analyses that should be conducted in support of system specification development. For guidance about human factors activities conducted before a system engineering baseline is available, consult the Human Factors Job Aid (FAA, 1997). For another description of the role of baselines in the larger ATC acquisitions process, consult Keegan, Skiles, Krois, and Merkle (1996).

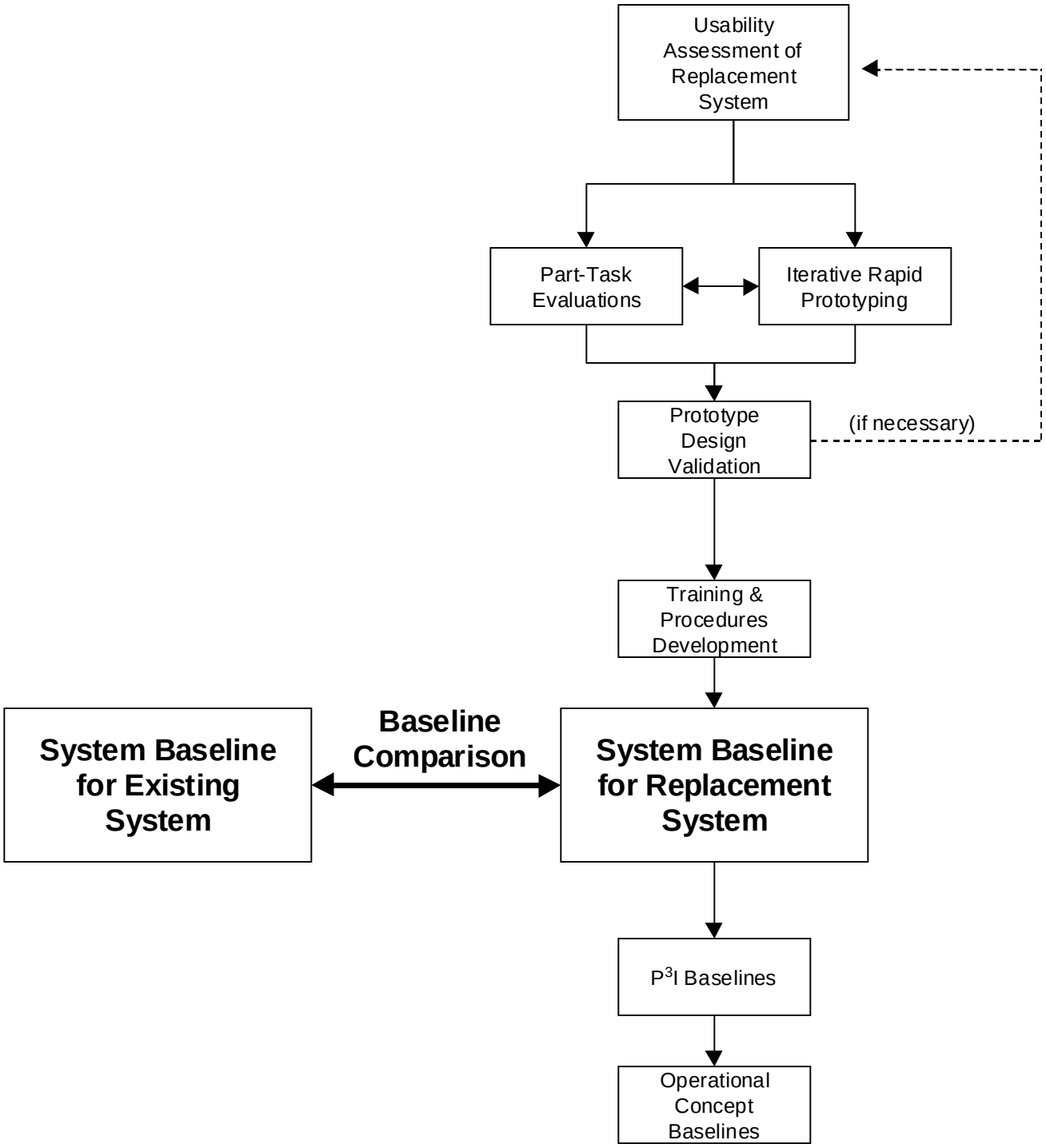

Figure 3. A process of human factors evaluations that can be conducted once the system engineering baseline is available.

7.1 Usability Assessment

A Usability Assessment (UA) is a medium-scale, human-in-the-loop simulation activity that is conducted soon after the engineering baseline becomes available to identify human factors issues. The UA does not use the baseline operational constructs nor does it require the extensive data collection and simulation realism of system baselines. Instead, researchers and SMEs develop a script of ATC activities that are relevant to the new system. These activities are designed to exercise the capabilities of the new system and to allow the participants to see and interact with it. The participants complete each of these activities using the new system under low-to-moderate traffic conditions. The participants are told that they are to focus on completing the scripted activities and that controlling the simulated traffic should not be their focus. As the participants complete the scripted activities, they provide feedback to human factors specialists about how successful they were. The human factors specialists then consolidate and categorize the participants feedback into a list of issues. This list guides the subsequent prototyping and part-task activities.

7.2 Part-Task Evaluations and Iterative Rapid Prototyping

Iterative Rapid Prototyping and Part-Task Evaluations are a series of activities conducted to develop and evaluate solutions to the issues identified in the UA. A multidisciplinary prototype team is convened containing human factors specialists, hardware and software engineers, prototype developers, and user representatives. The team categorizes the issues into several design threads such as target displays, console controls, and data entry. The team generates ideas that address the issues comprising each design thread. The prototype developers then implement these ideas into a realistic emulation prototype that allows rapid modification. Team members then are given the opportunity to see and interact with the prototype and to refine the design further. The success of each design is evaluated through small-scale, part-task evaluations that focus on the specific design thread. These evaluations allow precise measurement of speed, accuracy, heads-down time, reach envelopes, viewing angles, readability, and so on. The lessons learned from these part-task evaluations are incorporated into the prototype, and the part-task evaluations are repeated if necessary to assess design readiness.

