DOT/FAA/AM-01/5

Office of Aviation Medicine
Washington, D.C. 20591

Documentation of Validity for the AT-SAT Computerized Test Battery
Volume I.

R.A. Ramos
Human Resources Research Organization
Alexandria, VA 22314-1591

Michael C. Heil
Carol A. Manning

Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

March 2001

Final Report

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

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-01/5		

4. Title and Subtitle	5. Report Date
Documentation of Validity for the AT-SAT Computerized Test Battery, Volume I	March 2001
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Ramos, R.A.[1], Heil, M.C.[2], and Manning, C.A.[2]	

9. Performing Organization Name and Address		10. Work Unit No. (TRAIS)
[1]Human Resources Research Organization 68 Canal Center Plaza, Suite 400 Alexandria, VA 22314-1591	[2]FAA Civil Aeromedcial Institute P. O. Box 25082 Oklahoma City, OK 73125	11. Contract or Grant No.

12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aviation Medicine Federal Aviation Administration 800 Independence Ave., S. W. Washington, D.C. 20591	14. Sponsoring Agency Code

15. Supplemental Notes

Work was accomplished under approved subtask AM-B-99-HRR-517

16. Abstract

This document is a comprehensive report on a large-scale research project to develop and validate a computerized selection battery to hire Air Traffic Control Specialists (ATCSs) for the Federal Aviation Administration (FAA). The purpose of this report is to document the validity of the Air Traffic Selection and Training (AT-SAT) battery according to legal and professional guidelines. An overview of the project is provided, followed by a history of the various job analyses efforts. Development of predictors and criterion measures are given in detail. The document concludes with the presentation of the validation of predictors and analyses of archival data.

17. Key Words	18. Distribution Statement
Air Traffic Controllers, Selection, Assessment, Job Analyses	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	165	

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

ACKNOWLEDGMENTS

The editors thank Ned Reese and Jay Aul for their continued support and wisdom throughout this study. Also, thanks go to Cristy Detwiler, who guided this report through the review process and provided invaluable technical support in the final phases of editing.

TABLE OF CONTENTS
VOLUME I.

List of Figures and Tables

Figures

Tables

Appendices

CHAPTER 1

Air Traffic Selection and Training (AT-SAT) Project

Robert A. Ramos, HumRRO

INTRODUCTION

This document is a comprehensive report on a large-scale research project to develop and validate a computerized selection battery to hire Air Traffic Control Specialists (ATCSs) for the Federal Aviation Administration (FAA). The purpose of this report is to document the validity of the Air Traffic Selection and Training (AT-SAT) battery according to legal and professional guidelines. The Dictionary of Occupational Titles lists the Air Traffic Control Specialist Tower as number 193162018.

Background

The ATCS position is unique in several respects. On the one hand, it is a critically important position at the center of efforts to maintain air safety and efficiency of aircraft movement. The main purpose of the ATCS job is to maintain a proper level of separation between airplanes. Separation errors may lead to situations that could result in a terrible loss of life and property. Given the consequences associated with poor job performance of ATCSs, there is great concern on the part of the FAA to hire and train individuals so that air traffic can be managed safely and efficiently. On the other hand, the combination of skills and abilities required for proficiency in the position is not generally prevalent in the labor force. Because of these characteristics, ATCSs have been the focus of a great deal of selection and training research over the years.

Historical events have played a major role in explaining the present condition of staffing, selection and training systems for ATCSs. In 1981, President Ronald Reagan fired striking ATCSs. Approximately 11,000 of 17,000 ATCSs were lost during the strike. Individuals hired from August 1981 to about the end of 1984 replaced most of the strikers. A moderate level of new hires was added through the late 1980s. However, relatively few ATCSs have been hired in recent years due to the sufficiency of the controller workforce. Rehired controllers and graduates of college and university aviation training programs have filled most open positions. Starting in fiscal year 2005, a number of the post-1981 hires will start to reach retirement eligibility. As a consequence, there is a need for the Air Traffic Service to hire five to eight hundred ATCSs candidates a year for the next several years to maintain proper staffing levels. The majority of the new hires will have little background in ATCSs work. Further, it generally takes two to four years to bring ATCS developmentals to the full performance level (FPL).

In addition, the FAA Air Traffic Training Program has designed major changes in the staffing and training of new recruits for the ATCS position. In the past, training at the FAA Academy included aspects of a screening program. The newly developed AT-SAT selection battery is designed to provide the vehicle that will screen all candidates into the new Multi-Path Training Model. One of the important characteristics of the new training process is that it will no longer have a screening goal. The program will assume that candidates have the basic skills needed to perform the work of the ATCS. To implement the new training model, a selection process that screens candidates for the critical skills needed to perform the job is required. A Multi-path hiring model implemented with AT-SAT and augmented by a revised training program will likely reduce ATCS training length and time to certification.

Given this background, i.e., the demographics related to potential retirements, and new staffing requirements associated with training, there was a need to start the ATCS recruiting, selection, and training process in fiscal year 1997-1998. In spite of this immediate need to hire recruits, there were no currently feasible selection processes available to the FAA for use in the identification and selection of ATCSs. Test batteries that had been used in the past had become compromised, obsolete, or were removed from use for other reasons.

A two-stage selection process consisting of an OPM test battery and a nine-week Academy screen was introduced during the 1980s to select candidates for the position of air traffic controller. This two-stage process was both expensive and inefficient. First, candidates took a paper-and-pencil test administered by the Office

of Personnel Management (OPM). A rank-ordered list of candidates based on the OPM test scores was established. Candidates were listed according to their OPM test score plus any veteran's points. Candidates at the top of the list were hired, provided they could pass medical and security screening.

Once candidates were hired, they entered a nine-week screening program at the FAA Academy. Although modified several times during the 1980s, the basic program consisted of time spent in a classroom environment followed by work in laboratory-based, non-radar simulations. The classroom phase instructed employees on aircraft characteristics, principles of flight, the National Airspace System, and basic rules for separating aircraft in a non-radar situation. During the ATCS simulations phase, employees were taught and evaluated in an environment that emulated the work performed in an ATCS facility.

The OPM test had been in use, without revision, since 1981. In addition, test taking strategies and coaching programs offered by private companies increased the test scores of candidates without an apparent comparable increase in the abilities required to perform in the screen. The artificial increase in test scores apparently reduced the capability of the test to accurately identify the highest qualified individuals to hire. Due at least in part to the artificially inflated OPM scores, less than 40% of the ATCS trainees successfully completed the nine-week screen. A full discussion of prior selection procedures for ATCSs is provided in Chapter 6 on Archival Data Analyses.

Research and development efforts were begun to create a new selection device. One such research effort was the Separation and Control Hiring Assessment (SACHA) project initiated in 1991. SACHA focused on performing a job analysis of the air traffic controller position, developing ways to measure ATCS job performance, and identifying new tests suitable for selecting controllers. The SACHA contract expired in 1996.

Another research and development effort, the Pre-Training Screen (PTS), did produce a one-week selection test designed to replace the nine-week Academy screening process. However, the validity of the PTS was heavily weighted toward supervisor ratings and times to complete field training, along with performance in the Radar Training program. The FAA continued to use the PTS to screen candidates at a time when there was severe reduction in hiring, but there was no near-term potential to be hired. Meanwhile, the SACHA project was already underway and was partly envisioned as the "next stage" to the PTS.

In addition, the FAA had redesigned the initial qualification training program in anticipation that the PTS would filter candidates prior to their arrival at the Academy. The nine-week-long screening process was replaced with a program that focused on training, rather than screening candidates for ATCS aptitudes. As a result, the FAA had an initial training program but no pre-hire selection system other than the OPM written test.

The purpose of the AT-SAT project was to develop a job-related, legally defensible, computerized selection battery for ATCS's that was to be delivered to the FAA on October 1, 1997. The AT-SAT project was initiated in October of 1996. The requirement to complete the project within a year was dictated by the perceived need to start selecting ATCS candidates in 1997.

Organization of Report

A collaborative team, made up of several contractors and FAA employees, completed the AT-SAT project and this report. Team members included individuals from the Air Traffic Division of the FAA Academy and Civil Aeromedical Institute (CAMI) of the FAA, Caliber, Personnel Decisions Research Institutes (PDRI), RGI, and the Human Resources Research Organization (HumRRO). The Air Traffic Division represented the FAA management team, in addition to contributing to predictor and criterion development. CAMI contributed to the design and development of the job performance measures. Caliber was the prime contractor and was responsible for operational data collection activities and job analysis research. PDRI was responsible for research and development efforts associated with the job performance measures and development of the Experience Questionnaire (EQ). RGI was responsible for developmental activities associated with the Letter Factories Test and several other predictors. HumRRO had responsibility for project management, predictor development, data base development, validity data analysis, and the final report.

The final report consists of six chapters, with each chapter written in whole or part by the individuals responsible for performing the work. Contents of each chapter is summarized below:

Chapter 1 - Introduction: contains an overview of the project, including background and setting of the problem addressed, and methodology used to validate predictor measures.

Chapter 2 - Job Analysis: summarizes several job analyses that identified the tasks, knowledges, skills, and abilities required to perform the ATCS job. This chapter also contains a linkage analysis performed to determine the relationship between worker requirements identified in the job analysis to the predictor measures used in the validation study.

Chapter 3 - Predictor Development: focuses on how the initial computerized test battery was developed from job analysis and other information. This chapter also discusses construction of multi-aptitude tests and alternative predictors used to measure several unique worker requirements as well as the initial trial of the tests in a sample of students in a naval training school.

Chapter 4 - Criterion Development: discusses the development and construct validity of three criterion measures used to evaluate ATCS job performance.

Chapter 5 - Validation of Predictors: presents the predictor-criterion relationships, fairness analyses, and a review of the individual elements considered in deciding on a final test battery.

Chapter 6 - Analyses of Archival Data: discusses the results of analyses of historical data collected and maintained by CAMI and its relationship to AT-SAT variables.

Design of Validity Study

Step 1: Complete Predictor Battery Development

The tasks, knowledges, skills, and abilities (worker requirements) of the air traffic control occupation were identified through job analysis. Several prototype predictor tests were developed to cover the most important worker requirements of ATCSs. The management team held a predictor test review conference in November 1996 in McLean, Virginia. At the meeting, all prototype predictor tests were reviewed to determine which were appropriate and could be ready for formal evaluation in the validity study. Twelve individual predictor tests were selected. Step 1 of the management plan was to complete the development of the prototype tests and combine them into a battery that could be administered on a personal computer. This initial test battery was designated the Alpha battery.

It was also decided at this meeting to limit the validation effort to a sample of full performance level en route ATCSs to help ensure that the validity study would be completed within a year. Several considerations went into the decision to perform the validation study on en route controllers. Neither the development of a common criterion measure nor separate criterion measures for en route, tracon, and tower cab controllers was compatible with completing the validity study within one year. The solution was to limit the study to the single en route specialty. The SACHA job analysis concluded that there were not substantial differences in the rankings of important worker requirements between en route, tracon, and tower cab specialties. In addition, considerable agreement was found between subactivity ratings for the three specialties. Flight service, on the other hand, appeared to have a different pattern of important worker requirements and subactivity rankings than the other options. The en route option was viewed as reasonably representative of options that control air traffic, i.e., en route, tracon, and tower cab specialists. Further, the number of en route specialists was large enough to meet the sample size requirements of the validation study.

In addition, Step1 of the management plan included the requirement of a pilot test of the Alpha battery on a sample that was reasonably representative of the ATCS applicant population. Data from the pilot sample would be used to revise the Alpha test battery on the basis of the results of analyses of item, total score, and test intercorrelations. Beta, the revised test battery, was the battery administered to en route ATCSs in the concurrent validity sample and pseudo applicant samples. Test development activities associated with cognitive and non-cognitive predictors, the pilot sample description, results of the pilot sample analyses, and resultant recommendations for test modifications are presented in Chapter Three.

Step 2: Complete Criterion Measure Development

Three job performance measures were developed to evaluate en route job performance. By examining different aspects of job performance, it was felt that a more complete measure of the criterion variance would be obtained. The three measures included supervisor and peer ratings of typical performance, a computerized job sample, and a high-fidelity simulation of the ATCS job. Because the high-fidelity simulation provided the most realistic environment to evaluate controller performance, it was used to evaluate the construct validity of the other two criterion measures. The research and development effort associated with the criterion measures is presented in Chapter 4.

Step 3: Conduct Concurrent Validation Study

The job relatedness of the AT-SAT test battery was demonstrated by means of a criterion-related validity study. By employing the criterion-related validation model, we were able to demonstrate a high positive correlation between test scores on AT-SAT and the job performance of a large sample of en route ATCSs. Because of the amount of time required for ATCSs to reach full performance level status, i.e., two to four years, and the project requirement of completion within a year, a concurrent criterion-related design was employed in the AT-SAT study. In a concurrent validation strategy, the predictor and job performance measures are collected from current employees at approximately the same time.

The original goal for the number of total participants in the study was 750 en route ATCSs, including 100 representatives from each of the major protected classes. Over 900 pairs of predictor and criterion cases were collected in Phase I of the concurrent study. However, the goal of collecting 100 African American and Hispanic ATCS cases was not achieved. As a consequence, the FAA decided to continue the concurrent validity study to obtain a greater number of African American and Hispanic study participants. These additional data were required to improve statistical estimates of fairness of the AT-SAT battery. In Phase II, data were collected from en route sites that had not participated in Phase I. In addition, a second request for study participation was made to ATCSs in sites that had been a part of Phase I. All 20 en route sites participated in the AT-SAT study.

It should be noted that because of an ATCS employee union contract provision, each study participant had to volunteer to be included in the study. Consequently, the completion of the study was totally dependent on the good will of the ATCSs, and a significant amount of effort was expended in convincing them of the need and value of their participation in the AT-SAT project. A similar effort was directed at employee groups representing the protected classes. In the final analysis, however, each study participant was a volunteer. The FAA had no real control over the composition of the final ATCS sample. The data collection effort was anticipated to be highly intrusive to the operations of en route centers. There was substantial difficulty associated with scheduling and arranging for ATCS participation. The FAA's Air Traffic Training management team had the responsibility to coordinate acceptance and participation in the study of all stakeholders. The demographics of the obtained samples, corrected and uncorrected results of predictor and criterion analyses, and group difference and fairness analyses are discussed in Chapter 5.

Step 4: Conduct Pseudo-Applicant Study

Full-performance-level ATCSs are a highly selected group. As indicated earlier, even after the OPM selection battery was used to select candidates for ATCS training, there was still a 40% loss of trainees through the Academy screen and another 10% from on-the-job training. Under these conditions, it was highly likely that the range of test scores produced by current ATCSs would be restricted. Range restriction in predictor scores suggests that ATCSs would demonstrate a lower degree of variability and higher mean scores than an unselected sample. A restricted set of test scores, when correlated with job performance measures, is likely to underestimate the true validity of a selection battery. Therefore, to obtain validity estimates that more closely reflected the real benefits of a selection battery in an unselected applicant population, validity coefficients were corrected for range restriction. Range restriction corrections estimate what the validity estimates would be if they had been computed on an unselected, unrestricted applicant sample.

One method of obtaining the data required to perform the range restriction corrections is to obtain a sample of test scores from a group of individuals that is reasonably representative of the ATCS applicant pool. The best sources of data for this purpose are real applicants, but this information would not become available until the test battery was implemented. Both military and civilian pseudo-applicant samples were administered the AT-SAT battery for the purpose of estimating its unrestricted test variance and correcting initial validity estimates for range restriction. Pseudo applicant data were also used to obtain initial estimates of potential differences in test scores due to race and gender. The range restriction corrections resulted in moderate to large improvements in estimates of validity for the cognitive tests. These results are shown in Chapter 5. The final determination of these corrections will, by definition, require analyses of applicant data.

Step 5: Analyses and Validation of Predictors

Data management was a particularly critical issue on the AT-SAT project. Plans to receive, log in, and process data from 15 sites over an eight-week period were created. In addition, a final analysis database was devel-

oped so that the validity analyses of the predictors could be completed within a two-week time frame. Plans were also made on the methodology used to determine the validity of the predictors. These included predictor-criterion relationships and reviews of the individual elements to consider when deciding on the final test composite. There was a need to include special test composite analyses examining the interplay of differences between groups, the optimization of particular criterion variables, and coverage of worker requirements and their effect on validity. In particular, the final composition of the AT-SAT battery represented a combination of tests with the highest possible relation to job performance and smallest differences between protected classes. All validity and fairness analyses are presented in Chapter 5.

Step 6: Deliver Predictor Battery and Supporting Documentation

The final deliverable associated with the AT-SAT project was the AT-SAT test battery, version 1.0, on a compact disc (CD). The goal of developing a selection test battery for the ATCS that was highly job related and fair to women and minorities was achieved. Included with the CD are source code, documentation, and a user's manual. In addition, a database containing all raw data from the project was provided to the FAA.

CHAPTER 2

Air Traffic Controller Job Analysis

Ray A. Morath, Caliber Associates
Douglas Quartetti, HumRRO
Anthony Bayless, Claudet Archambault
Caliber Associates

PRIOR JOB ANALYSES

The foundation for the development of the AT-SAT predictor battery, as well as the job performance measures, was the Separation and Control Hiring Assessment (SACHA) job analysis (Nickels, Bobko, Blair, Sands, & Tartak, 1995). This traditional, task-based job analysis had the general goals of (a) supporting the development of predictor measures to be used in future selection instrumentation, (b) supporting the identification of performance dimensions for use in future validation efforts, and (c) identifying differences in the tasks and worker requirements (WRs; knowledges, skills, abilities, and other characteristics, or KSAOs) of the different ATCS options (Air Route Traffic Control Center, ARTCC; Terminal, and Flight Service) and ATCS job assignments (ARTCC, Terminal Radar Approach Control, TRACON; Tower Cab, and Automated Flight Service Station, AFSS).

Review of Existing ATCS Job Analysis Literature

Nickels et al. (1996) began by reviewing and consolidating the existing ATCS job analysis literature. This integration of the findings from previous job-analytic research served as the initial source of information on ATCS jobs prior to any on-site investigations and helped to focus the efforts of the project staff conducting the site visits. A core group of job analysis studies also provided much of the information that went into developing preliminary lists of tasks and WRs for the SACHA project. The following is a review of the major findings from selected studies that were most influential to SACHA and provided the greatest input to the preliminary task and WR lists.

Computer Technologies Associates (CTA)

CTA conducted a task analysis of the ARTCC, TRACON, and Tower Cab assignments with the goal not only of understanding how the jobs were currently performed but also of anticipating how these jobs would be performed in the future within the evolving Advanced Automation System (AAS).[1] They sought to identify the information processing tasks of ARTCC, TRACON, and Tower Cab controllers in order to help those designing the AAS to gain insight into controller behavioral processes (Ammerman et al., 1983).

An extensive assortment of documents was examined for terms suitable to the knowledge data base, including FAA, military, and civilian courses. Listed below are the sources of the documents examined for ATCS terms descriptive of knowledge topics and technical concepts:

- Civilian publications
- Community college aviation program materials
- Contractor equipment manuals
- FAA Advisory Circulars
- FAA air traffic control operations concepts
- FAA documents
- FAA orders
- Local facility handbooks
- Local facility orders
- Local facility training guides and programs
- NAS configuration management documents
- National Air Traffic Training Program (manuals, examinations, lesson plans, guides, reference materials, workbooks, etc.)
- Naval Air Technical Training Center air traffic controller training documents
- U.S. Air Force regulations and manuals

[1] Alexander, Alley, Ammerman, Fairhurst, Hostetler, Jones, & Rainey, 1989; Alexander, Alley, Ammerman, Hostetler, & Jones, 1988; Alexander, Ammerman, Fairhurst, Hostetler, & Jones, 1989; Alley, Ammerman, Fairhurst, Hostetler, & Jones, 1988; Ammerman, Bergen, Davies, Hostetler, Inman, & Jones, 1987; Ammerman, Fairhurst, Hostetler, & Jones, 1989.

Detailed task statements from each controller option were organized hierarchically into more global and interpretable subactivity and activity categories. Within CTA's framework, one or more tasks comprised a subactivity, with multiple subactivities subsumed under a single activity. Hence, this approach generated three levels of job performance descriptors. It was found that controllers in each of the three ATCS options (ARTCC, TRACON, and Tower Cab) performed about 6-7 of the more general activities, approximately 50 subactivities, and typically several hundred tasks.

The results of the CTA task analysis indicated that activity categories as well as subjectivity categories were similar across the assignments of ARTCC, TRACON, and Tower Cab, with only small variations in the tasks across the options. These findings suggested that the more global activities performed in each of these three controller jobs are almost identical. Additionally, job analysis results revealed 14 cognitively oriented worker requirements (WRs) that were found to be critical in the performance of tasks across the three assignments. These WRs were:

- Coding
- Decoding
- Deductive reasoning
- Filtering
- Image/pattern recognition
- Inductive reasoning
- Long-term memory
- Mathematical/probabilistic reasoning
- Movement detection
- Prioritizing
- Short-term memory
- Spatial scanning
- Verbal filtering
- Visualization

Human Technologies, Inc. (HTI)

HTI (1991) conducted a cognitive task analysis with ARTCC controllers to analyze mental models and decision-making strategies of expert controllers. An additional goal was to determine the effect of controller experience and expertise on differences in controller knowledges, skills, mental models, and decision strategies. Cognitive task analysis was performed by videotaping controllers during various traffic scenarios and having them describe in detail what they were thinking while they were handling the traffic scenarios.

The general findings of the cognitive task analysis were that ARTCC controllers' mental models for the control of air traffic could be broken down into three general categories, which were termed (a) sector management, (b) prerequisite information, and (c) conditions. These three categories roughly parallel the respective information processing requirements of short-term memory, long-term memory, and switching mechanisms. The resulting 12 cognitively oriented tasks were:

- Maintain situational awareness
- Develop and revise sector control plan
- Resolve aircraft conflict
- Reroute aircraft
- Manage arrivals
- Manage departures
- Receive handoff
- Receive pointout
- Initiate handoff
- Initiate pointout
- Issue advisory
- Issue safety alert

Another important finding from the study was that due to their more effective working memory, experts have access to more information than novices. That is, experts have a greater chunking capacity. Experts are also more efficient in the control of aircraft because they typically use a smaller number of strategies per traffic scenario, and have a greater number of strategies that they can employ. Finally, the study found that expert ARTCC controllers differed from novices in their performance on the two most important cognitive tasks: maintaining situational awareness and revising the sector control plan.

HTI's work also involved investigating the communications within teams of radar AT-SAT (radar and associate radar) as well as the communications between radar associates and controllers in other sectors. Teams were studied in both live and simulated traffic situations. Communication data were coded in relation to 12 major controller tasks that were found in the cognitive task analysis. The data indicated that nearly all communication between team members concerned the two tasks deemed most critical by the cognitive task analysis: maintain situational awareness, and develop and revise sector control plan.

A summary job analysis document (HTI, 1993) presents all the linkages from seven sets of prototype documents representative of air traffic controller job

analyses. The objective of this summary process was to systematically combine the results from the air traffic controller job analyses into a single document that emphasizes the task-to-KSAO linkages for the jobs of En Route, Flight Service Station, combined TRACON and Tower (Terminal), Tower, and TRACON controllers.

The results reported in the HTI summary job analysis were based on individual job analysis summaries, which included a cognitive task analysis and the use of the Position Analysis Questionnaire (Meecham & McCormick, 1969). The HTI analysis also utilized the CTA task analysis that had standardized task and KSAO data from existing air traffic control job analyses. In the individual job analyses, the controller tasks and KSAO data were translated into Standard Controller Taxonomies. Then, the linkages for each controller job were identified and placed in standard matrices based on these taxonomies.

The Task Taxonomy includes a total of 41 task categories grouped as follows:

- Perceptual Tasks
- Discrete Motor Tasks
- Continuous Psychomotor Tasks
- Cognitive Tasks
- Communication Tasks

The KSAO Taxonomy has a total of 48 KSAO categories divided into the following three groupings:

- Abilities
- Knowledge
- Personality Factors

This resulted in a 41-by-48 task-to-KSAO matrix that permits the standard listing of task-to-KSAO linkages from different job analyses.

The summary document (HTI, 1993), which incorporated the individual job analyses as well as the CTA report, was reviewed and utilized by the AT-SAT researchers in determining the contribution of these reports to the understanding of the air traffic controller job.

Embry-Riddle

Using a hierarchical arrangement of activities and tasks borrowed from CTA, Embry-Riddle researchers (Gibb et al., 1991) found that five activities and 119 tasks subsumed under those more global activities were identified as critical to controller performance in the non-radar training screen utilized with ARTCC and Terminal option controllers. The five global activities identified by Embry-Riddle investigators were:

- Setting up the problem
- Problem identification
- Problem analysis
- Resolve aircraft conflicts
- Manage air traffic sequences

Upon the basis of the results of the task inventory, existing documentation, and the information obtained from meetings with training instructors, the following 18 attributes were identified as critical in performing the activities and tasks involved in the training:

- Spatial visualization
- Mathematical reasoning
- Prioritization
- Selective attention
- Mental rotation
- Multi-task performance (time sharing)
- Abstract reasoning
- Elapsed time estimation and awareness
- Working memory - attention capacity
- Working memory - activation capacity
- Spatial orientation
- Decision making versus inflexibility
- Time sharing - logical sequencing
- Vigilance
- Visual spatial scanning
- Time-distance extrapolation
- Transformation
- Perceptual speed

In addition to identifying the tasks and abilities required for success in training, another goal of this project was to determine the abilities necessary for success on the ATCS job. The Embry-Riddle team employed Fleishman's ability requirements approach for this purpose. Utilizing the task-based results of the CTA job analysis (Ammerman et al., 1987), they had ARTCC, TRACON, and Tower Cab controllers rate CTA tasks on the levels of abilities needed to successfully perform those CTA-generated tasks. Using Fleishman's abilities requirements taxonomy (Fleishman & Quaintance, 1984), these subject matter experts (SMEs) rated the levels of perceptual-motor and cognitive abilities required for each of the tasks. It was found that the abilities rated by controllers as critical for

controller performance were highly similar to those found by the CTA study; they were also quite similar to those abilities identified by the Embry-Riddle team as important to success in the non-radar training screen. ARTCC controllers rated the following abilities from Fleishman's scales as necessary to perform the CTA-generated ARTCC tasks:

- Deductive reasoning
- Inductive reasoning
- Long-term memory
- Visualization
- Speed of closure
- Time sharing
- Flexibility of closure (selective attention)
- Category flexibility
- Number facility
- Information ordering

Those abilities rated by Terminal controllers as required to perform the Terminal option tasks were:

- Selective attention
- Time sharing
- Problem sensitivity
- All of Fleishman's physical abilities related to visual, auditory, and speech qualities
  - Oral expression
  -Deductive reasoning
  - Inductive reasoning
  - Visualization
  - Spatial orientation
  - All perceptual speed abilities.

The Embry-Riddle researchers presented no discussion on why differences in abilities between ARTCC and Terminal controllers were found.

Landon

Landon (1991) did not interview SMEs, observe controllers, or canvass selected groups to collect job analysis information. Rather, Landon reviewed existing documents and job analysis reports and summarized this information. Landon's focus was to identify and classify the types of tasks performed by controllers. Using CTA's hierarchical categorization of tasks, the ATCS tasks were organized into three categories based upon the type of action verb within each task:

I. Information Input Tasks
Receive, interpret, compare and filter information
Identify information needing storage or further processing
A. Scanning and monitoring
B. Searching
II. Processing Tasks

Organize, represent, process, store, and access information
A. Analytical planning
B. Maintain picture in active memory
C. Long-term memory
D. System regulation
III. Action/Output Tasks

Physical and verbal actions to communicate and record information
A. Communicate outgoing messages
B. Update flight records
C. Operate controls, devices, keys, switches

Myers and Manning

Myers and Manning (1988) performed a task analysis of the Automated Flight Service job for the purpose of developing a selection instrument for use with AFSS. Using the CTA hierarchy to organize the tasks of the job, Myers and Manning employed SME interviews and surveys to identify the activities, subactivities, and tasks of the job. They found 147 tasks, 21 subactivities, and the five activities that they felt comprised the jobs of the AFSS option. The activities that they identified were:

- Process flight plans
- Conduct pilot briefing
- Conduct emergency communications
- Process data communications
- Manage position resources

Using the subactivities as their focus, Myers and Manning identified those WRs required to successfully perform each subactivity. Unlike the CTA and Embry-Riddle job analyses, the WRs identified were much more specific in nature, as evidenced by the following examples:

- Ability to operate radio/receive phone calls
- Ability to use proper phraseology
- Ability to keep pilots calm
- Ability to operate Model 1 equipment

Summary of Previous Studies

SACHA project staff summarized the findings from the previous job analyses and identified the commonalities across those reports regarding the tasks and worker requirements. Specifically, they compared CTA's worker requirements with those reported by Embry-Riddle. Additionally, once the SACHA-generated lists were completed, the researchers mapped those worker requirements to those reported by CTA. In general, the global categories of tasks and the hierarchical organization of tasks for the ARTCC and Terminal options were common across the previous studies. Additionally, the sub-activities and worker requirements identified in previous research for those two ATCS options were similar. Finally, the previous job analyses illustrated differences in tasks and worker requirements between the AFSS option and the other two options.

SACHA Site Visits

After reviewing and summarizing the existing job analysis information, the SACHA project staff visited sites to observe controllers from the various options and assignments. More than a dozen facilities, ranging from ARTCC, Level II to V Tower Cab and TRACON, and AFSS facilities, were visited. The primary purpose of these initial site visits was to gain a better understanding of the ATCS job. SACHA project staff not only observed the controllers from the various options performing their job, but they also discussed the various components of the job with the controllers, their trainers, and supervisors.

Development of Task and WR Lists

Developing Preliminary Lists

On the basis of the results of the previous job analyses as well as the information obtained from the site visits, SACHA's project staff developed preliminary task and WR lists. Given the strengths of the CTA job analysis regarding (a) its level of specificity, (b) its hierarchical arrangement of tasks, and (c) its focus on both the current ATCS job and how the job is likely to be performed in the future, SACHA decided to use the task analysis results of CTA as the basis for preliminary task lists with the options of ARTCC, TRACON, and Tower Cab. Similarly, Myers and Manning performed a relatively extensive job analysis of the AFSS position, which had been modeled after CTA; they, too, had used a hierarchical categorization of tasks and had a high degree of specificity at the molecular, task level. Hence, four preliminary task lists were developed for the ATCS job assignments of ARTCC, TRACON, Tower Cab, and AFSS, with task-based findings from CTA and the Myers and Manning results serving as the primary source of task-based information for the respective options. Each of these lists contained from six to eight global activities, 36 to 54 subactivities, and hundreds of tasks.

Several steps were followed in the development of the list of ATCS worker requirements. On the basis of their review of existing literature and their knowledge of those relevant KSAO constructs, SACHA project staff developed an initial list of 228 WRs. They then conducted a three-day workshop dedicated to refining this initial list. Consensus judgments were used to eliminate WRs that were thought to be irrelevant or redundant. Finally, a single preliminary WR list was formulated that contained 73 WRs grouped into 14 categories (reasoning, computational ability, communication, attention, memory, metacognitive, information processing, perceptual abilities, spatial abilities, interpersonal, work and effort, stability/adjustment, self-efficacy, and psychomotor). This list of WRs is presented in Table 2.1 and will henceforth be termed the SACHA-generated WRs list.

Panel Review of Preliminary Lists

A panel of five controllers assigned to FAA Headquarters participated in a workshop to review and revise the SACHA materials and procedures for use in additional job analysis site visits. These five controllers, who represented each of the options of ARTCC, Terminal, and Flight Service, began the workshop by undergoing a "dry run" of the planned field procedures in order to critique the procedures and offer suggestions as to how they might be improved. Second, the panel reviewed and edited the existing task lists for the various options, mainly to consolidate redundant task statements and clarify vague statements. Finally, the panel reviewed the SACHA-generated WRs list. Upon discussion, the panel made no substantive changes to either the task lists for the various options or the SACHA-Generated WRs list.

Developing and Revising Task Lists in the Field

In field job analysis meetings held in different regions, the preliminary task lists were presented to sevennine SMEs (controllers with varying levels of job experience) from each of the four options (ARTCC, TRACON, Tower Cab and AFSS). Special attention was placed upon having groups of SMEs who were diverse in race/ethnicity, gender, and years of experience

controlling traffic, and who represented various levels of ATC facilities. Attempts were made to avoid mixing subordinates and their supervisor(s) in the same meeting.

Project staff instructed SMEs to review their respective task list (whether it be ARTCC, TRACON, Tower Cab, or AFSS) to confirm which tasks were part of their job and which were irrelevant. In addition, SMEs were asked to consolidate redundant tasks, to add important tasks, and to edit any task statements that needed rewording for clarification or correction of terminology. SMEs proceeded systematically, reviewing an entire group of tasks under a subactivity, discussing the necessary changes, and coming to a consensus regarding those changes before moving to the next subactivity. After editing of the tasks was completed, SMEs were asked to identify those tasks that they believed were performed by all ATCS options, as well as those tasks specific to one or more ATCS position(s).

These meetings produced four distinct task lists corresponding to ARTCC, TRACON, Tower Cab, and AFSS controllers.

Developing SME-Generated WRs Lists

SME meetings were also held for the purpose of having controllers generate their own lists of WRs that they felt were necessary for effective job performance. SMEs were not given the preliminary WR list that had been generated by SACHA job analysts but were instructed to generate their own list of skill and ability-related WRs. SMEs were utilized to identify and define WRs while project staff assisted by clarifying the differences and similarities among the WR definitions. That is, project staff tried to facilitate the development of the WRs without influencing the SMEs' judgments.

As a result of this effort, SME controllers across the three options of ARTCC, Terminal, and Flight Service generated 47 WRs. Based upon the input from the SMEs, definitions were written for each of the WRs. It was concluded that all 47 SME-generated WRs would be applicable to any position within the three ATCS options. Table 2.2 represents the list of 47 SME-generated WRs.

When the SACHA-generated and SME-generated WR lists are compared, they appear to be quite similar except for the lack of metacognitive and information processing WRs in the SME-generated list. SACHA staff reported that controllers generating the SME list of WRs lacked familiarity with metacognitive and information processing constructs and were probably not the best sources of information when it came to identifying these types of WRs. Because (as SACHA researchers stated) no "common language" existed with which to discuss these types of WRs, project staff did not try to pursue defining these types of WRs with the controller SMEs.