7.3 Prototype Design Validation

The Prototype Design Validation is conducted after all the prototype designs have been evaluated and refined. The purpose of this validation is to ensure that the prototype designs work as a cohesive system. The validation is similar in form to the UA, with the participants completing a series of scripted actions and providing feedback to human factors specialists. Ideally, the participants in this activity are the participants from the UA. If necessary, feedback from this evaluation can be given to the prototype team to further refine and improve the prototype.

7.4 Training and Procedures Development

All new technology requires some training and changes to existing procedures. In this phase, human factors specialists work with personnel from the Air Traffic Operations (ATO) and Air

Traffic Resource Management (ATX) Organizations to develop procedures and training that ease the transition to the new equipment.

In most cases, the potential human factors contribution to this activity will focus on mitigating the effects of negative transfer. Negative transfer is a performance decrement that occurs when skills or experience from one work environment contributes to human error in a new environment (Cardosi & Murphy, 1995). Negative transfer is of greatest concern in areas where controllers rely on well learned, nearly automatic actions and procedures such as data entries and display control modifications. Controllers are so experienced with these actions on their current equipment that they may have difficulty learning new procedures, especially under conditions of high volume or complexity. Human factors specialists, following the results of the UA and the prototyping activities, can provide input as to how to minimize this sort of problem.

In other cases, new training and procedures can mitigate the effects of a system design deficiency. Though intended to resolve all system deficiencies identified in the UA, it is possible that some of the solutions developed during the prototyping phase cannot be supported given cost and schedule considerations. As a result, some system deficiencies may remain at various stages of system deployment. Human factors specialists, following the results of the UA and the prototyping activities, can identify possible effects of those deficiencies on controller performance and workload.

7.5 System Baselines

System baselines are a high fidelity, human-in-the-loop simulation of ATC operations with many objective and subjective measures. These baselines provide data following the five operational constructs: safety, capacity, performance, workload, and usability. The data can be used to compare to the existing system and the replacement system. Comparisons are reviewed by an Operational Review Team consisting of psychologists, air traffic SMEs, and the participants from the study. The team identifies problems with the comparison and provides operationally meaningful explanations for any difference between systems. The focus of this evaluation is to ensure that the system provides a benefit over the system it replaces along the constructs and to identify areas where the new system is deficient. The data collected in the baselines guide further refinements to hardware, software, training, or procedures after deployment.

7.6 Pre-Planned Product Improvements Baseline Studies

After the system is deployed, the system baseline data serve as a basis for studying the effects of Pre-Planned Product Improvements (P^3I). P^3I are new system capabilities that were still under development at system deployment but are already scheduled and included as part of the program. Because the effort and expense of a baseline simulation are high, we recommend that baselines be conducted only for major P^3I or for a set of multiple, minor P^3I. For example, the upcoming Initial Conflict Probe (ICP) will provide major new capabilities (e.g., conflict prediction and resolution) to the baseline DSR system. The ICP will require not only major changes to hardware and software but also to how controllers work and interact with each other. Such a major change is suitable for a P^3I Baseline. Minor P^3I should be addressed through iterative rapid prototyping and part-task evaluations rather than full-scale baselines.

In these studies, researchers use the baseline scenarios, procedures, and the participants again but now also using the P^3I. Data collected from these baselines are compared directly to the system baseline data, and determinations are made about changes in safety, capacity, performance, workload, and usability resulting from the introduction of the P^3I. For example, a P^3I Baseline might show that the P^3I substantially improves system capacity while only moderately increasing controller workload.

As with system baselines, the P^3I Baselines should only be conducted using mature equipment and should not be used for design prototyping, requirement development, human-computer interface design, and so on. These are best addressed in small-scale prototyping and part-task evaluations conducted earlier in the acquisition process for the P^3I.

7.7 Operational Concept Baselines

As in other baselines, these studies examine the effect on safety, capacity, performance, workload, and usability of a proposed change in operational concept. A change in operational concept is a major procedural change or a set of multiple minor changes that affects what ATCSs do, especially their roles and responsibilities. The shared separation responsibility concept and the reduced vertical separation minima projects are good examples. Again, because the effort and expense of a baseline are high, we do not recommend a baseline-level simulation for most procedural changes that may be undertaken by a facility. Instead, these are better addressed through smaller-scale simulations that focus on the particular procedure change and its effects.

In these studies, the participants work the baseline scenarios with the original equipment but while operating under different procedures. Because of tight control over the simulation environment, data from these baselines can be compared to the replacement system baseline where the original procedures were in effect.

Like new equipment, baselines examining the effects of new procedures should use new procedures that are mature and developed. Small-scale, part-task evaluations and fast-time modeling may be more appropriate to test small modifications to the procedure.

8. Conclusion

The Methodology Guide provides information for researchers involved in ATC system baselines. The authors would like to extend an invitation to all readers and users of the Methodology Guide to submit their own lessons learned and information for inclusion in future editions of the Methodology Guide. These suggestions and information should be sent to

Air Traffic Control System Baseline Methodology Guide
NAS Human Factors Branch, ACT-530
Federal Aviation Administration
William J. Hughes Technical Center, Building 28
Atlantic City International Airport, NJ 08405

References

Allendoerfer, K. R., Mogford, R. H., & Galushka, J. J. (1999*). Comparison of the Plan View Display and Display System Replacement system baselines.* Manuscript in preparation.