Linking Tasks to SME-Generated WRs

SME meetings were then held to have controllers provide linkage judgments (obtained via group discussion) relating the tasks subsumed under a particular subactivity to the SME-generated WRs required to perform that subactivity. SMEs from the ARTCC and Terminal options reviewed the task and SME-generated WR lists from their respective options and identified those WRs needed to perform each subactivity. SMEs focused upon one subactivity at a time and obtained consensus regarding the most important WRs for that subactivity before moving on to the next. Linkages were made at the subactivity level because the large number of tasks precluded linkages being made at the task level. Due to scheduling problems, the SACHA project staff were unable to hold a linkage meeting with AFSS SMEs, so no data were obtained at this stage linking AFSS tasks to AFSS WRs.

While two SME-generated WRs (Motivation and Commitment to the Job) were not linked to any of the subactivities, controllers stated that these two WRs were related to the overall job. Thus, even though these WRs could not be directly linked to the tasks of any specific subactivity, controllers felt that their importance to overall job performance justified the linkage of these two requirements to every subactivity. Results of the linkage meetings revealed that every SME-generated WR could be linked to at least one subactivity and that each subactivity was linked to at least one WR.

Developing Consolidated List of WRs

At this stage, SACHA job analysts generated a consolidated WR list combining the SME (controller) and SACHA-generated WRs. They began the consolidation process by including 45 of the 47 original SME-generated WRs. Two of the original 47 were dropped (Aviation Science Background and Geography) because they were job knowledges rather than skills or abilities. Next, the SACHA-generated list of WRs was reviewed to add WRs that had not been identified in the SME-generated list but were considered important to the job of ATCS. Finally, the project staff added two WRs (Recall from Interruption and Translation of Uncertainty into Prob-

ability) that had not been identified in either the SACHA or the controller lists but were deemed necessary to perform the job from job analytic suggestions.

Thus, a final consolidated list of 66 WRs was created (45 SME-generated and 21 SACHA-generated) that included skills and abilities in the areas of communication, computation, memory, metacognition, reasoning, information processing, attention, perceptual/spatial, interpersonal, self-efficacy, work and effort, and stability/adjustment (Table 2.3).

Job Analysis Survey

Utilizing the information gained from the site visits and SME meetings, the SACHA staff developed a job analysis survey and disseminated it to a cross-section of ATCSs from the various options and assignments located throughout the country. The main goals of the mail-out survey were to identify the most important WRs for predictor development, to explore criterion measures, and to identify possible differences in the subactivities being performed across job assignments. Of the 1009 surveys sent out to ATCSs in February 1994, 444 were returned, with usable data obtained from 389 respondents.

Content of the Survey

The survey was divided into four sections: an introduction, a subactivity ratings section, a WRs rating section, and a background information section. The introduction explained the purpose of the survey to the ATCSs, provided instructions on completing the survey, and encouraged participation. The background section gathered information on such things as the respondents' gender, race, job experience, facility type and facility level.

The subactivity rating section was comprised of 108 entries, the combined list of all sub-activities across the ATCS options of ARTCC, Terminal, and Flight Service. The instructions informed respondents that the survey contained subactivities from ARTCC, TRACON, Tower Cab, and AFSS jobs, and thus it would be unlikely that all entries would be relevant for a particular job assignment. Respondents were asked to rate each subactivity on (a) its importance in successfully performing their job, and (b) the time spent performing this subactivity relative to the other job duties they perform. A single task criticality index was also created by combining the importance and relative time spent ratings. This index provided an indication of the relative criticality of each subactivity with respect to job performance.

The WR rating section of the survey was comprised of 67 items, which included (a) the 45 controller-generated items, (b) the two SME-generated job knowledges that had been reintroduced into the survey, (c) the 16 SACHA-generated items (five of the 21 SACHA-generated items dealing with information processing were left off the survey due to controllers' lack of understanding and familiarity with these constructs), and four items to identify random responses. Respondents were instructed to rate each of the 67 WRs on both its relative importance in learning the job and its relative importance in doing the job.

Overview of Survey Findings

WRs. Findings revealed very little difference between the WRs seen as important for doing the job and those needed to learn the job. Rank orderings of the WR mean scores for doing and learning the job were highly similar. This result appeared to hold across job options and job assignments. Mean rankings of the WRs for all ATCS job assignments are shown in Table 2.4. These scores reflect the mean rankings of the WRs for learning and for doing the job.

The results also suggested that, while there seemed to be no substantial difference between the WR ratings of the ARTCC and the Terminal option controllers (TRACON and Tower Cab), the Flight Service controllers appeared to rate the WRs differently. They rated WRs dealing with information collection and dissemination as relatively more important than did the ARTCC and Terminal option controllers, and rated WRs dealing with metacognitive functions as relatively less important.

As a result of the findings, SACHA staff felt that there were no substantive differences between the ARTCC and the Terminal options in the ordering of the WRs, which would influence predictor development for these two options. However, they advised that any future work dealing with test development for the Flight Service option should take into consideration their different rank ordering of WRs. Tables 2.5 and 2.6 list the mean ratings of the WRs (from each of the four job assignments) for doing and learning the job.

Subactivities. As with the ratings of the WRs, the results of the subactivity ratings revealed that ARTCC and Terminal option controllers shared similar profiles with respect to the relative criticality of the subactivities. While the ARTCC and Terminal option controllers share more common subactivities than they do each share with the Flight Service option, 11 subactivities were given relatively high ratings by all three options. These common sub-activities were associated with the safe and expeditious flow of traffic, as well as responding to emergencies or special conditions and contingencies. Table 2.7 contains the ranked mean rating of the subactivities across all ATCS options.

The SACHA staff also felt that another important finding from the controller ratings of the subactivities was that, regardless of job option or assignment, those subactivities dealing with multitasking were consistently seen as important to the ATCS job. The project staff operationalized multitasking as those times when controllers must (a) perform two or more job tasks simultaneously, (b) continue job tasks despite frequent interruptions, and (c) use multiple sensory modalities to collect information simultaneously or near simultaneously. When dealing with the ratings of the subactivities across all ATCS options, it was found that ten of the 11 sub-activities dealing with multitasking had criticality scores that placed them in the top third of all subactivities.

Conclusions

Considering the results of the SACHA job analysis survey and taking into account the goals of this selection-oriented job analysis, the project staff arrived at several general conclusions.

· There appeared to be no substantial differences in the rankings of the important WRs between ARTCC, TRACON, and Tower Cab controllers. However, the differences in the rankings found between Flight Service option controllers and the other options did appear to be substantive enough that any future efforts to develop selection instrumentation should take these differences into account.

· Considerable agreement was found between the subactivity rankings for the ARTCC, TRACON, and Tower Cab controllers, while the rank ordering of the subactivities for the Flight Service option appears to be different from all other options and job assignments.

· Regardless of job option or assignment, multitasking is an important component of the ATCS job.

Linkage of Predictors to Work Requirements Overview

To determine whether the various instruments comprising the AT-SAT predictor battery were indeed measuring the most important WRs found in the SACHA job analysis, linkage judgments were made by individuals familiar with the particular AT-SAT measures, as well as the WRs coming out of SACHA. Linkage analysis data were collected through surveys. Surveys contained (a) a brief introduction describing why the linkage of tests to WRs was necessary, (b) a brief background questionnaire, (c) instructions for completing the linkage survey, (d) definitions of the WRs from SACHA's revised consolidated list, and (e) linkage rating scales for each of the AT-SAT measures. Survey results showed that each of the measures comprising the AT-SAT battery was successfully capturing at least one or more WRs from SACHA's revised consolidated list. Additionally, the vast majority of those WRs being captured by AT-SAT measures were those SACHA found to be most important for both learning and doing the job.

The linkage analysis made use of 65 of the 66 WRs from SACHA's final revised consolidated list. Due to an oversight, two of the WRs from SACHA's list were labeled Rule Application (one SME-generated and the other SACHA-generated), and both were listed under the Information Processing category. When the SACHA list of WRs and their respective definitions were being transcribed for use in the linkage analysis, only one of the two Rule Application WRs was transcribed. Hence, the linkage analysis collected linkage ratings only on the SACHA-generated version of Rule Application, defined as the ability to efficiently apply transformational rules inferred from the complete portions of the stimulus array to the incomplete portion of the array. The SME-generated version of Rule Application, which was defined as the ability to apply learned rules to the real world work situation, was not included in the linkage survey.

Respondent Background Questionnaire

Project staff created the 7-item background questionnaire to be completed by the survey respondents. Items measured the respondent's highest educational degree, area of study, experience in data collection, experience in test construction, familiarity with AT-SAT, role in developing AT-SAT predictors and/or criterion measures, and familiarity with ATCS job. One

purpose of the items was to serve as a check in making sure the individuals were qualified raters. Additionally, these items could serve to identify subgroups of raters based upon such things as testing experience, educational background, and/or educational degree. In the event that rater reliability was low, attempts could be made to determine whether the lack of rater agreement was due to any one of these subgrouping variables.

Descriptions of AT-SAT Measures

The AT-SAT test battery used in the concurrent validity study contained the following 12 predictor measures:

- Dials
- Sound
- Letter Factory
- Applied Math
- Scanning
- Angles
- Analogies
- Memory
- Air Traffic Scenarios
- Experience Questionnaire
- Time Wall/Pattern Recognition
- Planes

An important element of the linkage survey consisted of operational descriptions of each of the AT-SAT predictor tests. These descriptions were meant not to replace the respondent's familiarity and experience with each of the measures but to highlight the most basic features of each measure for each respondent. While one of the criteria for inclusion as a survey respondent was a familiarity with each of the measures (typically gained through helping to create and/or taking the test), it was felt that a general description of the features of each measure would facilitate a more complete recall of the test characteristics. Respondents were instructed to read each test description before rating the degree to which that test measures each of the WRs. Figure 2.1 is an example of a description for one of the AT-SAT measures and its accompanying rating scale.

The only predictor measure for which no operational description was provided was the Experience Questionnaire (EQ). This measure was a biodata inventory comprised of 14 subscales, with individual subscales containing anywhere from 9 to 15 items. In the place of test descriptions, respondents making linkage ratings on the EQ received the actual items from the individual scales (but did not receive the construct labels for these scales). Respondents were to use the items comprising each scale to determine the construct being measured by that particular scale and then make their ratings as to the degree to which the scale successfully measured each WR.

Definitions of WRs

The survey contained an attachment listing the WRs and their accompanying definitions from SACHA's revised consolidated WR list (except for the SME-generated WR of Rule Application). It was felt that, in order for respondents to make the most informed linkage rating between a test and a WR, they should not only have a clear understanding of the properties of the test, but also possess a firm grasp of the WR. Survey respondents were instructed to read through the attachment of WRs and their respective definitions before making any linkage ratings and to refer back to these definitions throughout the rating process (Table 2.8).

Survey Respondents

To qualify as raters, individuals had to be familiar with the measures comprising the AT-SAT battery, and they had to have an understanding of each of the WRs being linked to the various measures. Potential respondents were contacted by phone or E-mail and informed of the nature of the rating task. A pool of 25 potential respondents was identified. The individuals in this pool came primarily from the organizations contracted to perform the AT-SAT validation effort but also included FAA personnel directly involved with AT-SAT.

Survey Methodology

Those who had agreed to participate in the linkage process received the packet of rating materials via regular mail. Each packet contained the following items:

(1) An introduction, which outlined the importance of linking the AT-SAT predictor tests to the WRs identified in the SACHA job analysis. It included the names and phone numbers of project staff who could be contacted if respondents had questions concerning the rating process.
(2) The 7-item background questionnaire.
(3) The attachment containing the list of WRs and their definitions.

(4) Rating scales for each of the AT-SAT tests. Each rating scale contained the operational description of the measure (or for the EQ, those items comprising an EQ sub-scale), the Likert scale response options, and the WRs to be rated (Figure 2.2.).

In view of the inordinate amount of time it would take a respondent to rate all the WRs on each of the 12 tests, tests were divided in half, with one group of respondents rating six tests and the other group rating the other six tests. The 25 potential raters who had been identified were split into two groups. Thirteen respondents were responsible for linkage ratings for the Angles, Analogies, Memory, AT Scenarios, Planes, and Experiences Questionnaire; the remaining 12 were to make linkage ratings for the Dials, Sound, Letter Factory, Applied Math, Scanning, and Time Wall. The subset of tests sent to the 13 respondents was labeled Version 1, and the second subset of tests sent to the remaining 12 respondents was labeled Version 2.

Results of the Survey

The surveys were returned, and the data were analyzed by project staff. Twenty-four respondents completed the background questionnaire, as well as all or portions of the six tests they were to link with the respective WRs. Nineteen of the 24 respondents classified themselves as Industrial/Organizational Psychologists; all but one of the 24 had obtained at least a master's degree. In general, results of the questionnaire indicated that the raters were experienced in test construction and administration and were familiar with the AT-SAT test battery.

All 12 respondents who were asked to rate Version 1 of the linkage survey completed and returned their ratings. Two of these individuals volunteered to complete the linkage ratings of the Version 2 tests and followed through by completing these ratings as well. Completed ratings were returned by 11 of the 13 respondents who were tasked with making linkage ratings for Version 2 tests, with one rater choosing not to rate either the Memory or the Planes tests. Hence, 12 complete sets of linkage ratings were obtained for the Version 1 tests, and 14 complete sets of linkage ratings were obtained for all but two of the Version 2 tests (Table 2.9).

Scale Reliability

Reliability indices were computed for each rating scale. Scale reliabilities ranged from .86 to .96. Hence, the intraclass correlations (Shrout & Fleiss, 1979) for each of the rating scales revealed a high level of agreement between the respondents as to which WRs were being successfully measured by the respective tests. These reliability coefficients are listed in Table 2.9. In view of the high level of agreement, it appeared that such factors as the rater's highest educational degree, educational background, and familiarity with the ATCS job did not influence the level of agreement among the raters.

Angles

The Angles test measures the participant's ability to recognize angles. This test contains 30 multiple-choice questions and allows participants up to 8 minutes to complete them. The score is based on the number of correct answers (with no penalty for wrong or unanswered questions). There are two types of questions on the test. The first presents a picture of an angle and the participant chooses the correct answer of the angle (in degrees) from among four response options. The second presents a measure in degrees and the participant chooses the angle (among four response options) that represents that measure. For each worker requirement listed below, enter the rating best describing the extent to which this test and/or its subtests measure that particular worker requirement.

5= This test measures this worker requirement to a **very great extent**
4= This test measures this worker requirement to a **considerable extent**
3= This test measures this worker requirement to a **moderate extent**
2= This test measures this worker requirement to a **limited extent**
1= This test measures this worker requirement to a **slight extent**
0= This test **does not measure** this worker requirement

Linkage Results

Mean linkage scores between tests and WRs were computed, representing a summary of the extent to which the raters felt each test measured each WR. It was

decided that linkage means greater than or equal to 3 suggested that the raters felt a test was able to measure the WR to at least a moderate extent. Therefore, a criterion cutoff mean ≥ 3 was established to determine if a test was indeed able to successfully measure a particular WR. Table 2.10 presents the following information: (a) a full list of the WRs (rank ordered by their importance for doing the job), (b) those tests rated as measuring the WR to at least a moderate extent (based upon the mean ≥ 3 cutoff), and (c) the mean linkage rating corresponding to that test/WR pair.

The rank-ordered listing of WRs used in Table 2.10 was derived from the SACHA job analysis and consisted of the ARTCC controllers' ratings of the extent to which the WR was seen as being important for doing the job. Hence, on the basis of ARTCC controllers' rating from SACHA, prioritization was seen as the most important WR for doing the job. AT-SAT project staff chose to use the ARTCC rank-ordered listing of WRs on their importance for doing the job for three reasons. First, the ARTCC list was chosen over the list generated by the ratings from all ATCSs because the latter included the ratings of controllers from the Flight Service option. SACHA had clearly stated that the job analysis findings for the Flight Service option were different enough from the ARTCC and Terminal options to require its own predictor development. Second, the ARTCC list was selected over the Terminal option list because the AT-SAT validation effort was using controllers from the ARTCC option. Finally, the ARTCC list for doing the job was chosen over the list for learning the job because the focus of AT-SAT was on developing a selection battery that would predict performance in the job—not necessarily in training.

It should be mentioned that no data on importance for learning or doing the job existed for five of the SACHA-generated information processing WRs (Confirmation, Encoding, Rule Inference, Rule Application, and Learning). This was because the SACHA project staff believed that controllers could not adequately comprehend these WRs well enough to rate them. Because of this lack of importance data, project staff placed these WRs at the end of the list of WRs (see Table 2.10.).

Based upon these criteria for inclusion, it was found that 14 of the 15 most important WRs (as rated by ARTCC controllers in the SACHA job analysis) were successfully measured by one or more of the tests of the AT-SAT battery. Similarly, the mean linkage ratings suggest that the vast majority of the more important WRs were successfully measured by multiple tests.

The linkage survey results indicated that all important WRs were not successfully measured by the AT-SAT battery. Four WRs (Oral Communication, Problem Solving, Long-Term Memory, and Visualization) from the top third of SACHA's rank-ordered list did not have linkage means high enough to suggest that they were being measured to at least a moderate extent. None of the AT-SAT tests were specifically designed to measure oral communication and, as a result, linkage means between this WR and the tests were found to be at or near zero. Problem Solving had mean linkage ratings that approached our criterion for inclusion for the Applied Math and the Letter Factory tests. Similarly, the mean linkage ratings between the Memory test and Long-Term Memory, and between the Letter Factory test and Visualization also approached but failed to meet the mean criterion score of 3.