Cardosi, K. M., & Murphy, E. D. (1995). *Human factors in the design and evaluation of air traffic control systems* (DOT/FAA/RD-95/3). Washington, DC: Federal Aviation Administration Office of Aviation Research.

Federal Aviation Administration. (1996). *Project configuration management guidelines*. Unpublished manuscript. Atlantic City International Airport, NJ: DOT/FAA William J. Hughes Technical Center.

Federal Aviation Administration. (1997). *Human factors job aid*. Washington, DC: Office of Chief Scientific and Technical Advisory for Human Factors.

Galushka, J., Frederick, J., Mogford, R., & Krois, P. (1995). *Plan View Display baseline research report* (DOT/FAA/CT-TN95/45). Atlantic City International Airport, NJ: DOT/FAA Technical Center.

Guttman, J. A., Stein, E. S., & Gromelski, S. (1995). *The influence of generic airspace on air traffic controller performance* (DOT/FAA/CT-TN95/38). Atlantic City International Airport, NJ: DOT/FAA Technical Center.

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

Hering, H., & Coatleven, G. (1996). *ERGO (Version 2) For instantaneous self assessment of workload in a real-time ATC simulation environment* (Report No. 10/96). Brétigny-sur-Orge Cedex, France: Eurocontrol Experimental Centre.

Keegan, C., Skiles, T., Krois, P., & Merkle, M. (1996). Baseline measurement approach to ATC acquisitions. *The Journal of Air Traffic Control*, *38*, 33-37.

Krois, P., & Marsden, A. (1997*). FAA ODID VI: En route baseline comparison simulation final report* (Report No. 311). Brétigny-sur-Orge Cedex, France: Eurocontrol Experimental Centre.

Mogford, R. H., Allendoerfer, K. R., & Galushka, J. J. (1999). *ARTS IIIA terminal baseline research report* (DOT/FAA/CT-TN99/7). Atlantic City International Airport, NJ: DOT/FAA William J. Hughes Technical Center.

Skiles, T., Graham, R., Marsden, A., & Krois, P. (1997). En route ODID-PVD baseline comparisons. *The Journal of Air Traffic Control*, *39*, 38-41.

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 International Airport, NJ: DOT/FAA William J. Hughes Technical Center.

Stein, E. S. (1985). *Air traffic controller workload: An examination of workload probe* (DOT/FAA/CT-TN84/24). Atlantic City International Airport, NJ: DOT/FAA Technical Center.

Acronyms

AMP	Aircraft Management Program
AOC	Airline Operations Center
ARINC	Aeronautical Radio, Incorporated
ARTCC	Air Route Traffic Control Center
ARTS	Automated Radar Terminal System
ATC	Air Traffic Control
ATCS	Air Traffic Control Specialist
ATWIT	Air Traffic Workload Input Technique
CDR	Continuous Data Recording
DART	Data Analysis and Reduction Tool
DEDS	Data Entry and Display Subsystem
DRAT	Data Reduction and Analysis Tool
DSR	Display System Replacement
DYSIM	Dynamic Simulation
EARTS	En Route Automated Radar Tracking System
EDC	Early Display Capability
ESSF	En Route Simulation Support Facility
ETG	Enhanced Target Generator
ETVS	Enhanced Terminal Voice Switch
FAA	Federal Aviation Administration
FACO	Facility Control Office
FDAD	Full Digital ARTS Display
FPL	Full Performance Level
HCS	Host Computer System
I^2F	Integration and Interoperability Facility
ICP	Initial Conflict Probe
ISC	Initial System Capability
KDR	Keyboard Data Recorder
LOA	Letter of Agreement
MERR	Mobile Experimental Recording Rack
NATCA	National Air Traffic Controllers Association
ODID	Operational Display and Input Development
ODL	Oceanic Data Link
OT&E	Operational Test & Evaluation
P^3I	Pre-Planned Product Improvements
PTT	push-to-talk
PVD	Plan View Display
SAR	System Analysis Recording
SATORI	Systematic Air Traffic Operations Research Initiative
SCIP	Surveillance Communications Interface Processor
SME	Subject Matter Expert
STARS	Standard Terminal Automation Replacement System
TGF	Target Generation Facility
TRACON	Terminal Radar Approach Control
TSSF	Terminal Simulation Support Facility

UA	Usability Assessment
VSCS	Voice Switching and Control System
WAK	Workload Assessment Keypad
ZDC	Washington ARTCC

Appendix A
Questionnaires

Notes:

The following questionnaires represent the most recent versions of the baseline questionnaires. Because ACT-530 is constantly revising and improving these questionnaires, the items and wording contained here do not necessarily correspond to those used in earlier baselines. We recommend that researchers interested in comparing data to earlier baselines examine the questionnaires used in the earlier baseline to determine what changes and refinements have been made and these changes will affect validity.

When using these questionnaires, researchers should replace the pseudonym "ATCView System" with the name of the system they are researching. In addition, other revisions to these questionnaires will be necessary to tailor them to the specific system in question. We have purposely included more information on these questionnaires, particularly the Background Questionnaire, than will be necessary in every baseline. Some areas that are likely to need revision for future baselines are marked with brackets and bold characters. Example: [include specifics here]

BACKGROUND QUESTIONNAIRE

Participant Code: _____ Date: _____

Controller Team: _____

Instructions

The purpose of this questionnaire is to obtain information about your experience and background as an air traffic controller. We will use this information to describe the participants in this baseline as a group rather than as individuals. So that your identity can remain anonymous, please do not write your name anywhere on this form. The data you provide on this questionnaire, as with all data you provide for this study, will be identified only by a participant code known only to you and the experimenters.

---

1) What is your age?

_____ years

2) What is your current position as an air traffic controller?

❑ Full Performance Level ❑ Other (specify) _____

3) How many of the past 12 months have you actively controlled traffic?

_____ months

4) Please indicate the number of years experience you have in the following air traffic control domains.

En Route: ____ Terminal: ____ Tower: ____

Oceanic: ____ Military: ____

5) Please indicate the number of years experience you have using the following air traffic control systems.

Host: _____ ARTS: _____ EARTS: _____

STARS: _____

BACKGROUND QUESTIONNAIRE (CONTINUED)

6) Please indicate the number of years experience you have using the following radar display systems.

PVD/M1: _____ FDAD: _____ DEDS: _____

DSR: _____ STARS: _____

7) If you wear corrective lenses, will you wear them during the simulations?

❑ Yes ❑ No ❑ I don't wear corrective lenses

8) Circle the number that best describes your current state of health.

1	2	3	4	5	6	7
Not Very Healthy						Extremely Healthy

9) How many hours of training and experience have you received using the ATCView System?

_____ hours

10) Circle the number which best describes your level of satisfaction with the ATCView System.

1	2	3	4	5	6	7
Not Very Satisfied						Extremely Satisfied

11) Circle the number which best describes your level of experience with personal computers.

1	2	3	4	5	6	7
Not Very Experienced						Extremely Experienced

POST-RUN QUESTIONNAIRE

Participant Code: _____ Date: _____

Controller Team: _____

Sector: 26 38 27 35 Run: 1 2 3 4

Position: Radar Data

Instructions

The purpose of this questionnaire is to obtain information about the simulation you just completed. We will use this information to determine how the simulation experience affected your opinions. As you answer each question, feel free to use the entire numerical scale. Please be as honest and as accurate as you can. So that your identity can remain anonymous, please do not write your name anywhere on this form and use only your participant code.

---

1) How well did you control traffic during this problem?

1	2	3	4	5	6	7
Not Very Well						Extremely Well

2) What was your average workload level during this problem?

1	2	3	4	5	6	7
Very Low Workload						Very High Workload

3) How difficult was this problem compared to other simulation training problems?

1	2	3	4	5	6	7
Not Very Difficult						Extremely Difficult

4) How good do you think your air traffic control services were from a pilot's point of view?

1	2	3	4	5	6	7
Not Very Good						Extremely Good

POST-RUN QUESTIONNAIRE (CONTINUED)

5) To what extent did technical problems with the simulation equipment interfere with your ability to control traffic?

1	2	3	4	5	6	7
Not Very Much						A Great Deal

6) To what extent did problems with pseudopilots interfere with your normal air traffic control activities?

1	2	3	4	5	6	7
Not Very Much						A Great Deal

7) How realistic was this simulation problem compared to actual air traffic control?

1	2	3	4	5	6	7
Not Very Realistic						Extremely Realistic

OBSERVER LOG

Observer: _____ Date: _____

Sector: 26 38 35 27 Run: 1 2 3 4

Instructions

Please record any unusual events by noting the system time, the nature of the event, and the aircraft involved. Please also note any technical problems and other safety-critical or otherwise important events. Use back of page for explanations, if necessary.

System Time	Event	Aircraft

FINAL QUESTIONNAIRE

Participant Code: _____ Date: _____

Controller Team: _____

Instructions

The purpose of this questionnaire is to obtain information about the [ATCView System] overall. We will use this information to determine how effectively the system performed during this simulation. As you answer each question, feel free to use the entire numerical scale. Please be as honest and as accurate as you can. So that your identity can remain anonymous, please do not write your name on this form and use only your participant code.

Section A

Please circle the number which best describes your level of agreement with each of the following statements concerning the [ATCView System].

1) The flight progress strips are easy to access in the strip bays.

1	2	3	4	5	6	7
Strongly Disagree						Strongly Agree

2) The flight progress strips are easy to read and mark in the strip bays.

1	2	3	4	5	6	7
Strongly Disagree						Strongly Agree

3) The [on-screen] controls are easy to access.

1	2	3	4	5	6	7
Strongly Disagree						Strongly Agree

4) The operation and functions of the [on-screen] controls are intuitive.

1	2	3	4	5	6	7
Strongly Disagree						Strongly Agree

5) The controller keyboard is easy to use.

1	2	3	4	5	6	7
Strongly Disagree						Strongly Agree

6) The radar and map displays are easy to read.

1	2	3	4	5	6	7
Strongly Disagree						Strongly Agree

7) The radar and map displays are easy to understand.

1	2	3	4	5	6	7
Strongly Disagree						Strongly Agree

8) There is plenty of space to work within the workstation.

1	2	3	4	5	6	7
Strongly Disagree						Strongly Agree

9) The equipment, displays, and controls allow me to control traffic in the most efficient way possible.

1	2	3	4	5	6	7
Strongly Disagree						Strongly Agree

10) The equipment, displays, and controls allow me to control traffic without any awkward limitations.

1	2	3	4	5	6	7
Strongly Disagree						Strongly Agree

11) Overall, the equipment, displays and controls are effective in meeting the needs of controllers.

1	2	3	4	5	6	7
Strongly Disagree						Strongly Agree

Section B

Please circle the number which best describes your overall interaction with the equipment, displays, and controls of the [ATCView System].