Quality of Individual Tests in the AT-SAT Battery

Results of the linkage survey were also summarized to enable project staff to gain insight into how well individual tests were measuring the most important WRs. Based upon the criterion of mean linkage score ≥ 3 for demonstrating that a test successfully measures a particular WR, project staff determined the number of WRs successfully measured by each test. This score provided some indication of the utility of each test. Project staff also computed two additional scores to indicate the utility of each measure. Some WRs were rated as being successfully measured by many tests, and other WRs were measured by only one or two tests. Two other indicators of the utility of a measure were developed: (a) the number of WRs a test measured that are only measured by *one (or fewer) other test(s)*, and (b) the number of WRs that are not measured by *any other test*. Scores based upon these criteria were computed for each measure and are listed in Table 2.11.

In addition to the indicators of each test's utility, it was felt that indicators of each test's utility and quality in measuring WRs could also be computed. To provide some indication of each test's quality, project staff again utilized SACHA findings— the ARTCC controller ratings of the importance of each WR for doing the job. Each WR's mean importance rating (from SACHA) was multiplied by those WR/test linkage ratings meeting criteria. The product of these two scores (mean WR importance for doing the job x mean linkage rating of WR for a test) factored in not only how well the test was capturing the WR but the importance of that WR as well. The mean and sum of these products were com-

puted for each (Table 2.11.). The mean of the products can be viewed as an indicator of the average quality of a measure factoring in both how well the test was measuring the WR and the importance of the WR. The sum of the products provides some indication of the overall utility of the measure in that the more WRs a test captures, the better it captures those WRs, and the more important these WRs are for the doing the job, the higher a test's total score on this factor.

Given that no data were collected in SACHA for five of the WRs (Confirmation, Encoding, Rule Inference, Rule Application, and Learning) on their importance for doing the job, the mean importance score across all the WRs was imputed for these five WRs. This was done so that some indication of a test's ability to measure these WRs could be computed and factored into its overall quality and utility scores (Table 2.11.).

Results suggest that some tests - Letter Factory, AT Scenarios, and to a lesser degree the Time Wall and Analogies tests - measured numerous WRs, while the remaining tests measured from one to three WRs. Some tests, such as Applied Math and Analogies, measured multiple WRs that were not measured by other tests, while other tests (Letter Factory, Air Traffic Scenarios, and Time Wall) measured many WRs but none uniquely. It should be mentioned that one of the reasons the Letter Factory, Air Traffic Scenarios, and Time Wall did not uniquely capture any WRs was that there was so much overlap in the WRs successfully measured by these three tests—especially between the Letter Factory and the Air Traffic Scenarios.

Conclusion

Based upon the results of the linkage survey, every test within the AT-SAT battery appeared to be successfully measuring at least one WR, and many of the tests were rated as measuring multiple WRs. While not every WR was thought to be successfully measured by the AT-SAT battery, the vast majority of the WRs considered most important for doing the job was successfully measured by one or more predictors from the battery.

CHAPTER 3.1

Predictor Development Background

Douglas Quartetti, HumRRO
William Kieckhaefer, RGI, Inc.
Janis Houston, PDRI, Inc

Following the air traffic controller strike and the subsequent firing of a significant portion of that workforce in 1981, the Federal Aviation Administration was forced to hire en masse to ensure safety of the airways. Cooper, Blair, and Schemmer (1994) reported on the selection procedures used after the strike. Their work is summarized below.

SELECTION PROCEDURES PRIOR TO AT-SAT

The OPM Battery

In October 1981, the FAA introduced a two-stage process for selecting Air Traffic Control Specialists (ATCSs). The first stage was a paper-and-pencil aptitude test battery administered by the Office of Personnel Management (OPM), known as the OPM Battery. This battery consisted of three tests: the Multiplex Controller Aptitude Test (MCAT), the Abstract Reasoning Test (ABSR), and the Occupational Knowledge Test (OKT). The second stage was called the Academy Screen.

The first test of the OPM Battery, the MCAT, simulated aspects of air traffic control. Applicants were required to solve time, distance, and speed problems, plus interpret tabular and graphical information to identify potential conflicts between aircraft. Extensive research at the Civil Aeromedical Institute (CAMI) indicated that the MCAT scores were significantly correlated with performance during the Academy Screen and later field status (Manning, Della Rocco, & Bryant, 1989; Rock, Dailey, Ozur, Boone, & Pickerel, 1978).

The second test of the OPM Battery, the ABSR, was developed by the U.S. Civil Service Commission to examine the abstract relationships between symbols and letters. Research indicated a relationship between scores on this test and the Academy Screen training performance (Boone, 1979; Rock et al., 1978).

The final test in the OPM Battery, the OKT, contained questions on air traffic phraseology and procedures. It was designed to provide credit for prior ATCS experience. It has been reported that OKT scores correlated with many of the indices of training success (Boone, 1979; Buckley, O'Connor, & Beebe, 1970; Manning et al., 1989; Mies, Coleman, & Domenech, 1977).

The scores on the MCAT and the ABSR were combined with weights of .80 and .20 applied, respectively. These scores were then transmuted to have a mean of 70 and maximum of 100. The passing score varied with education and prior experience. Applicants who received passing scores on the first two predictors could receive up to 15 additional points from the OKT.

The second stage in the hiring process was the Academy Screen. Applicants who passed the OPM Battery were sent to the FAA Academy for a 9-week screen, which involved both selection *and* training (Manning, 1991a). Students spent the first 5 weeks learning aviation and air traffic control concepts and the final 4 weeks being tested on their ability to apply ATC principles in non-radar simulation problems. Applicants could still be denied positions after the 9 weeks on the basis of their scores during this phase. The reported failure rate was 40 percent (Cooper et al., 1994).

This hiring process received much criticism, despite its reported effectiveness and links to job performance. The criticisms revolved around the time (9 weeks for the Academy screen) and cost of such a screening device ($10,000 per applicant). In addition to the FAA investment, applicants made a substantial investment, and the possibility remained that after the 9 weeks an applicant could be denied a position. Finally, there was concern that the combination of screening and training reduced training effectiveness and made it impossible to tailor training needs to individual students.

As a result of these criticisms, the FAA separated selection and training, with the idea that the training atmosphere of the Academy Screen would be more supportive and oriented toward development of ATCSs

once it was separated from selection. This necessitated developing a new selection device to replace the Academy Screen.

The Pre-Training Screen

The FAA introduced the Pre-Training Screen (PTS) in June 1992 to replace the second stage of the hiring process, the Academy Screen. The PTS was developed from a cognitive psychology perspective by Aerospace Sciences, Inc. (ASI) in 1991. It was computer administered and consisted of two parts: the Complex Cognitive Battery and the Air Traffic Scenario Test. For complete descriptions of the components of the PTS and the critical aptitudes covered by these tests, the reader is referred to ASI (1991).

The first part of the PTS, the Complex Cognitive Battery, included five test components: Static Vector/Continuous Memory, Time Wall/Pattern Recognition, Visual Search, Stroop Color-Word Test, and Letter Rotation Test. According to ASI (1991), the Static Vector/Continuous Memory Test was a multimeasure test designed to assess the critical aptitudes of spatial relations, working memory, verbal/numerical coding, attention switching, and visualization. The Time Wall/Pattern Recognition test was designed to assess filtering, movement detection, prioritizing, short-term memory, image/pattern recognition, and spatial scanning. The Visual Search test measured short-term memory and perceptual speed. The Stroop Color-Word Test assessed the critical aptitudes of decoding, filtering, and short-term memory. Finally, the Letter Rotation Test assessed the critical aptitudes of decoding, image/pattern recognition, and visualization. It should be noted that each test in this battery could yield multiple scores.

The second part of the PTS, the Air Traffic Scenario Test, was a low-fidelity work sample test (Broach & Brecht-Clark, 1994). Applicants were given a synthetic, simplified air space to control. This test was designed to assess nearly all of the critical aptitudes of the ATCS job.

Two attempts were made to validate the PTS. The first (ASI, 1991) correlated PTS performance with training criteria (the Academy Screen Comprehensive Test score). Based on correlation analyses, the full set of test scores was reduced to ten (Safety and Delay scores from Air Traffic Scenario; Percent Correct and Mean Correct Reaction Time from Static Vector; Percent Correct and Mean Correct Reaction Time from Continuous Memory; Mean Absolute Time Error from Time Wall; Mean Correct Reaction Time from Pattern Recognition; Stroop Mean Reaction Time for Conflict Stimuli from the Stroop Color-Word Test; and Visual Search Mean Correct Reaction Time from Visual Search). These scores were retained based on their single-order correlation with the criterion, their intercorrelations with other predictor scores, and the multiprocessing nature of the paired test scores (e.g., Air Traffic Safety and Delay).

Multiple regression analyses showed that the Safety score from Air Traffic Scenario and the Percent Correct and Correct Reaction Time scores from the Static Vector test had significant power in predicting the Academy Screen Comprehensive Test score. The betas for the remaining subtest scores were not significant in the context of the other tests. ASI (1991) reported the regression model shown in Tables 3.1.1 and 3.1.2.

The Pre-Training Screen was intended to be used as a secondary screening procedure. Incremental validity was estimated for the OPM battery score, and for the OPM and PTS scores where the OPM score was entered in step 1 and the PTS scores were entered in a block at step 2. The OPM score alone produced an $\underline{R}$ = .226, $\underline{R}$ square = .05. The model using the OPM and PTS scores produced a multiple correlation of $\underline{R}$ = .505, $\underline{R}$ square = .26. The difference of variance accounted for by the addition of the PTS (.26 vs. .05) was significant ($\underline{F}$ = 24.18, $\underline{p}$ <.01). This indicated that the Pre-Training Screen added significantly to the prediction of the Academy Screen Comprehensive Test score, over and above the OPM battery alone.

The second validation attempt (Weltin, Broach, Goldbach, & O'Donnell, 1992) was a concurrent criterion-related study using a composite measure of on-the-job training performance. Scores obtained from the Air Traffic Scenario Test, Static Vector, Continuous Memory, Stroop Color-Word Test, and Letter Rotation Test correlated significantly with the criterion. Using two weighting schemes, the regression-based weighting scheme yielded a correlation of .21, whereas the unit weighting yielded a correlation of .18.

Use of the PTS as a screening device was discontinued in February 1994. The defensibility of the PTS was questioned since it was validated only against training performance criteria. The perception was that the test security of the OPM test, in use since 1981 without revision, had been compromised. Further, several coaching schools provided guarantees to students that they

would pass the OPM battery. For a more complete discussion of prior programs used to screen ATCS candidates before AT-SAT, see Chapter 6 of this report.

Separation and Control Hiring Assessment (SACHA)

In September 1991, the FAA awarded a contract to University Research Corporation for the development and validation of a new test battery for selection of ATCSs. (The outcomes of the SACHA job analysis were covered in more detail in Chapter 2.) By 1996, a comprehensive job analysis was completed on four ATCS options, and the construction of possible predictor tests had begun. The FAA terminated the SACHA contract late in 1996.

A meta-analytic study of much of the previous validation research on the ATCS job was performed as part of SACHA (Schemmer et al., 1996). This study reported on predictors ranging from traditional cognitive ability tests, and personal characteristics instruments, to air traffic control simulations and psychomotor ability measures. The validity studies are summarized in Table 3.1.3.

As reported by Schemmer et al. (1996), for most of the predictor measure categories, the validity coefficients exhibited substantially greater variability than would be expected under the simple explanation of sampling error. This suggests that, in general, some specific predictor measures are relatively more predictive of job performance than others. For example, simulations and math tests have historically been good predictors of controller job performance.

On the basis of the SACHA job analysis, Schemmer et al. (1996) proposed an overall model of the ATCS worker requirements that included a Cognitive Model and a Temperament/Interpersonal Model. The Cognitive Model contained two higher-order constructs, *g* and Processing Operations. Table 3.1.4 displays the construct categories, worker requirements under the higher order construct of *g*, and the tests purported to measure the worker requirements. Schemmer et al. recommended at least one test per worker requirement. As Table 3.1.4 shows, there were some worker requirements for which the project still had not developed tests. For example, their predictor battery did not account for any of the requirements under the rubric of Communication. Additionally, much of the Applied Reasoning construct remained untested, and Numeric Ability (Multiplication/Division), Scanning, and Movement Detection were not addressed.

Table 3.1.5 displays the construct categories, worker requirements under the higher order construct of Processing Operations, and the tests that Schemmer et al. hypothesized would assess the worker requirements. Table 3.1.5 reveals that, for the construct labeled Metacognitive, no tests had been recommended. In addition, Schemmer et al. did not account for Sustained Attention, Timesharing, Scanning, or Movement Detection worker requirements.

Finally, a Temperament/Interpersonal Model was proposed to provide coverage of the worker requirements that did not fit into the Cognitive Model (Table 3.1.6.).

As noted earlier, due to a compromised OPM battery and the elimination of use of PTS, the FAA decided to support the development and validation of a new test battery against job performance criteria. With this decision, a contract was awarded to Caliber Associates, and support for the AT-SAT project was initiated.

AIR TRAFFIC SELECTION AND TRAINING (AT-SAT) PROJECT

One of the challenges facing the AT-SAT research team was to decide what SACHA-generated materials would be adequate for the new battery and how many new tests needed to be developed. This section describes the procedures undertaken to review existing SACHA materials and documents the evaluation of and comments on the battery's coverage of the predictor space. Recommendations for the AT-SAT project were made, based on the review process.

Test by Test Evaluation

A panel of nine individuals was asked to review each test currently available on computer for possible inclusion in the air traffic control predictor test battery. Evaluation sheets were provided for each test, requesting information about the following criteria:

(1) Does the test measure the worker requirement(s) it purports to measure?
(2) Is it a tried-and-true method of assessing the worker requirement(s)?
(3) Does the scoring process support the measurement of the worker requirement(s)?
(4) Is the time allocation/emphasis appropriate?
(5) Is the reading level consistent with job requirements?

(6) Does the test have potential adverse impact?
(7) Is the test construction ready for validation administration?

Short descriptions of each test were provided for this evaluation, along with test information such as number of items, and scoring procedures. The worker requirement definitions used throughout this evaluation process were those listed for the Revised Consolidated Worker Requirements on pages 115-119 of the *SACHA Final Job Analysis Report* (January 1995). Sixteen tests were independently reviewed by the panel members. The results of the independent reviews were discussed at a 3-day meeting. An additional four tests (Letter Factory and the three PTS tests) were reviewed in a similar fashion during the meeting. The 20 tests reviewed were:

Sound	Stix
Scan	Time
Angles	Syllogism
Map	Analogy
Dial Reading	Classification
Headings	Personal Experiences and Attitude
Projection	Questionnaire (PEAQ)
Memory 1 and 2	Letter Factory
Direction and Distance	Air Traffic Scenario (from PTS)
Planes	Time Wall/Pattern Recognition (from PTS) Static Vector/Continuous Memory (from PTS)

The project staff met with the panel members on 5-7 November, l996 to discuss the predictor battery. For each test, independent ratings on each evaluation criterion were collected, and the relative merits and problems of including that test in the predictor battery were discussed. The comments were summarized and recorded.

After the group discussion, panel members were asked to provide independent evaluations on whether or not each test should be included in the predictor battery. For each test, panel members indicated "*Yes*" for inclusion, "*No"* for exclusion, and "*Maybe*" for possible inclusion. The *Yes-No-Maybe* ratings were tallied and summarized.

Selection of a Subset of Tests

The next step involved selecting a subset of the 20 tests for inclusion in the predictor battery. Considerations included both the *Yes-No-Maybe* ratings (based on multiple, test-specific criteria), and how well different subsets of tests covered the predictor domain. The list of worker requirements, rank ordered by incumbent importance ratings, as described in Chapter 2, was used to help determine whether different tests or subsets of tests covered the critical job requirements. These investigations and recommendations are summarized below.

Nine tests that received a preponderance of *Yes* ratings were measuring critically important job requirements and appeared to be relatively non-overlapping. These were Scan, Letter Factory, Sound, Dial Reading, PEAQ, Analogy, Air Traffic Scenario (ATS), Time Wall/Pattern Recognition (TW), and Static Vector/ Continuous Memory (SV). These nine tests were recommended for inclusion in the predictor battery. All required modifications before they were deemed ready for administration. Examples of the recommended modifications follow.

· Scan: Increase clarity of figures, increase number of test items, and possibly use mouse to decrease keyboard skills requirement.
· Letter Factory: Increase planning/thinking ahead requirement (e.g., by adding boxes at top of columns).
· Sound: Investigate possibility of changing scoring to allow partial credit for partially correct answers.
· Dial Reading: Increase number of items, decrease time limit, investigate fineness of differentiation required.
· PEAQ: Decrease number of items (focus on only critically important worker requirements), replace random response items, edit response options for all items.
· Analogy: Delete information processing component, possibly add some of the Classification test items.
· ATS, TW, SV: Separate individual tests from PTS administration, shorten tests.

Three additional tests were strongly considered for inclusion, but with further modifications: Planes, Projection (perhaps modified to represent above-ground stimuli), and a numerical ability test. The latter represented a worker requirement that was otherwise not measured by the set of "included" tests. The plan for the numerical ability test was initially to include items modified from several existing tests: Headings, Direction and Distance, and Time, all of which include components of on-the-job numerical computation. Angles and Dials would be added to round out the numeric ability construct.