1)	1	2	3	4	5	6	7
	Not Very Limited						Extremely Limited
2)	1	2	3	4	5	6	7
	Not Very Frustrating						Extremely Frustrating
3)	1	2	3	4	5	6	7
	Not Very Effective						Extremely Effective
4)	1	2	3	4	5	6	7
	Not Very Efficient						Extremely Efficient
5)	1	2	3	4	5	6	7
	Not Very Easy to Operate						Extremely Easy to Operate
6)	1	2	3	4	5	6	7
	Not Very Easy to Understand						Extremely Easy to Understand

Section C

This section should address any specific improvements that have been suggested during system development. The items in this section should follow the format below.

Please circle the number which best represents your opinion about the following potential improvements to the [ATCView System].

1) To what extent do you think [possible improvement] would improve your effectiveness with the [ATCView System]?

1	2	3	4	5	6	7
Not Very Much						A Great Deal

Section D

For each the following questions, indicate your opinion by marking one or more of the provided boxes. Then, please provide any additional comments that you think are appropriate.

1) Which aspects of the [ATCView System] need improvement?

- ❑ Radar and Map Displays
- ❑ Flight Strip Bays
- ❑ Keyboard
- ❑ Trackball
- ❑ On-Screen Controls
- ❑ Volume of Workspace
- ❑ Other (specify) _____
- ❑ Other (specify) _____

Please provide some details about why you think each of these aspects needs improvement.

_____

_____

_____

_____

_____

_____

_____

2) What are the most common mistakes you encountered using the [ATCView System]?

- ❑ Misreading Radar Display Information
- ❑ Misreading Map Display Information
- ❑ Misreading Flight Progress Strips
- ❑ Making Entries with Keyboard
- ❑ Selecting Targets with Trackball
- ❑ Adjusting On-screen Controls
- ❑ Other (specify) _____
- ❑ Other (specify) _____

Please provide some details about what you think caused you to make each of these mistakes.

_____

_____

_____

_____

_____

_____

3) Please comment on the positive aspects of the system.

Section E

If there are any other comments or suggestions that you have regarding this baseline study of the [ATCView System], please write your ideas in the space provided below.

Section F

During this baseline study, you have used the Workload Assessment Keypad (WAK) to rate your workload during the simulation runs. This technique is known as the Air Traffic Workload Input Technique (ATWIT), which has been extensively researched at the Technical Center. Please indicate below how you defined the lowest (1) and highest (7) workload rating on the scale.

To me, the lowest ATWIT rating (1) meant my workload was:

To me, the highest ATWIT rating (7) meant my workload was:

Did responding to the WAK prompts interfere with performing your primary function?

1	2	3	4	5	6	7
Not at all						A great deal

SUBJECT MATTER EXPERT
OBSERVER RATING FORM
FOR EN ROUTE OPERATIONS

Observer Code: _____ Date: _____

Instructions

This form is designed to be used by supervisory controllers to evaluate the effectiveness of controllers working in simulation environments. SATCSs will observe and rate the performance of controllers in several different performance dimensions using the scale below as a general purpose guide. Use the entire scale range as much as possible. You will see a wide range of controller performance. Take extensive notes on what you see. Do not depend on your memory. Write down your observations. Space is provided after each scale for comments. You may make preliminary ratings during the course of the scenario. However, wait until the scenario is finished before making your final ratings and remain flexible until the end when you have had an opportunity to see all the available behavior. At all times please focus on what you actually see and hear. This includes what the controller does and what you might reasonably infer from the actions of the pilots. Try to avoid inferring what you think may be happening. If you do not observe relevant behavior or the results of that behavior, then you may leave a specific rating blank. Also, please write down any comments that may help improve this evaluation form. Do not write your name on the form itself. Your identity will remain anonymous, as your data will be identified by an observer code known only to yourself and researchers conducting this study. The observations you make do not need to be restricted to the performance areas covered in this form and may include other areas that you think are important.

Assumptions

ATC is a complex activity that contains both observable and unobservable behavior. There are so many complex behaviors involved that no observational rating form can cover everything. A sample of the behaviors is the best that can be achieved, and a good form focuses on those behaviors that controllers themselves have identified as the most relevant in terms of their overall performance. Most controller performance is at or above the minimum standards regarding safety and efficiency. The goal of the rating system is to differentiate performance above this minimum. The lowest rating should be assigned for meeting minimum standards and also for anything below the minimum since this should be a rare event. It is important for the observer/rater to feel comfortable using the entire scale and to understand that all ratings should be based on behavior that is actually observed.

Rating Scale Descriptors

Remove this Page and keep it available while doing ratings

SCALE	QUALITY	SUPPLEMENTARY
1	Least Effective	Unconfident, Indecisive, Inefficient, Disorganized, Behind the power curve, Rough, Leaves some tasks incomplete, Makes mistakes
2	Poor	May issue conflicting instructions, Doesn't plan completely
3	Fair	Distracted between tasks
4	Low Satisfactory	Postpones routine actions
5	High Satisfactory	Knows the job fairly well
6	Good	Works steadily, Solves most problems
7	Very Good	Knows the job thoroughly, Plans well
8	Most Effective	Confident, Decisive, Efficient, Organized, Ahead of the power curve, Smooth, Completes all necessary tasks, Makes no mistakes

I - MAINTAINING SAFE AND EFFICIENT TRAFFIC FLOW

1. Maintaining Separation and Resolving Potential Conflicts 1 2 3 4 5 6 7 8
   - using control instructions that maintain appropriate aircraft and airspace separation
   - detecting and resolving impending conflicts early
   - recognizing the need for speed restrictions and wake turbulence separation