Possible Gaps in Coverage

Viewing the 12 tests on the preliminary list as a whole, the coverage of worker requirements appeared quite good. However, a few important worker requirements remained unrepresented: reading comprehension, memory, and street physics. There was some discussion about including measures of these three requirements. A reading test could be prepared, using very short, face valid passages, where the passage and the test question could be displayed on screen at the same time. Discussions about adding a memory test primarily focused on a modification of the Map test, which would require candidates to indicate whether the stimulus had changed or remained the same since they "memorized" it. The possibility of finding a published test measuring street physics was also discussed. If such a published test could not be found, some kind of mechanical or abstract reasoning test might be included as a close approximation.

Excluded Tests

Three of the 20 tests reviewed were deleted from further consideration: Stix, Map (except as it might be revised to cover memory), and Syllogism. These tests were deleted because of problems with test construction, and/or questionable relevance for important job requirements, and/or redundancy with the included measures.

Additional Recommendations

Several additional recommendations were made concerning the predictor battery and its documentation. The first was that all tests, once revised, be carefully reviewed to ensure that the battery adheres to good test construction principles such as consistency of directions and keyboard use, reading/vocabulary level, and balancing keyed response options.

A second recommendation was that linkages be provided for worker requirements that do not currently have documented linkages with ATCS job duties. The current documentation (from the Job Analysis report) was incomplete in this regard.

A third recommendation was to "pilot test" the predictor set in February l997. It was thought that this would yield the kind of data needed to perform a final revision of all predictors, select the best test items, shorten tests, reduce redundancy across tests, ensure clarity of instructions, and so on.

AT-SAT ALPHA BATTERY

Based on the reviews and recommendations of the expert panel, the AT-SAT researchers developed the predictor battery to be pilot tested, called the Alpha Battery. It consisted of 14 tests given across five testing blocks. They were:

Block A:	Air Traffic Scenarios
Block B:	Sound test and Letter Factory test
Block C:	Dials test, Static Vector/Continuous Memory test, and Experiences Questionnaire (formerly PEAQ)
Block D:	Time Wall/Pattern Recognition test, Analogy test, and Classification test
Block E:	Word Memory test, Scan test, Planes test, Angles test, and Applied Mathematics test

A short description of the tests follows. In a few instances, details reflect modifications made in the alpha pilot tests for use in the beta (validation) testing.

Air Traffic Scenarios Test

This is a low-fidelity simulation of an air traffic control radar screen that is updated every 7 seconds. The goal is to maintain separation and control of varying numbers of simulated aircraft (represented as data blocks) within the participant's designated airspace as efficiently as possible. Simulated aircraft either pass through the airspace or land at one of two airports within the airspace. Each aircraft indicates its present heading, speed, and altitude via its data block. There are eight different headings representing 45-degree increments, three different speed levels (slow, moderate, fast), and four different altitude levels (1=lowest and 4=highest).

Separation and control are achieved by communicating and coordinating with each aircraft. This is accomplished by using the computer mouse to click on the data block representing each aircraft and providing instructions such as heading, speed, or altitude. New aircraft in the participant's airspace have data blocks appear in white that turn green once the participant has communicated with them. Rules for handling aircraft are as follows: (1) maintain a designated separation distance between planes, (2) land designated aircraft at their proper airport and in the proper landing direction flying at the lowest altitude and lowest speed, (3) route aircraft passing through the airspace to their designated exit at the highest altitude and highest speed. The

version of ATST that was incorporated in the alpha battery was modified to operate in the windows environment (Broach, 1996).

Analogy Test

The Analogy test measures the participant's ability to apply the correct rules to solve a given problem. An analogy item provides a pair of either words or figures that are related to one another in a particular way. In the analogy test, a participant has to choose the item that completes a second pair in such a way that the relationship of the items (words or figures) in the second pair is the same as that of the first.

The test has 57 items: 30 word analogies and 27 visual analogies. Each item has five answer options. The scoring is based primarily on the number of correct answers and secondarily on the speed with which the participant arrived at each answer. Visual analogies can contain either pictures or figures. The instructions inform the participant that the relationships for these two types of visual analogies are different. Picture analogies are based on the relationships formed by the meaning of the object pair (e.g., relationships of behavior, function, or features). Figure analogies are based on the relationships formed by the structure of the object pair (e.g., similar parts or rotation).

Angles Test

The Angles test measures the participant's ability to recognize angles. This test contains 30 multiple-choice questions and allows participants up to 8 minutes to complete them. The score is based on the number of correct answers (with no penalty for wrong or unanswered questions). There are two types of questions. The first presents a picture of an angle, and the participant chooses the correct answer of the angle (in degrees) from among four response options. The second presents a measure in degrees, and the participant chooses the angle (among four response options) that represents that measure.

Applied Mathematics Test

This test contains 30 multiple-choice questions and allows participants up to 21 minutes to complete them. The score is based on the number of correct answers (with no penalty for wrong or unanswered questions). The test presents five practice questions before the test begins. Questions such as the following are contained on the test:

> A plane has flown for 3 hours with a ground speed of 210 knots. How far did the plane travel?

These questions require the participant be able to factor in such things as time and distance to identify the correct answer from among the four answer choices.

Dials Test

The Dials test is designed to test the participant's ability to quickly identify and accurately read certain dials on an instrument panel. The test consists of 20 items completed over a total time of 9 minutes. Individual items are self-paced against the display of time left in the test as a whole. Participants are advised to skip difficult items and come back to them at the end of the test. The score is based on the number of items answered correctly. The test screen consists of seven dials in two rows, a layout which remains constant throughout the test. Each of the seven dials contains unique flight information. The top row contains the following dials: Voltmeter, RPM, Fuel-air Ratio, and Altitude. The bottom row contains the Amperes, Temperature, and Airspeed dials.

Each test item asks a question about one dial. To complete each item, the participant is instructed to (1) find the specified scale on the instrument panel; (2) determine the point on the scale represented by the needle; (3) find the corresponding value among the five answer options; (4) use the numeric keypad to press the number corresponding to the option.

Experiences Questionnaire

The Experiences Questionnaire assesses whether participants possess certain work-related attributes by asking questions about past experiences. There are 201 items to be completed in a 40-minute time frame. Items cover attitudes toward work relationships, rules, decision-making, initiative, ability to focus, flexibility, self-awareness, work cycles, work habits, reaction to pressure, attention to detail, and other related topics. Each question is written as a statement about the participant's past experience and the participant is asked to indicate their level of agreement with each statement on the following 5-point scale: 1= Definitely true, 2= Somewhat true, 3= Neither true nor false, 4= Somewhat false, 5= Definitely false.

Letter Factory Test

This test simulates a factory assembly line that manufactures letters A to D of the alphabet. Examinees perform multiple and often concurrent tasks during the

test with aid of a mouse. Tasks include: (1) picking up letters of various colors from a conveyor belt and loading them into boxes of the same color; (2) moving empty boxes from storage to the loading area; (3) ordering new boxes when supplies become low; (4) calling Quality Control when defective letters appear; and (5) answering multiple-choice questions about the factory floor display. The test is comprised of 18 test parts; each part begins when the letters appear at the top of the belts and ends with four multiple-choice questions. Awareness questions assess the state of the screen display. Easier questions are presented during lulls in assembly line activity and assess the current state of the display. More difficult questions are asked during peak activity and assess what a future display might look like.

Overall scores on the LFT are based on (1) the number of boxes correctly moved to the loading area; (2) the time it takes to move a box after it is needed; (3) the number of letters correctly placed into boxes; and (4) answers to the awareness questions. The following actions lower test scores: (1) allowing letters to fall off the end of a belt; (2) placing letters in an incorrect box; (3) not moving a box into the loading area when needed; and (4) attempting to move the wrong box into the loading area.

Planes Test

The Planes test contains three parts, each with 48 items to be completed in 6 minutes. Each individual item must be answered within 12 seconds. *Part 1*: Participants perform a single task. Two planes move across a screen; one plane is red, the other is white. Each plane moves toward a "destination" (a vertical line) at a different speed. The planes disappear before they reach their destinations, and the participant must determine which plane would have reached its destination first. To answer each item, the participant presses the "red" key if the red plane would have reached the destination first, and the "white" key if the white plane would have arrived first. Participants can answer while the planes are still moving, or shortly after they disappear. *Part 2:* Part 2 is similar to Part 1, but participants must now perform two tasks at the same time. In this part of the test, participants determine which of two planes will arrive at the destination first. Below the planes, a sentence will appear stating which plane will arrive first. The participant must compare the sentence to their perception of the planes' arrival, and press the "true" key to indicate agreement with the statement, or the "false" key to indicate disagreement. *Part 3:* Participants perform the same tasks as in Part 2, but the statements below the planes are a little more difficult to analyze. In all other respects, the participants perform in the same manner.

Scan Test

In the Scan test, participants monitor a field that contains discrete objects (called data blocks) which are moving in different directions. Data blocks appear in the field at random, travel in a straight line for a short time, then disappear. During the test, the participant sees a blue field that fills the screen, except for a 1-inch white bar at the bottom. In this field, up to 12 green data blocks may be present. The data blocks each contain two lines of letters and numbers separated by a horizontal line. The upper line is the identifier and begins with a letter followed by a 2-digit number. The lower line contains a 3-digit number. Participants are scored on the speed with which they notice and respond to the data blocks that have a number on the lower line outside a specified range. Throughout the test, this range is displayed at the bottom of the screen (e.g., 360-710). To "respond" to a data block, the participant types the 2-digit number from the upper line of the block (ignoring the letter that precedes it), then presses "enter."

Sound Memory Test

The Sound Memory test measures a participant's listening comprehension, memory, and hand-eye coordination. Participants must hear, remember, and record strings of numbers varying in length from 5 to 10 digits. After the digits have been read, there is a brief pause. Then a yellow box will appear on screen, and participants must type in the numbers they heard and remembered, in the order presented orally. Participants may use the backspace to delete and correct the numbers they enter, and press the "enter" key to submit the answer.

Each participant's score equals the total number of digits the participant remembers correctly. If the participant transposes two digits then half-credit is given. Items must be answered in the order presented—participants cannot skip and return to previous items. If too *few* digits are typed then the missing digits are scored as incorrect; if too *many* digits are typed then the extra digits are ignored. The object is simply to recall digits heard in the correct order.

Time Wall/Pattern Recognition Test

The Time Wall/Pattern Recognition test consists of two tasks that measure the examinee's ability to judge the speed of objects and to compare visual patterns at the

same time. In the time judgment task, the participant watches a square move from left to right and estimates when it will hit a wall positioned on the right side of the display screen. In the pattern comparison task, the participant determines whether two patterns are the same or different from each other. Each exercise begins with a square moving toward a wall at a fast, medium, or slow speed. After a short while, the square disappears behind the pattern recognition screen. The participant must hit the stop key at the exact moment the square hits the wall.

In the pattern comparison task, the participant is shown two blue circles, each with an overlay pattern of white dots. Test takers are requested to press the "same" key if the patterns are the same or press the "differ" key if the patterns are different. Concurrently, participants should press the "stop" key when they think the square will hit the wall, even if they are in the middle of comparing two patterns. Participants are scored upon how quickly they respond without making mistakes. The score is lowered for each incorrect judgment.

Word Memory Test

The Word Memory test presents a series of 24 words in an artificial language (i.e., "SPRON") and their associated English equivalents. The goal is to memorize the 24 SPRON words and their English equivalents and then recall these at two different testing times: one immediately following a practice session and another in a subsequent testing block. The practice session lasts 4 minutes, during which the list of 24 SPRON words and their English equivalents are displayed in a box to the right of the display screen while the multiple-choice items are displayed on the left. The practice items allow the test takers to apply their memory by allowing them to review the SPRON-English list of words as a reference. The first testing session starts immediately following the practice session and lasts 5 minutes. The second testing session starts in a subsequent testing block (after a break time) and also lasts for 5 minutes. Each multiple-choice item displays the SPRON word as the item stem and displays five different English equivalents as the five response alternatives.

Conclusion

The initial AT-SAT test battery (Alpha) was professionally developed after a careful consideration of multiple factors. These included an examination of the SACHA job analysis and prior job analyses that produced lists of worker requirements, prior validation research on the ATCS job, and the professional judgment of a knowledgeable and experienced team of testing experts.

CHAPTER 3.2

Air Traffic - Selection and Training Alpha Pilot Trial After-Action Report

Claudette Archambault, Robyn Harris
Caliber Associates

INTRODUCTION

The purpose of this report is to document the observations of the Air Traffic - Selection and Training Completion (AT-SAT) predictor battery (alpha version) pilot trial. The AT-SAT predictor battery is a series of tests in five blocks (A through E) of 90 minutes each and four different ending blocks of 20 minutes each. The pilot test was administered February 19 through March 2, 1997, in the Air Traffic Control School at the Pensacola Naval Air Station in Pensacola, Florida. Participants consisted of 566 students stationed at the Naval Air Technical Training Center (NATTC). Of the 566 students, 215 of the participants were currently enrolled in the Air Traffic Control School and 346 were students waiting for their classes at NATTC to begin. (The status of five participants was unknown.)

This report contains the following sections:

- Pilot Test Description and Procedures
- General Observations
- Feedback on Test Block A
- Feedback on Test Block B
- Feedback on Test Block C
- Feedback on Test Block D
- Feedback on Test Block E
- Feedback on the Ending Block

The report concludes with a summary of all of the feedback and observations.

The At-SAT Pilot Test Description and Administration Procedures

The following sections describe the AT-SAT pilot test and pilot test administration procedures.

AT-SAT Pilot Test Description

The AT-SAT Pilot Test is a series of five test blocks (Blocks A through E) and Ending Blocks. (There are four different Ending Blocks.) The tests are designed to measure different aptitudes required for successfully performing the job of air traffic controller. Tests are subdivided as follows:

- Block A contains one test entitled Air Traffic Scenarios (ATS).
- Block B contains the Sound Test and the Letter Factory Test (LFT).
- Block C contains the Dials Test, Static Vector/Continuous Memory Test (SVCM), and Experiences Questionnaire.
- Block D contains the Time Wall/Pattern Recognition Test (TWPR), the Analogy Test, and the Classification Test.
- Block E contains the Word Memory Test, the Scan Test, the Planes Test, the Angles Test, and the Applied Mathematics Test.

Depending on the Participant's group number, the Ending Block consisted of one of the following.

- the LFT
- the ATS
- the SVCM and Word Memory tests
- the Word Memory and TWPR tests

The following section describes the test administration procedures including the sequence of the testing blocks for groups of participants.

Pilot Test Administration Procedures

Participants were arranged in five groups of ten (Groups 1 through 5). Test Administrators (TAs) supplied the testing rooms with 55 computers. Fifty of the computers were used for testing stations; five were

failure-safe or recovery stations. Recovery stations were reserved for use by participants when TAs were not able to restore operation to a malfunctioning computer.

In one classroom, there were 33 computers (30 for testing and three failure-safe computers): Groups 1, 2, and 3 were tested on computers 1 through 30 (See Exhibit 3.2.1). In a second classroom, there were 22 computers (20 for testing and two failure-safe computers). Groups 4 and 5 were tested in the second room on computer numbers 31 through 50. Exhibit 3.2.1 displays the sequencing of test blocks. (The exhibit does not reflect breaks.)

Participants were offered a minimum of a ten-minute break between each of the five testing sections. Because the tests are self-paced, participants were not required to take the 10-minute breaks between blocks. They were required to take a 1.5 hour meal break between sessions two and three.

GENERAL OBSERVATIONS

This section presents some general observations about the entire AT-SAT Battery Pilot Test. The remarks in this section address the instructions, the test ending, the purpose of the tests, and the introductory block.

Instructions

Instructions for several of the tests in the battery need improved clarity. Participants often did not understand the test instructions as written but proceeded with the tests, anticipating that the objective of the tests would become more clear as the tests proceeded. Too often, however, participants still did not understand the objective even after attempting a few examples. (After participants completed the examples, they would often raise their hand and ask for further instructions.) Therefore, any practice sessions for the tests did not clarify the confusing instructions. The test instructions that need revision and feedback for specific test blocks are discussed in the following sections.

Purpose of the Tests

Participants also required further clarification of the purpose of tests within the blocks during the practice session (instead of before or after the test). Perhaps a short paragraph including the aptitudes that are being tested would clarify the purpose of certain tests.

In addition to more specific test instructions, an introductory screen at the start of each block, to include the number of different tests within the specific block, the names of the tests and a short description of each test, the aptitudes that are being tested, and the time allotted for each test should be added. This screen may eliminate discrepancies where participants are unclear as to whether to continue with the other tests in the block when they reach the end of a test.

Test Ending

The end of each test should include a brief statement in a text box stating that the participant has completed the current test and should press enter, or click on the continue button (with the mouse pointer) to proceed to the next test in the block. The text box could also state the number of tests completed and the number of tests that remain for each block.

Currently, some blocks do not indicate the end of the block with a text box. Some tests simply go to a blue screen and do not indicate that the test has indeed ended. The final test in a block should indicate not only that the current test is finished but also that the participant has indeed completed all tests within the block and that they should raise their hand to speak with the Test Administrator.

Not all of the tests indicate the results of the tests and/or practice sessions. For consistency, either all tests should display results, or all tests should not display results.

Introductory Block

The addition of an Introductory Block (IB) is recommended. The IB could explain of the general purpose of the testing a modified version of the Keyboard Familiarization section and the current Background Information questions.

The explanation of the general purpose of the test might also include a brief description of the evolution of the test (how the FAA came to design this specific testing procedure). This section could describe the types of tests and the general purpose of the tests (i.e., ability to multitask, ability to follow instructions, skill with plane routing procedures, etc.). Finally, general grading/scoring procedures could be explained with more specific explanations within each of the tests.

The Keyboard Familiarization (KF) currently includes instruction and practice for the number keys and the "A, B, C" keys (after the Test Administrator exchanges the slash, star, and minus keys with the A, B, and C keys) on the numeric pad on the right side of the keyboard. Instructions should be modified to include the names of the tests requiring the use of these keys.

Directions under KF should also include the names of the tests that will require the use of the numerical keys on the top of the keyboard. A practice session should also be included for these keys to allow participants to become acquainted with the placement of their hands at the top of the keyboard.

Background information questions should be included within the Introductory Block. This will allow participants to practice using the keyboard outside of a testing block. It will also allow them to ask the Test Administrator questions about the test, use of the keyboard, the placement of hands on the keyboard, and so on.

FEEDBACK ON TEST BLOCK A

Block A was the only block that consisted of only one test. Therefore, the comment below applies to the Air Traffic Scenarios Test (ATST).

The ATST requires participants to manipulate the heading, speed, and level (or altitude) of planes in their airspace. On the testing screen, participants see the airspace for which they are responsible, two airports for landing planes, and four exits for routing planes out of the airspace. The screen also displays the controls for directing planes: (1) the heading (to manipulate individual plane direction), (2) the speed (slow, medium, or fast), (3)

the level (1, 2, 3, 4). Finally, a landing heading indicator is displayed that informs the participant of the direction to land planes at each of the airports.

Instructions

Instructions for the ATST may need more clarification. Often, participants required further clarification on:

· the meaning of each of the plane descriptors that appear on the screen
· the difference between white and green planes
· the need to click on the graphic (depicted as an arrow) that represents the plane (versus the text descriptors of the plane) to change the heading, level, and speed.

Instructions need to be rewritten to include more details on the descriptors accompanying the planes. Perhaps in the instructions section, the descriptors can be enlarged on the screen with an arrow pointing to the definition of the letters and number as in Exhibit 3.2.2.

New Planes

New planes that appear in the participant's airspace are white (while all other planes are green). The white planes remain circling at the point where they entered the airspace until they receive acknowledgment from the controller (by clicking on the graphic with the mouse pointer). Often during testing participants did not understand the purpose of the white planes in their airspace. They would leave the white planes circling and never manipulate their heading, speed, or level. White planes need to be more clearly defined as new planes in the controller's airspace that require acknowledgment by the controller.

Countdown

At the start of a scenario, participants often did not notice the countdown (on the counter at the bottom right-hand corner of the screen) before the beginning of a test. There is a delay (of approximately 7 seconds) between the time the test initially appears on the screen and the time the participant can perform an action to the planes on the screen.

During this delay, some participants continuously pushed the "enter" button, which would often result in one of two consequences: (1) The computer screen would permanently freeze (such that the system would need to be rebooted); (2) at the end of the test, the participant received the plain blue screen (indicating that the test was complete). However, once the Test Administrator closed-out the blue screen and returned to the program manager, there would remain a row of several icons with each icon indicating an air traffic scenario. The Test Administrator would need to presume that the next correct test in the sequence was the first in the row of icons and double click on that icon to begin a scenario. At the end of each scenario, the Test Administrator would double-click on the next scenario in the row of icons until all scenarios were complete.

For participants to clearly see that there is a delay before they can manipulate the planes on the screen, perhaps the countdown timer can be moved to a more conspicuous position in the middle of the screen (as in the Static Vector/Continuous Memory Test). An alternative would be to display the counter in a brightly colored text box (still in the bottom right-hand corner of the screen). After the countdown timer had finished, the text box could change colors and blend with the other instructions on the screen.

Landing Heading Indicator

Participants often did not notice the landing heading indicator located on the bottom right-hand corner of the screen. Others noticed the arrow but did not understand its purpose. Further instruction on the location and the purpose of the landing heading indicator may be necessary. Perhaps during the practice scenario, a text box can flash on the screen pointing out when the participant has landed a plane in the incorrect direction. The same idea may be useful to point out other participant errors during the practice session(s).

FEEDBACK ON TEST BLOCK B

This section details the observations for the two tests in Block B: the Sound Test and the Letter Factory Test. Specific comments about each test are provided below.

Sound Test

For this test, the participant uses headphones to listen to a sequence of numbers. Then the participant must repeat the sequence of the numbers heard using the right-hand numeric keypad to record the sequence of numbers.

Failures

It was found in the first day of testing that computers would lock or fail if the Sound Test was run after any other blocks. In other words, unless Block B was first in the sequence of testing, computers would fail (at the moment participants are prompted for sound level adjustment) and need to be rebooted. This proved disruptive to other participants and delayed the start of the test (since Test Administrators can only aid one or two participants at a time). To prevent failures during the testing, Test Administrators would reboot every computer before the start of Block B. Still, the software would sometimes fail at the moment the participant is requested to adjust the volume to their headphones via the keyboard (versus the sound level adjustment located directly on the headphones). On occasion, the Sound test would still fail, but after again rebooting the computer, the program recovered.

After several attempts to restore the program where there were repeated failures, the computer still did not allow the participant to continue with the test. In these cases where a computer failed repeatedly, participants would be moved to a failure-safe computer. It is likely that such failures are the result of the hardware or hardware configuration, rather than the software.

There is another possible reason for the failure of the Sound Test. Some participants would attempt to repeatedly adjust the volume of their headsets with the numbers on the top of the keyboard rather than using the number keys on the right-hand side of the keyboard (as instructed). It is possible that the use of these keys caused some of the failures.

Removal of Headphones

Upon completion of the Sound Test, participants often keep the headphones on their ears throughout the second test in Block B. The addition of some text at the end of the test to instruct participants to remove their headphones might be useful.

Letter Factory Test

This test measures four abilities required to perform air traffic controller jobs. These abilities are: (1) planning and deciding what action to take in a given situation through the application of specific rules; (2) thinking ahead to avoid problems before they occur; (3) continuing to work after being interrupted; and (4) maintaining awareness of the work setting.

Test Instruction

The test instructions are clear and well-written. Few participants had questions in reference to the tasks they were to perform once the test began.

Demonstration

Participants were often confused during the demonstration because the pointer would move when they moved the mouse, but they could not "click" and manipulate the screen. Participants would ask Test Administrators if they had already begun the test since they could move the pointer. Perhaps the mouse can be completely disabled during the demonstration to eliminate confusion. Disabling the mouse would allow participants to concentrate on the instructions since they would not be distracted by movement of the mouse.

Mouse Practice Instructions

Instructions for the mouse practice session are not clear. The objective of the mouse practice is for the participant to click on the red box in the middle of the

screen and then click on the conveyer belt that illuminates. Participants are often unsure of the objective. Perhaps text box instructions can be displayed on the screen that direct the participant to click on the red box. As the participant clicks on the red box, another instruction screen would appear, telling the participant to click on the illuminated conveyer belt. After a few sequences with text box instruction, the instructions could be dropped.

Some participants had difficulty completing the mouse practice session. They continuously received messages instructing them to "...move the mouse faster and in a straight line." Perhaps there should be a limit to the number of mouse practice exercises. It is possible that some participants are not capable of moving the mouse quickly enough to get through this section.

FEEDBACK ON TEST BLOCK C

This section details the observations and suggestions for the three tests in Block C: the Dial Test, Static Vector/Continuous Memory Test, and Experiences Questionnaire. Specific comments about each test are provided below.

Dial Test

This measures the participant's ability to quickly and accurately read dials on an instrument panel. Participants did not appear to have difficulties with this test. Test Administrators rarely received questions from participants about this test.

Static Vector/Continuous Memory Test

This measures the participant's ability to perform perceptual and memory tasks at the same time. The perceptual task involves determining whether two planes are in conflict. The memory task involves remembering flight numbers. On each trial, the screen displays a plane conflict problem on the left side of the screen and a memory problem on the right side of the screen. An attention director indicates which problem the participant is to work on and is located in the middle at the bottom of the screen.

Instructions

Participants do not understand the instructions for the Memory part of the test. Numerous participants asked for clarity on what numbers they were to compare. They often think they should be comparing the two numbers that are on the screen at that moment, rather than comparing the top number of the current screen to the bottom number of the previous screen.

The example for determining the conflict for the Static Vector questions is not clear. The rule about the planes requiring 2000 feet difference was confusing because they did not understand that, although the altitude is actually displayed in hundreds of feet, the altitude represents thousands of feet.

Keyboard Issues

Many participants attempted to use the numerical keys on the right-hand side of the keyboard to answer the items rather than the using the keys on the top of the keyboard as instructed. When participants use the right-hand keypad, their answers are not recorded. The keys to be used for this test need to be stated more explicitly.

Participants may be using the right-hand keypad because of the instruction they receive in the Keyboard Familiarization (KF) section at the beginning of the testing. The current version of the KF only provides instruction for use of the keypads on the right-hand side of the keyboard. The KF does not instruct participants on the use of the numerical keys on the top of the keyboard.

As noted previously, the KF needs to be modified to include instructions on the use of the keys on the top of the keyboard. For data to be properly collected, it is critical for participants to use the top keys.

Experiences Questionnaire

The Experiences Questionnaire determines whether the participant possesses work-related attributes needed to be an air traffic controller. Participants generally did not ask any questions about the Experiences Questionnaire. The occasional inquiry was in reference to the purpose of certain questions. Test Administrators did not receive questions about the wording of the items.

FEEDBACK ON BLOCK D

This section details the observations and suggestions for the three tests in Block D: the Time Wall/Pattern Recognition Test; the Analogy Test; and the Classification Test. Specific comments about each test are provided below.

Time Wall/Pattern Recognition Test

This test measures the participant's ability to judge time and motion and make perceptual judgments at the same time. The time judgment task involves watching a ball move (from the far left-hand side of the screen) and estimating when it will hit a wall (located on the far right of the screen). The pace of the ball is different for every scenario. The perceptual task involves determining whether two patterns are the same or different. These tasks must be performed concurrently by the participant. The following paragraphs provide observations and suggestions for this test. This section includes observations and suggestions for improving the Time Wall/Pattern Recognition test in Block D.

Location of the Broken Wall

When the participant does not stop the ball from hitting the wall in a timely manner, the screen displays a broken wall. However, the broken wall appears in the middle of the screen, rather than on the right-hand side of the screen. In reference to this, participants often asked how they were to determine when the ball would hit the wall if the wall was always moving. Test Administrators had to explain that the wall did not move, but that once the ball broke through the wall, the screen displayed the distance past the wall the ball had moved. To eliminate confusion, perhaps the broken wall can remain on the right-hand side of the screen and just appear broken rather than being moved to the center of the screen.

Keyboard

As with the Static Vector/Continuous Memory Test, many participants attempted to use the numerical keys on the right-hand side of the keyboard to answer the items rather than using the keys on the top of the keyboard as instructed. When participants use the right-hand keypad, their answers are not recorded. The keys to be used for this test need to be stated more explicitly.

Participants may be using the keypad because of the instruction they receive in the Keyboard Familiarization (KF) section at the beginning of the first block of testing. The current version of the KF only provides instruction for using the keypad. The KF does not instruct participants on the use of the numerical keys on the top of the keyboard.

Analogy Test

The Analogy Test measures the participant's reasoning ability in applying the correct rules to solve a given problem. The participant is asked to determine the relationship of the words or pictures in set A and use this relationship to complete an analogy in set B. The following paragraph provides observations and suggestions for this test.

Level of Difficulty

The vocabulary level and the types of relationships depicted in the Analogy Test may have been too difficult for the pilot test participants. Perhaps the questions can be revised to require a lower level of vocabulary and reasoning skills for the participants.

Classification Test

This also measures the participant's reasoning ability in applying the correct rules to solve a given problem. The Classification Test is similar to the Analogy Test, except that the participant is required to determine the relationship of three words or pictures and use this relationship to complete the series with a fourth word or picture. The following paragraph provides observations and suggestions for the improvement of this test.

Level of Difficulty

Similar to the issues discussed with the Analogy Test, many of the items in the Classification Test appeared to be difficult for the pilot test population. The Classification Test could be revised to allow a lower level of vocabulary and reasoning skills.

FEEDBACK ON TEST BLOCK E

This section details the observations and suggestions for the five tests in Block E. Specific comments about each test are provided below.

Word Memory Test

The Word Memory Test requires the participant to remember the English equivalents for words in an artificial language called "Spron." The following paragraphs provide observations and suggestions for this test.

Level of Difficulty

The majority of participants appeared to understand how to respond to this test. The practice session for this test seemed to work well in preparing participants for the actual test questions.

Erroneous Text Boxes

Several text boxes appear during this test that should be removed for future versions of the Word Memory Test. The test provides the participant with a text box at the end of the test that displays a total score. This is inconsistent with many of the other tests in the AT-SAT Battery that provide no final scores to the participants. Also, when the test begins, a text box appears, which prompts the participant to "Press 'Enter' to begin." Once the participant presses enter, another text box appears that prompts the participant to "Please be sure Num Lock is engaged." Because these text boxes are irrelevant, the software should eliminate this message in future versions.

Scan Test

The Scan Test measures a participant's ability to promptly notice relevant information that is continuously moving on the computer screen. Participants are provided with a number range and asked to type the identifier for numbers that appear on the screen outside of that range. A revised version of this test was installed midway through the pilot test, which changed the process for recording data but did not change the appearance or the performance of the test for the participants. The following paragraphs provide observations and suggestions for the improvement of this test.

Instructions

While the instructions for the test seemed clear, participants had some common misunderstandings with the test instructions. First, participants typed the actual numbers which were outside of the number range instead of the identifier numbers. This confusion might be alleviated by revising the text that appears on the bottom of the screen during the test. It currently states, "Type the identifier numbers contained in the data blocks with the lower line numbers falling beyond the range." It could be revised to state, "Type the identifier numbers contained in the data blocks (following the letter) with the lower line numbers falling beyond the range." Second, participants did not know to push "Enter" after typing the identification numbers. This confusion might be alleviated by highlighting the text that appears at the bottom of the screen during the test to "Press 'Enter' to record this selection." Third, participants did not know whether the instructions to identify numbers "outside the range" were inclusive of the numbers at the top and bottom of the range. This issue should be explicitly stated in the test instructions.

Computer Keyboards

Since the directions instructed the participants to respond as quickly as possible, in their haste, many participants were pressing the numeric keys very hard. The banging on the keyboard was much louder with this test than with any of the other tests; this affect the longevity of the numeric keys when this test is repeated numerous times.

Planes Test

The Planes Test measures the participant's ability to perform different tasks at the same time. The Planes Test consists of three parts. In Part one, the participant uses the "1" and the "3" keys to determine whether the red plane (1) or the white plane (3), which are at varying distances from their destinations, will reach its destination first. In Part two, the participant uses the "1" and the "3" keys to determine if a statement about the red and white planes as they are in motion is true (3) or false (1). In Part three, the participant uses the "1" and the "3" keys to determine if a statement about the arrival of the red and white planes at their destination are true (3) or false (1), but unlike in Part two, the planes are at varying distances from their destinations. The following paragraphs provide observations and suggestions for the improvement of this test.

Practice Sessions

The practice sessions preceding the first two parts of the Planes Tests are somewhat lengthy. There are 24 practice items that the participant must complete before the actual test of 96 items. If the number of practice items were reduced by one half, the participants would still have enough practice without becoming bored before the actual test begins.

Level of Difficulty

Participants appeared to be challenged by the Planes Test. One factor that added to the level of difficulty for the participants was that the response keys for Parts two and three of this test are: 1 = "false" and 3 = "true." It was more intuitive for many participants that 1 = "true" and 3 = "false" thus, they had a difficult time remembering

which keys to use for true and false. This might have caused participants more difficulty than actually determining the correct answer to the statements. If the labeling of the true and false response keys cannot be modified in future software versions, a message box can be created to remain on the screen at all times that indicates 1 = "false" and 3 = "true."

Test Results

Once the participant provides a response to an item on the Planes Test, a results screen appears indicating whether the response was "right" or "wrong." This is inconsistent with many of the other tests in the AT-SAT Battery that do not indicate how a participant performs on individual test items, in addition to further lengthening an already lengthy test.

Angles Test

This measures a participant's ability to recognize angles and perform calculations on those angles. The following paragraph provides observations and suggestions for this test.

Level of Difficulty

Participants appeared to be challenged by this test, although it seemed as if they could either very quickly determine a response about the measure of an angle, or it took them some time to determine their response.

Applied Mathematics Test

This measures the participant's ability to apply mathematics to solve problems involving the traveling speed, time, and distance of aircraft. The following paragraphs provide observations and suggestions for the improvement of this test.

Instructions

A sentence should be included in the instructions that no pencils, paper, or calculators may be used during this test. Many pilot test participants assumed that these instruments were allowed for this portion of the test.

Level of Difficulty

Many participants appeared to have difficulty determining the best answer to these mathematical questions. Several participants spent so much time trying to calculate an answer that they ran out of time and were not able to complete this test. Perhaps the level of difficulty of the applied mathematics questions can be reduced.