   Comments:

2. Sequencing Aircraft Efficiently 1 2 3 4 5 6 7 8
   - using efficient and orderly spacing techniques for arrival, departure, and en route aircraft
   - maintaining safe arrival and departure intervals that minimize delays

   Comments:

3. Using Control Instructions Effectively/Efficiently 1 2 3 4 5 6 7 8
   - providing accurate navigational assistance to pilots
   - issuing economical clearances that result in need for few additional instructions to handle aircraft completely
   - ensuring clearances use minimum necessary flight path changes

   Comments:

4. Overall Safe and Efficient Traffic Flow Scale Rating 1 2 3 4 5 6 7 8

   Comments:

II - MAINTAINING ATTENTION AND SITUATION AWARENESS

5. Maintaining Awareness of Aircraft Positions..... 1 2 3 4 5 6 7 8
   - avoiding fixation on one area of the radar scope when other areas need attention
   - using scanning patterns that monitor all aircraft on the radar scope

   Comments:

6. Ensuring Positive Control..... 1 2 3 4 5 6 7 8
   - tailoring control actions to situation
   - using effective procedures for handling heavy, emergency, and unusual traffic situations

   Comments:

7. Detecting Pilot Deviations from Control Instructions..... 1 2 3 4 5 6 7 8
   - ensuring that pilots follow assigned clearances correctly
   - correcting pilot deviations in a timely manner

   Comments:

8. Correcting Own Errors in a Timely Manner..... 1 2 3 4 5 6 7 8
   - acting quickly to correct errors
   - changing an issued clearance when necessary to expedite traffic flow

   Comments:

9. Overall Attention and Situation Awareness Scale Rating..... 1 2 3 4 5 6 7 8

   Comments:

III – PRIORITIZING

10. Taking Actions in an Appropriate Order of Importance.....1 2 3 4 5 6 7 8
    - resolving situations that need immediate attention before handling low priority tasks
    - issuing control instructions in a prioritized, structured, and timely manner

    Comments:

11. Preplanning Control Actions..... 1 2 3 4 5 6 7 8
    - scanning adjacent sectors to plan for future and conflicting traffic
    - studying pending flight strips in bay

    Comments:

12. Handling Control Tasks for Several Aircraft..... 1 2 3 4 5 6 7 8
    - shifting control tasks between several aircraft when necessary
    - communicating in timely fashion while sharing time with other actions

    Comments:

13. Marking Flight Strips while Performing Other Tasks 1 2 3 4 5 6 7 8
    - marking flight strips accurately while talking or performing other tasks
    - keeping flight strips current

    Comments:

14. Overall Prioritizing Scale Rating..... 1 2 3 4 5 6 7 8

    Comments:

IV – PROVIDING CONTROL INFORMATION

15. Providing Essential Air Traffic Control Information..... 1 2 3 4 5 6 7 8
    - providing mandatory services and advisories to pilots in a timely manner
    - exchanging essential information

    Comments:

16. Providing Additional Air Traffic Control Information..... 1 2 3 4 5 6 7 8
    - providing additional services when workload is not a factor
    - exchanging additional information

    Comments:

17. Providing Coordination..... 1 2 3 4 5 6 7 8
    - providing effective and timely coordination
    - using proper point-out procedures

    Comments:

18. Overall Providing Control Information Scale Rating.....1 2 3 4 5 6 7 8

    Comments:

V – TECHNICAL KNOWLEDGE

19. Showing Knowledge of LOAs and SOPs 1 2 3 4 5 6 7 8
    - controlling traffic as depicted in current LOAs and SOPs
    - performing handoff procedures correctly

    Comments:

20a. Showing Knowledge of Aircraft Capabilities and Limitations..... 1 2 3 4 5 6 7 8
    - using appropriate speed, vectoring, and/or altitude assignments to separate aircraft with varied flight capabilities
    - issuing clearances that are within aircraft performance parameters

    Comments:

20b. Showing Effective Use of Equipment.....1 2 3 4 5 6 7 8
    - updating data blocks
    - using equipment capabilities

    Comments:

21. Overall Technical Knowledge Scale Rating..... 1 2 3 4 5 6 7 8

    Comments:

VI – COMMUNICATING

22. Using Proper Phraseology..... 1 2 3 4 5 6 7 8
    - using words and phrases specified in the 7110.65
    - using phraseology that is appropriate for the situation
    - using minimum necessary verbiage

    Comments:

23. Communicating Clearly and Efficiently 1 2 3 4 5 6 7 8
    - speaking at the proper volume and rate for pilots to understand
    - speaking fluently while scanning or performing other tasks
    - ensuring clearance delivery is complete, correct and timely
    - speaking with confident, authoritative tone of voice

    Comments:

24. Listening to Pilot Readbacks and Requests 1 2 3 4 5 6 7 8
    - correcting pilot readback errors
    - acknowledging pilot or other controller requests promptly
    - processing requests correctly in a timely manner

    Comments:

25. Overall Communicating Scale Rating 1 2 3 4 5 6 7 8

    Comments:

SUBJECT MATTER EXPERT
OBSERVER RATING FORM
FOR TERMINAL SIMULATIONS

Observer Code: _____ Date: _____

Instructions

This form is designed to be used by supervisory controllers to evaluate the effectiveness of controllers working in simulation environments. SATCSs will observe and rate the performance of controllers in several different performance dimensions using the scale below as a general purpose guide. Use the entire scale range as much as possible. You will see a wide range of controller performance. Take extensive notes on what you see. Do not depend on your memory. Write down your observations. Space is provided after each scale for comments. You may make preliminary ratings during the course of the scenario. However, wait until the scenario is finished before making your final ratings and remain flexible until the end when you have had an opportunity to see all the available behavior. At all times please focus on what you actually see and hear. This includes what the controller does and what you might reasonably infer from the actions of the pilots. Try to avoid inferring what you think may be happening. If you do not observe relevant behavior or the results of that behavior, then you may leave a specific rating blank. Also, please write down any comments that may help improve this evaluation form. Do not write your name on the form itself. Your identity will remain anonymous, as your data will be identified by an observer code known only to yourself and researchers conducting this study. The observations you make do not need to be restricted to the performance areas covered in this form and may include other areas that you think are important.