FEEDBACK ON THE ENDING BLOCK

This section details the observations and suggestions for the four retests included in the Ending Block. Specific comments about each ending test block is provided below.

Letter Factory Re-Test

Participants in Group One (computers 1-10) and Group Five (computers 41-50) completed a re-test of the Letter Factory as their Ending Block. This version of the Letter Factory Test does not provide the participant with any test instructions or opportunities to practice before beginning the test. However, participants appeared to have little difficulty remembering the instructions for this test from Block B.

Air Traffic Scenarios Re-Test

Participants in Group Two during the pilot test (computers 11-20) completed a re-test of the Air Traffic Scenarios as their Ending Block. This re-test allows the participant to review the instructions before beginning the abbreviated-length version of the Air Traffic Scenarios. The proposed revisions to the Air Traffic Scenarios Test in Section 4 of this report also apply to this version of the test in the Ending Block.

Static Vector/Continuous Memory and Word Memory Re-Test

Participants in Group Three (computers 21-30) completed a re-test of the Static Vector/Continuous Memory Test and the Word Memory Test as their Ending Block. The re-test of the Static Vector/Continuous Memory Test allows the participant to review the instructions but does not provide a practice session before the actual test begins. The proposed revisions to the Static Vector/Continuous Memory Test in Section 6.2 of this report and to the Word Memory Test in Section 8.1 of this report also apply to these versions of the tests in the Ending Block.

Word Memory and Time Wall/Pattern Recognition Re-Test

Participants in Group Four (computers 31-40) completed a re-test of Word Memory and the Time Wall/ Pattern Recognition as their Ending Block. The re-test of the Time Wall/Pattern Recognition Re-Test allows the participant to review the instructions and complete a practice session before beginning the test. The proposed revisions to the Word Memory Test in Section 8.1 of this report and the Time Wall/Pattern Recognition Test in Section 7.1 of this report also apply to these versions of the tests in the Ending Block.

SUMMARY OF THE FEEDBACK ON THE AT-SAT PILOT TEST BATTERY

This section of the report summarizes the feedback on all the test blocks within the AT-SAT Pilot Test Battery. Overall, we are recommending relatively few changes to the entire battery of tests. The majority of the recommended changes are intended to enhance the clarity of test instructions, increase the value of the test practice sessions, and revise some of the questions for the ability level of the participants. Exhibit 3.2.3, on the following page, displays a summary of the proposed revisions to the pilot test software.

The information provided by the Test Administrators was one of the information sources used to revise the test battery. A significant effort on the part of the project team went into revising the instructions for the tests and the other changes recommended by the Test Administrators. The next section discusses the psychometric information used to revise the battery. Both sources of information provided the test developers the information necessary to build the Beta Battery, which was used in the concurrent validation study.

CHAPTER 3.3

Analysis and Revisions of the AT-SAT Pilot Test

Douglas Quartetti and Gordon Waugh, HumRRO
Jamen G. Graves, Norman M. Abrahams, and William Kieckhaefer, RGI, Inc
Janis Houston, PDRI, Inc
Lauress Wise, HumRRO

This chapter outlines the rationale used in revising the tests and is based on the pilot test data gathered prior to the validation study. A full description of the samples used in the pilot study can be found in Chapter 3.2. It is important to note that some of the tests were developed specifically for use in the AT-SAT validation study, and therefore it was imperative that they be pilot-tested for length, difficulty, and clarity. There were two levels of analysis performed on the pilot test data. First, logic and rationale were developed for the elimination of data from further consideration in the analyses. After the were elimination process, an item analysis of each test was used to determine the revisions to tests and items that were needed.

Exclusionary decision rules were based on available information, which varied from test to test. For example, in some instances, item latency (time) information was available as the appropriate method for exclusion; in other cases, the timing of the tests were computer driven and other criteria for exclusion were developed. An item was considered a candidate for deletion if it exhibited any of the following characteristics:

- ***Low Discrimination:*** The item did not discriminate between those individuals who received high versus low total scores, stated as a biserial correlation.
- ***Check Option:*** One or more incorrect response options had positive biserial correlations with total test score.
- ***Too Hard:*** The percent correct was low.
- ***Too Easy:*** The percent correct was high.
- ***High Omits:*** The item was skipped or not reached, with these two problems being distinguishable from each other.

Applied Math Test

Case Elimination

To determine reasonable average and total latencies for the items attempted, the original sample of 435 was restricted to those individuals who completed all 53 items (N=392) of the Applied Math test (AM). Examining the average latency in seconds for the items revealed a mean time of 14.7 and a standard deviation of 10. After review of the actual test items, it was decided that any individual spending less than 4 per item was probably responding randomly or inappropriately. Review of a scatter plot of average latency by percentage correct revealed that those individuals taking less than 5 scored at the extreme low end, about half scoring below chance. To corroborate this information, a comparison of scores on the Applied Math test and scores on the ASVAB Arithmetic Reasoning test (AS_AR) identified individuals who had the mathematical ability but were not motivated to perform on the Applied Math test (i.e., high ASVAB score but low AM score).

Based on this information, three guidelines for eliminating individuals were formulated:

(1) High Omits: It was determined that any individual attempting fewer than 35 items AND receiving a percent correct score of less than 39 percent was not making a valid effort on this test.
(2) Random Responders: After reviewing and comparing the percent correct scores for the Applied Math test and the AS_AR scores, it was determined that any individual whose AS_AR was greater than 63, but whose percent correct was less than 23%, was not putting forth an honest effort on this test.
(3) Individuals whose average latency was less than 4 were excluded from further item analysis.

Application of these exclusion rules further reduced the sample size to 358 for the item analysis.

Item Analysis

On the Applied Math test, all of the items that were characterized as High Omits were items that the participants did not reach because of test length, not items that they merely skipped. Additionally, this test has four response options with each item, and therefore the chance level of a correct response is 25%.

After review of the item analysis, 18 items were deleted, reducing the test length to 30 items. The item analysis printout for the deleted items can be found in Appendix A. An extensive review of the items by content and computation type was conducted to ensure fair representation of relevant item types. The item types represented were Computing Distances, Computing Travel Time, Computation Given Multiple Point Distances, Computing Ascending/Descending Rates, and Computing Speed.

Summary and Recommendations

The test was shortened from 53 items to 30. Textual changes were made to four items for clarification. The items were re-ordered with the five easiest items first, then the rest of the 30 items randomly distributed throughout the test. This ensured that the test taker would reach at least some of the most difficult items.

Dials Test

Case Elimination

For the Dials test, 406 of the 449 participants completed the entire test. A scatter plot of average latency per item in seconds by percent correct for attempted items was created for the reduced sample (406). The mean and standard deviation for average latency were 12.47 and 4.11, respectively. The mean and standard deviation for percentage correct were 78.89% and 12.96%, respectively. A grid overlay based on the means and standard deviations revealed that individuals who were more than two standard deviations below the mean for average latency (4.25 per item) were scoring more than two standard deviations below the mean for percent correct (52.97%). It appears that these individuals were not taking the time to read the items or put forth their best effort. Following an exclusion rule of eliminating participants who had an average latency per item of 4.25 or less, the sample was reduced from 449 to 441.

Item Analysis

After review of the item analysis and of specific items, 13 items were deleted from the original test. All had low discrimination and/or another response option that was chosen more frequently than the correct response. In many instances, the graphics made it difficult to discriminate between correct and incorrect dial readings. The revised test consists of 44 items. The item analysis printout for the deleted items can be found in Appendix A.

Summary and Recommendations

The 13 items that had low discrimination or response options that were chosen more often than the correct response were eliminated, reducing the test length to 44 items. An additional recommendation was that 17 - inch display monitors be used in the beta version to ensure the integrity of the graphics.

Angles Test

Case Elimination

For the Angles test, all 445 individuals completed the entire test (30 items). A scatter plot was created of the average latency per item in seconds by the percent correct for attempted items. The mean and standard deviation for average latency were 8.2 and 2.67, respectively. The mean and standard deviation for percent correct were 67.73% and 17.7%, respectively. A grid based on the means and standard deviations of each axis revealed that, of the four individuals who were more than two standard deviations below the mean for average latency (2.86 per item), three scored more than two standard deviations below the mean for percentage correct (32.33%). The other individual was about 1.5 standard deviations below the mean for percent correct. It appears that these individuals were not taking the time to read the items or put forth their best effort. By eliminating those individuals with an average item latency of less than 2.86, the item analysis sample was reduced to 441.

Item Analysis

The item analysis did not reveal any problem items and there appeared to be a good distribution of item difficulties. No text changes were indicated. After reviewing the item analysis and the items in the test, none of the items were deleted.

Summary and Recommendations

This test appears to function as it was intended. There were no item deletions and no textual changes.

Sound Test

Case Elimination

On the Sound test, 437 participants completed 17 or 18 items. Of the remaining five participants, one completed only two items (got none correct) and was deleted from the sample. The other four participants made it to the fourth set of numbers (8 digits). All the scores of this group of four were within one standard deviation (15%)

of the mean of the percentage correct for attempted items (35.6%). Additionally, five other participants did not get any items correct. It was determined that two of them were not trying, and they were deleted. The remaining three seemed to "be in the ballpark" with their responses (i.e., many of their responses were almost correct). With the exclusion of three participants, the total sample for the item analysis was 439.

Alternative Scoring Procedure

The Sound test consists of numbers of set lengths (5, 6, 7, 8, 9, 10 digits) being read and then participants recalling them. There are three trials associated with each number length (i.e., number length 5 has three trials, number length 6 has three trials, etc.) for a total of 18 items. Examinees receive a point for every item answered correctly. An alternative scoring procedure would be to award a point for each digit they get correct and one point for a digit reversal error. For example, in the 5-digit case, a correct response may be 12345, but a participant may answer 12354 (digit reversal). In this case, the participant would receive 3 points for the first three digits and 1 point for the digit reversal, for a total of 4 points on that trial. This scoring procedure was examined as an alternative to the number correct score.

Item Analysis

After review of the item analysis, none of the items were removed. However, the biserial correlations of the items from digit length 5 and digit length 10 were appreciably lower than the rest of the items. The reliability of this test with the original scoring procedure was .70, while the alternative scoring procedure improved reliability to .77. Using the alternative scoring procedure, in a comparison of the original version and a revised version with digit length 5 and digit length 10 removed, the revised version had a slightly higher reliability (.78).

Summary and Recommendations

Since digit lengths of 5 and 10 had lower biserial correlations than the rest of the items, it was recommended that the number of trials associated with these items be reduced to two each. The alternative scoring procedure, based on the number of within-item digits correct with partial credit for digit reversals, was recommended for the beta version.

Memory Test

Case Elimination

A scatter plot of Memory test items answered by percent correct revealed a sharp decline in the percent correct when participants answered fewer than 14 items. It was decided that participants who answered fewer than 14 items were not making an honest effort on this test. Additionally, it was felt that participants who scored less than 5% correct (about 1 of 24 correct) probably did not put forth their best effort, and therefore, they were removed from the item analyses. These two criteria eliminated 14 participants, leaving a sample of 435 for the item analyses.

Item Analysis

After review of the item analysis, none of the items were removed. Item 1 had low discrimination, low percent correct, and a high number of omits. However, there were no such problems with the remaining items, and given that these are non-sense syllables, one explanation may attribute the poor results to first-item nervousness-acclimation. All items were retained for the beta version, and no editorial changes were made.

Summary and Recommendations

This test performed as expected and had a normal distribution of scores. One item had problem characteristics, but a likely explanation may be that it was the first item on the test. The recommendation was to leave all 24 items as they were but to re-examine the suspect item after beta testing. If the beta test revealed a similar pattern, then the item should be examined more closely.

Analogy Test

Case Elimination

For the Analogy test, cases were eliminated based on three criteria: missing data, pattern responding, and apparent lack of participant motivation.

Missing Data. The test software did not permit participants to skip items in this test, but several (12.8%) did not complete the test in the allotted time, resulting in missing data for these cases. Those missing 20% or more of the data (i.e., cases missing data for 11 items or more) were omitted. Five cases were eliminated from the sample.

Pattern Responders: The chance level of responding for this test was 20%. An examination of those participants near chance performance revealed one case where the responses appeared to be patterned or inappropriate.

Unmotivated Participants: Identifying participants who appeared to be unmotivated was based on the average latency per item, which was 5.4. It was determined, because of the complexity of the items, that participants spending 5.4 or less were not taking the test seriously or were randomly responding, and therefore were eliminated from the item analyses. As a cross check, an examination of the percentage correct for those participants whose average latency was 5.4 seconds or less showed that their scores were near chance levels. Four participants were eliminated.

In summary, 10 participants were eliminated from further analyses, reducing the sample size from 449 to 439.

Scale and Item Analyses

An examination of the biserial correlations for the 53 items revealed 12 items that had biserial correlations of .10 or less. This reduced the number of items within three of the four test scales as follows (the original number of items appears in parentheses): Non-Semantic Words 9 (15), Semantic Words 12 (15), and Semantic Visuals 7 (10). Tables 3.3.1 to 3.3.4 present these corrected item-total correlations and the alphas for the items within each scale. As Table 3.3.4 indicates, all 13 items within the Non-Semantic Visual scale met the criterion for retention. After omitting items based on the above criteria, the number of items in this test dropped from 53 to 41.

Construct Validity

A multitrait-multimethod matrix was constructed to assess whether the information processing scores and the number-correct scores measure different constructs or traits. Test scores based on two traits (i.e., Information Processing and Reasoning) and four methods (i.e., Word Semantic, Word Non-Semantic, Visual Semantic, and Visual Non-Semantic) were examined. The results provided the following median correlations:

- The median convergent validity (i.e., same trait, different method) for information processing scores was .49.
- The median convergent validity for number-correct scores was .34,
- The median divergent validity (i.e., different trait, different method) was .18.

These preliminary results suggest keeping separate the information-processing and number-correct scores for the Analogy test, pending further investigation.

Testing Time

Based on the sample of 439 participants, 95% of the participants completed the test and instructions in 33 minutes (Table 3.3.5). Table 3.3.6 shows time estimates for two different levels of reliability.

Test Revisions

A content analysis of the test revealed four possible combinations of semantic/non-semantic and word/visual item types. The types of relationships between the word items could be (a) semantic (word-semantic), (b) based on a combination of specific letters (word - non-semantic), (c) phonetic (word - non-semantic), and (d) based on the number of syllables (word - non-semantic). The types of relationships between visual items could be based on (a) object behavior (visual-semantic), (b) object function (visual-semantic), (c) object feature (visual-semantic, (d) adding/deleting parts of the figures (visual - non-semantic), (e) moving parts of the figures (visual - non-semantic), and (f) rotating the figures (visual - non-semantic).

After categorizing the items based on item type, an examination of the item difficulty level, item-total correlations, the zero-order intercorrelations between all items, and the actual item content revealed only one perceptible pattern. Six non-semantic word items were removed due to low item-total correlations, five being *syllable* items (i.e., the correct solution to the analogy was based on number of syllables).

Seven more items were removed from the alpha Analogy test version due to either very high or low difficulty level, or to having poor distractor items.

Word Items. The time allocated to the Analogy test items remained approximately the same (35 minutes and 10 minutes for reading instructions) from the alpha version to the beta version. The number of word items did not increase; however, nine items were replaced with items that had similar characteristics of other well-performing word items. There were equal numbers of semantic and non-semantic items (15 items each).

Since the analogy items based on the number of syllables performed poorly, this type of item was not used when replacing the non-semantic word items. Instead, the five non-semantic word items were replaced with *combinations of specific letters* and *phonetic* items. Additionally, three semantic items were replaced with three new semantic items of more reasonable (expected) difficulty levels.

Visual Items. Since the non-semantic picture items demonstrated a relatively stable alpha (.67) and high item-total correlations, no items were removed. In an effort to stabilize the alpha further, three non-semantic picture items were added, increasing the non-semantic visual subtest from 13 to 16 items.

One item was dropped because it contained poor distractors. Two other semantic visual items that appeared to have poor distractors were modified to improve the clarity of the items (without lowering the difficulty level). In addition, two newly created items were added to this scale. Thus, one item was replaced with two new items, and two others were modified.

Instructional Changes. Based on feedback from site Test Administrators, portions of the Analogy test instructions were simplified to reduce the required reading level of the text. Also, the response mode was changed from use of a keyboard to use of a mouse. The Viewing an Item section of the instructions was revised accordingly.

Summary and Recommendations

The Analogy test assesses inductive reasoning and information processing abilities in four areas: Non-Semantic Word, Semantic Word, Non-Semantic Visual, and Semantic Visual. The number-correct scores that reflected reasoning ability proved less reliable than the information processing scores. Of the 53 items in the alpha battery, 41 contributed sufficiently to test reliability to warrant inclusion in the revised version. It was estimated that to achieve a reliability level of 0.80 it would be necessary to increase the test length to 150 items. Given time limits in the validation version, the overall test length was limited to 57 items.

Classification Test

Case Elimination Analyses

Cases in the Classification test were eliminated based on three criteria: missing data, pattern responding, and apparent lack of participant motivation.

Missing Data. As with the Analogy test, the Classification test software did not permit participants to skip items. However, some participants (7.5%) did not complete the test in the allotted time, resulting in missing data. Of these cases, those missing 20% or more of the data (i.e., cases missing data for nine items or more) were omitted. A total of 10 cases were eliminated.