Assumptions

ATC is a complex activity that contains both observable and unobservable behavior. There are so many complex behaviors involved that no observational rating form can cover everything. A sample of the behaviors is the best that can be achieved, and a good form focuses on those behaviors that controllers themselves have identified as the most relevant in terms of their overall performance. Most controller performance is at or above the minimum standards regarding safety and efficiency. The goal of the rating system is to differentiate performance above this minimum. The lowest rating should be assigned for meeting minimum standards and also for anything below the minimum since this should be a rare event. It is important for the observer/rater to feel comfortable using the entire scale and to understand that all ratings should be based on behavior that is actually observed.

Rating Scale Descriptors

Remove this Page and keep it available while doing ratings

SCALE	QUALITY	SUPPLEMENTARY
1	Least Effective	Unconfident, Indecisive, Inefficient, Disorganized, Behind the power curve, Rough, Leaves some tasks incomplete, Makes mistakes
2	Poor	May issue conflicting instructions, Doesn't plan completely
3	Fair	Distracted between tasks
4	Low Satisfactory	Postpones routine actions
5	High Satisfactory	Knows the job fairly well
6	Good	Works steadily, Solves most problems
7	Very Good	Knows the job thoroughly, Plans well
8	Most Effective	Confident, Decisive, Efficient, Organized, Ahead of the power curve, Smooth, Completes all necessary tasks, Makes no mistakes

I - MAINTAINING SAFE AND EFFICIENT TRAFFIC FLOW

1. Maintaining Separation and Resolving Potential Conflicts 1 2 3 4 5 6 7 8
   - using control instructions that maintain appropriate aircraft and airspace separation
   - detecting and resolving impending conflicts early
   - recognizing the need for speed restrictions and wake turbulence separation

   Comments:

2. Sequencing Aircraft Efficiently 1 2 3 4 5 6 7 8
   - using efficient and orderly spacing techniques for arrival and departure aircraft
   - maintaining safe arrival and departure intervals that minimize delays

   Comments:

3. Using Control Instructions Effectively/Efficiently 1 2 3 4 5 6 7 8
   - providing accurate navigational assistance to pilots
   - issuing economical clearances that result in need for few additional instructions to handle aircraft completely
   - ensuring clearances use minimum necessary flight path changes

   Comments:

4. Overall Safe and Efficient Traffic Flow Scale Rating 1 2 3 4 5 6 7 8

   Comments:

II - MAINTAINING ATTENTION AND SITUATION AWARENESS

5. Maintaining Awareness of Aircraft Positions..... 1 2 3 4 5 6 7 8
   - avoiding fixation on one area of the radar scope when other areas need attention
   - using scanning patterns that monitor all aircraft on the radar scope

   Comments:

6. Ensuring Positive Control..... 1 2 3 4 5 6 7 8
   - tailoring control actions to situation
   - using effective procedures for handling heavy, emergency, and unusual traffic situations

   Comments:

7. Detecting Pilot Deviations from Control Instructions..... 1 2 3 4 5 6 7 8
   - ensuring that pilots follow assigned clearances correctly
   - correcting pilot deviations in a timely manner

   Comments:

8. Correcting Own Errors in a Timely Manner..... 1 2 3 4 5 6 7 8
   - acting quickly to correct errors
   - changing an issued clearance when necessary to expedite traffic flow

   Comments:

9. Overall Attention and Situation Awareness Scale Rating..... 1 2 3 4 5 6 7 8

   Comments:

III – PRIORITIZING

10. Taking Actions in an Appropriate Order of Importance.....1 2 3 4 5 6 7 8
    - resolving situations that need immediate attention before handling low priority tasks
    - issuing control instructions in a prioritized, structured, and timely manner

    Comments:

11. Preplanning Control Actions..... 1 2 3 4 5 6 7 8
    - scanning adjacent sectors to plan for future and conflicting traffic
    - studying pending flight strips in bay

    Comments:

12. Handling Control Tasks for Several Aircraft..... 1 2 3 4 5 6 7 8
    - shifting control tasks between several aircraft when necessary
    - communicating in timely fashion while sharing time with other actions

    Comments:

13. Marking Flight Strips while Performing Other Tasks 1 2 3 4 5 6 7 8
    - marking flight strips accurately while talking or performing other tasks
    - keeping flight strips current

    Comments:

14. Overall Prioritizing Scale Rating..... 1 2 3 4 5 6 7 8

    Comments:

IV – PROVIDING CONTROL INFORMATION

15. Providing Essential Air Traffic Control Information..... 1 2 3 4 5 6 7 8
    - providing mandatory services and advisories to pilots in a timely manner
    - exchanging essential information

    Comments:

16. Providing Additional Air Traffic Control Information..... 1 2 3 4 5 6 7 8
    - providing additional services when workload is not a factor
    - exchanging additional information

    Comments:

17. Providing Coordination..... 1 2 3 4 5 6 7 8
    - providing effective and timely coordination
    - using proper point-out procedures

    Comments:

18. Overall Providing Control Information Scale Rating.....1 2 3 4 5 6 7 8

    Comments:

V – TECHNICAL KNOWLEDGE

19. Showing Knowledge of LOAs and SOPs 1 2 3 4 5 6 7 8
    - controlling traffic as depicted in current LOAs and SOPs
    - performing handoff procedures correctly

    Comments:

20. Showing Knowledge of Aircraft Capabilities and Limitations..... 1 2 3 4 5 6 7 8
    - using appropriate speed, vectoring, and/or altitude assignments to separate aircraft with varied flight capabilities
    - issuing clearances that are within aircraft performance parameters

    Comments:

21. Overall Technical Knowledge Scale Rating 1 2 3 4 5 6 7 8

    Comments:

VI – COMMUNICATING

22. Using Proper Phraseology..... 1 2 3 4 5 6 7 8
    - using words and phrases specified in the 7110.65
    - using phraseology that is appropriate for the situation
    - using minimum necessary verbiage

    Comments:

23. Communicating Clearly and Efficiently 1 2 3 4 5 6 7 8
    - speaking at the proper volume and rate for pilots to understand
    - speaking fluently while scanning or performing other tasks
    - ensuring clearance delivery is complete, correct and timely
    - speaking with confident, authoritative tone of voice

    Comments:

24. Listening to Pilot Readbacks and Requests 1 2 3 4 5 6 7 8
    - correcting pilot readback errors
    - acknowledging pilot or other controller requests promptly
    - processing requests correctly in a timely manner

    Comments:

25. Overall Communicating Scale Rating 1 2 3 4 5 6 7 8

    Comments:

Appendix B

STATEMENT OF CONFIDENTIALITY AND INFORMED CONSENT

Researchers from the NAS System Engineering and Analysis Division (ACT-500) of the William J. Hughes Technical Center and its contractors maintain strict standards regarding participant confidentiality and informed consent. Our standards are based on the *Ethical Principles in the Conduct of Research with Human Participants* by the American Psychological Association. Our standards are structured around four main principles:

- Your participation is voluntary. You may withdraw from this research at any time without consequence. If you feel you must withdraw for whatever reason, please inform researchers immediately.

- Your responsibilities will be clear. Researchers will clearly explain what is expected of you during the simulation. They will answer any and all questions about the objectives of the research, the simulation design, and the data collection techniques.

- Your data will remain anonymous. Your responses will be identified by a code known only to you and the researchers. Your identity will be kept separate from the data you provide. To facilitate this, please do not write your name or any other identifying marks on the questionnaires. Please do not share your participant code with anyone other than the researchers. No names will be associated with data in any reports.

- Your data will be confidential. The *raw* data collected in this study will become the property of the NAS Human Factors Branch (ACT-530). The raw data will be analyzed by specialists from this organization and its contractor employees. The raw data will not be made available to other organizations without your permission. The *aggregate* data from this study will be published in a Technical Note by the William J. Hughes Technical Center, which will be distributed throughout the FAA and elsewhere. These data will take the form of averages, standard deviations, and other statistics.

Additional considerations for this baseline simulation:

- Please be aware that we are making video and audio recordings during the runs. The video cameras will be positioned above and behind you. At some sectors, a video camera will also be recording your hands as you type on the keyboard. Audio recordings will come from wireless microphones that you will wear during the simulation. If you strongly object to having yourself recorded in this way, please inform researchers immediately.
- Please be aware that we are making recordings of your keystrokes including any typographical errors. If you strongly object to having yourself recorded in this way, please inform researchers immediately.

Good research requires good data. We hope that by protecting your rights, we are encouraging you to be as accurate and honest in your responses as possible. If you have questions at any time regarding the study, researchers will be glad to answer them.

Thank you for your participation!

Appendix C

WORKLOAD ASSESSMENT KEYPAD
INSTRUCTIONS FOR THE PARTICIPANTS

One purpose of this research is to obtain an accurate evaluation of controller workload. By workload, we mean all the physical and mental effort that you must exert to do your job. This includes maintaining the "picture," planning, coordinating, decision making, communicating, and whatever else is required to maintain a safe and expeditious traffic flow.

The way you will tell us how hard you are working is by pressing the buttons numbered from 1 to 7 on the keypad located at your controller workstation. These buttons correspond to the following levels of workload. At the low end of the scale (1 and 2), your workload is low—you can accomplish everything easily. As the numbers increase, your workload is getting higher. Numbers 3 and 4 represent increasing levels of moderate workload where the chance of error is still low but steadily increasing. Numbers 5 and 6 reflect relatively high workload where there is some chance of making mistakes. The high end of the scale (7) represents a very high workload, where it is likely that you will have to leave some tasks incomplete.

Beginning at minute 10 of the simulation run, you will hear the keypad chirp and it will illuminate its lights. Please press the key of your choice as soon as possible and the lights will extinguish and the chirping will stop. The WAK will prompt again every four minutes. We realize that this requirement may be somewhat annoying at first, but please give it a chance for the purposes of this project.

All controllers, no matter how proficient and experienced, will be exposed at one time or another to all levels of workload. It does not detract from controllers' professionalism when they indicate that they are working very hard or that they are hardly working. Feel free to use the entire scale and tell us honestly how hard you are working!

Thank you again for your cooperation, and remember that this data is being collected without any information that could later be used to identify you. Your privacy is protected.

For more information about the ATWIT and measuring air traffic controller workload, we recommend *Air Traffic Controller Workload: An Examination of Workload Probe* by Earl S. Stein, FAA Technical Center Technical Note DOT/FAA/CT-TN84/24.