Pattern Responding. From examination of the pattern of responses of participants who scored at or near chance levels (20%), eight participants were identified as responding randomly and were eliminated.

Unmotivated Participants. It was decided that participants spending less than 3 per item were not making a serious effort. Four participants fell into this category. An examination of their total scores revealed that they scored at or near chance levels, and thus they were eliminated from further analyses.

In summary, 22 participants were eliminated from further analyses, reducing the sample size for this test from 449 to 427.

Scale Reliabilities and Item Analyses

Reliability analyses were conducted to identify the items within each of the four test scales that did not contribute to the internal consistency of that scale. The corrected item-total correlation was computed for each item within a scale, as well as the overall alpha for that scale.

An examination of the item-total correlations revealed that the Non-Semantic Word scale items had an average correlation of .179, and therefore the entire scale was omitted from further analyses. This reduced the number of items within the three remaining test scales as follows (the original number of items appears in parentheses): Semantic Word 9 (11), Non-Semantic Visual 10 (13), and Semantic Visual 3 (10). Note that the greatest number of items were removed from the semantic visual scale. Tables 3.3.7 to 3.3.10 present the corrected item-total correlations for the items within each scale. After omitting items based on the above criteria, the number of items in this test was reduced from 46 to 22.

Construct Validity

In assessing the construct validity of the information processing measures independent of the number correct scores, a multitrait-multimethod matrix was constructed. Two traits (i.e., information processing and reasoning) and four methods (i.e., Word Semantic, Word Non-Semantic, Visual Semantic, and Visual Non-Semantic) were examined. The results of this analysis provided the following median correlations:

- The median convergent validity (i.e., same trait, different method) for information processing scores was .48.
- The median convergent validity for number-correct scores was .20.

· The median divergent validity (i.e., different trait, different method) was .09.

These preliminary results suggested a separation of the information-processing and number-correct scores for the Classification test, pending further investigation.

Time Limit Analyses

Based on the sample of 427 participants, 95% of the participants completed the instructions and the 46 test items in 22 minutes (Tables 3.3.11). Table 3.3.12 shows estimates of test time limits assuming two different levels of reliability and test lengths for the three test parts. These estimates assume keeping all aspects of the test the same (i.e., all four classification schemes).

Summary and Recommendations

Of the original 46 items, only three of the four scales (i.e., Semantic Word, Semantic Visual, and Non-Semantic Visual) and a total of 22 items contributed sufficiently to test reliability to warrant inclusion in a revised test version. To construct a test having the same three parts and increase the reliability to about .80 (for number-correct scores), the number of items would need to increase from the 22 to 139. It was further found that the Classification test correlates highly with the Analogy test. Given that the Classification test had lower reliability scores than the Analogy test, it was recommended that the Classification test be eliminated from the AT-SAT battery.

Letter Factory Test

Analysis of Initial LFT

Case Elimination. Two criteria were used in eliminating Letter Factory Test participants: apparent lack of participant motivation and inappropriate responding.

Unmotivated Participants. Unmotivated participants were considered to be those who responded to very few or none of the items. An examination of performance on the number correct across all Planning/Thinking (P/T) items in the test sequences (Table 3.3.13) reveals a gap in the distribution at 28% correct. It was decided that participants scoring below 28% were not making a serious effort on this test, and they were eliminated from further analysis.

Inappropriate Responding. Inappropriate responders were identified as participants who either selected a box of the wrong color or selected boxes when none were needed. During the test, there were 86 times when a participant should have placed a box in the loading area. The computer software recorded the number of times a participant tried to place a box incorrectly (i.e., to place a box when one was not needed or to place an incorrectly colored box). This measure serves as an indicator of inappropriate responding. Table 3.3.14 shows the distribution of the number of unnecessary attempts to place a box in the loading area across the entire sample of 441 cases.

Several participants had a very high number of these erroneous mouse clicks. There are two possible reasons for this. Feedback from the Test Administrators indicated that some participants were double-clicking the mouse button, instead of clicking once, in order to perform LFT test tasks. Every instance a participant erroneously clicks the mouse button is recorded and compiled by the computer to generate an inappropriate response score. Thus, if a participant orders the correct number of boxes (86) by double-clicking instead of single-clicking, 86 points will be added to his or her inappropriate response score.

A few participants had a random-response score higher than 86. These participants may have developed a strategy to increase their test score. The test instructions explained the importance of placing a box in the loading area as soon as one was needed. This may have caused some participants to continuously and consistently attempt to place boxes in the loading area. Participants received an error signal each time they unnecessarily attempted to place a box; however, they may not have realized the negative impact of this error on their test score.

Cases were eliminated where the inappropriate response score was higher than 86. This allowed using the data from participants who were motivated but might have misunderstood the proper way to use a mouse during this test. However, to prevent an inappropriate response strategy from interfering with a true measure of Planning/Thinking Ahead (P/T), information from the inappropriate response variable must be used when calculating a participant's P/T test score. Omitting cases based on the above criteria reduced the sample from 441 to 405.

Item Analyses

Recall From Interruption (RI). The proposed measure of RI was a difference score between a participant's number-correct score across a set of items presented immediately after an interruption and number-correct score across a set of items presented just before the interruption. Four of the test sequences (sequences 4, 6,

8, and 11) contained RI items. Table 3.3.15 shows the number of items within each sequence that make up each pre-interruption and post-interruption scale score, as well as the score means and standard deviations. The mean scores for each sequence are very high, indicating a ceiling effect. However, increasing the difficulty of the RI items would require increasing either the number of letters on belts or the belt speed. Either of these methods would alter the task so that psychomotor ability would become a very significant component of the task. Therefore, it was concluded that this task is not suited to providing a measure of RI.

Table 3.3.15 also provides a summary of the reliability of the pre-interruption, post-interruption, and difference scores. Notice that the reliability of the pre- and post-interruption scores (Alpha = .79 and .73, respectively) is much higher than the reliability of the difference scores (Alpha = .10). Plans for the recall from interruption score were abandoned due to low reliability.

Planning/Thinking Ahead. To prevent an inappropriate response strategy from interfering with a true measure of P/T, the inappropriate response score must be used in calculating a participant's P/T test score. The test design prevented the association of unnecessary mouse clicking with a specific P/T item. (Participants do not have to respond to P/T test items in the order in which they receive them; instead, they may wait and then make several consecutive mouse clicks to place multiple boxes in the loading area.) However, the software recorded the number of times each participant inappropriately attempted to place a box during a test sequence. Also, the number of unnecessary mouse clicks has a low but significant negative correlation ($r = -.20$, $p < .001$) with the number-correct scores on a sequence. Therefore, the P/T scale score per sequence was computed by subtracting the number of unnecessary mouse clicks in a sequence from the number-correct score across all P/T items in that sequence.

Table 3.3.16 summarizes findings from the reliability analysis on the seven sequences designed to measure P/T. The second column indicates the number of P/T items involved in calculating the P/T sequence scores. Notice that the sequence-total correlations are .60 or higher. Therefore, none of the P/T sequences were deleted. The alpha computed on the seven sequences was .86.

Situational Awareness (SA). As noted earlier, the Letter Factory Test contained multiple-choice questions designed to measure three levels of SA. Fourteen of these items were designed to measure SA Level 1, 16 items to measure SA Level 2, and 14 items to measure SA Level 3. Table 3.3.17 summarizes findings from the analyses on the items within the scale for each of these three levels. If an item revealed a corrected item-total correlation (column 3) of less than .10, it was removed from the scale. This reduced the number of Level 1, Level 2, and Level 3 items to 8, 11, and 7, respectively. A reliability analysis on the remaining SA items showed alphas on the three new scales of .42, .61, and .47.

Next, a scale score was computed for each of the three SA levels. These scores were used in a reliability analysis to determine whether the three scales were independent or could be combined into one scale that measured the general construct of SA. The alpha computed on the three scale scores was .53. The results indicated that removal of any one of the three scale scores would not increase the alpha. These results supported the notion that all remaining SA items should be combined into one scale.

Table 3.3.18 presents findings from a reliability analysis on the 26 remaining SA items scored as one scale. The alpha computed on this overall scale was .68. The corrected item-total correlations computed in the reliability analysis on the three separate SA scales (Table 3.3.17, "After Item Deletion") were very similar to the corresponding corrected item-total correlations computed for the combined scale (Table 3.3.18). This also supports the notion of using a single number-correct score across all remaining SA items.

Analysis of LFT Retest

The LFT Form B or retest contains five test sequences that are parallel to the following sequences in the LFT Form A: 3, 6, 7, 10, and 11. Form B contains 37 P/T items, 54 RI items (27 pre-interruption items and 27 post-interruption items), and 20 SA items.

Case Elimination

First, we identified participants who had been classed as unmotivated or random responders while taking Form A of the LFT. Twenty-three of those participants also received Form B of the LFT. Since Form B was designed to serve as a retest, we eliminated the Form B 23 cases that had been eliminated from Form A.

Next, the same criteria used in Form A were used to eliminate unmotivated participants and inappropriate responders in Form B. To look for unmotivated participants, we considered performance on the number correct across all P/T items in the test sequences. Table

3.3.19 provides an overview of the distribution of those number-correct scores across the entire sample of 217 cases.

A natural gap was evident in the distribution and cases where the number-correct score was lower than 30 were eliminated. Then, participants who were responding inappropriately to items were identified. During the test, there were 37 times when a participant should have placed a box in the loading area. Table 3.3.20 provides an overview of the distribution of the number of inappropriate attempts to place a box in the loading area across the entire sample of 217 cases. Cases where the inappropriate response score was higher than 37 were eliminated. After omitting cases based on the above criteria, the sample size for this test was reduced from 217 to 184.

Item Elimination

Again, since Form B was designed to serve as a retest, the findings from analyses performed on LFT Form A were used to determine which test items to eliminate from Form B. We removed 8 SA items so 12 SA items remain in Form B. Similarly, a P/T score was computed for Form B by subtracting the number of unnecessary mouse clicks from the number-correct score across all P/T items.

Performance Differences

Form B was used to assess whether participants had reached an asymptote of performance during Form A. Different sequences could not be used in Form A for this test because the item types are very heterogeneous, and little information is available on item or sequence difficulties. By matching test sequences, we can control for manageable aspects of the test that impact test performance. Table 3.3.21 presents the results of dependent *t*-tests comparing two performance measures. Those results show no support for a change in participants' performance on Situational Awareness. However, the roughly 8% performance increment on Planning and Thinking Ahead was a significant increase in performance. This suggests that participants would benefit from more practice before beginning the test.

Time Limit Analyses

Table 3.3.22 presents the distribution of test completion times for the LFT, showing that 95% of participants completed the LFT in 64.9 minutes or less. When we use the normal curve to compute the 95th percentile (i.e., take 1.96 times the standard deviation (7.63) and add that product to the mean), we estimate a slightly higher amount of time (66.7 minutes) for the 95th percentile participant. A test completion time of 67 minutes, then, seems appropriate for the test at its current length. Of this, 95% of participants completed the LFT test sequences in 27.1 minutes. This leaves about 39.9 minutes for instructions and practice.

The Spearman-Brown formula was used to estimate the number of items needed to raise the reliability of Situational Awareness to .80 (49 items) and .90 (110 items). Because the measure of Planning and Thinking Ahead already has a higher estimated reliability of .86, it would automatically go up to a sufficient level when the number of sequences is raised to increase the reliability of Situational Awareness. Table 3.3.23 presents a recommended composition of sequence lengths and number of Situational Awareness test items per sequence. It was estimated that this recommended composition would yield a test reliability in the low .90s for Planning and Thinking Ahead, and a reliability in the low .80s for Situational Awareness. Based on experience with the alpha LFT, it was estimated that participants would spend 45 minutes on the test portion of the new test version.

The amount of practice before the test also needed to be increased. The initial practice sequence was an easy 30-second sequence, followed by sequences of 2 minutes and 2.25 minutes. Adding three more sequences of the same difficulty as the test, together with four SA questions after each sequence, was proposed. One 30-second sequence and two 2.5-minute sequences would add an additional 7.5 minutes to the practice and instruction time. Improving the instructions to emphasize the penalty that occurs for error clicks on the Planning/Thinking Ahead measure would add about half a minute. Instruction and practice time, then, should increase by 8 minutes from 39.9 to about 48 minutes. With a 45-minute time for the test sequences, this amounted to 93 minutes for the recommended beta version of the LFT.

Test Revisions

Test Sequences. To reduce the number of inappropriate responses by participants who double-click or continuously click on the box stack, the associated error signal (one red arrow) was changed to an error message. The error message appears in red above the box stack when participants try to move a box to the loading area when one is not needed. The new error message reads, "You did not need to move a box."

To increase visibility, the error signal was also changed to an error message when participants try to place a letter in a box that is not the fullest. This error message reads, "You did not place the letter in the fullest box."

Analyses described above showed that the mean scores for RI items had a ceiling effect. Also, it was indicated earlier that the difficulty level of the RI items could not be increased without making psychomotor ability a significant component of the task. For these reasons, all RI items were removed from the test.

To increase test reliability for the P/T and SA measures, the number of test sequences and SA questions was increased. Test sequences were increased from 11 to 18. Level 1 SA questions were increased from 14 to 26, Level 2 SA questions from 16 to 24, and Level 3 SA questions were increased from 14 to 26. SA questions also were revised.

Test Instructions. Based on feedback from the project team and the FAA, several sections of the instructions were changed. The pilot version of the test included several short practice sequences in the middle of the instructions and three complete practice sequences at the end of the instructions but before the test. To increase the amount of practice, two practice sequences were added in the middle of the instructions, allowing participants to practice moving boxes to the loading area and placing letters into boxes. In addition, the mouse exercise appears before the practice sequences so participants can learn how to use the mouse before they practice taking the test. In addition, the mouse exercise was changed so participants can choose to receive additional practice using the mouse (up to nine trials), or move to the next set of instructions.

Other changes in response to project team and FAA feedback include (a) not showing the mouse cursor (arrow) on the screens in which the mouse is disabled; (b) adding a screen that identifies the six items/areas that make up the test display; (c) simplifying and increasing the speed of some of the examples; (d) changing the "Call Quality Control" screen label to "Quality Control"; (e) and simplifying some words and sentences in portions of the instructional text.

Changes were also made in parts of the instructions in response to the data analyses. To reduce the number of inappropriate responses due to double-clicking or constant attempts to place boxes in the loading area, instructions were added telling participants to click on the mouse only once and to not double-click. Corrective feedback also was added to the practice sequences that appear in the middle of the instructions. For these practice sequences, if a participant clicks more than once to move a box to the loading area, the screen freezes and the following corrective feedback appears: "To move a box to the loading area, you only need to click on the box once. Do not double-click." The following corrective feedback appears when a participant does not place a box in the loading area when one is needed: "The computer moved a box into the loading area for you. Keep track of new letters that appear on the conveyor belt and move boxes to the loading area as soon as they are needed." If the computer places a letter into a box for the participant, the following corrective feedback appears: "A letter fell off a conveyor belt and the computer placed it in a box. Make sure you track the letters as they move down the belts." The last corrective feedback appears when a participant tries to place a letter in a box that is not closest to being full: "You need to place the letters in the boxes that are closest to being full."

Finally, since the RI measure had been eliminated from the test, any reference to RI was removed from the instructions. Also, in view of the significant increase in participant performance on the P/T measure on the retest, three practice sequences were added at the end of the instructions. This provides participants with a total of six practice sequences at the end of the instructions.

Situational Awareness Items. SA items at the end of each LFT sequence were revised on the basis of the item analyses described above. The following subsections below provide an overview of how item analyses guided development of new SA items.

For all levels of SA, care was taken to not ask any questions about letters below availability lines on the belts. The reason was that examinees who worked quickly would have completed tasks associated with letters below availability lines; for such examinees, those letters might not even be on the belts, but rather in boxes. For all examinees, though, the letters above the belts would all be in the same places.

Level 1 Items. Level 1 Situational Awareness items assess an examinee's ability to perceive elements in the environment, together with their status, attributes, and dynamics. The following are examples of item stems (i.e., the questions) that item analysis supported:

- Which belt moved letters the SLOWEST?
- Which belt had the most letters ABOVE the availability line?
- Which two belts have their availability lines closest to the BOTTOM of the belt?

· How many ORANGE letters were ABOVE the availability lines?
· Which belt had no letters on it?

Listed below are examples of item stems that item analysis did NOT support:

· Which belt moved letters the FASTEST?
· Which belt had its availability line CLOSEST to the TOP of the belt?
· Which letter was CLOSEST to the TOP of the belt?
· Which letter was CLOSEST to crossing the availability line?

Item analysis, therefore, suggested that Level 1 SA is more reliably measured in the LFT by asking about the number and color of letters above the availability lines, which belt was slowest, and which belts had their availability lines closest to the bottom. These questions are consistent with targeting items toward a lower average level of planning and thinking ahead. Generally, the items that did not contribute well to scale reliability included those about the fastest belt, availability lines closer to the top of the belt, and letters closer to the top of the belt. Those items are more consistent with a higher level of planning and thinking ahead. Therefore, development of new Level 1 SA items was focused on the types of areas listed above that required lower amounts of planning and thinking ahead.

Level 2 Items. Level 2 Situational Awareness items assess an examinee's ability to comprehend the situation. This requires a synthesis of disjointed Level 1 elements. In asking about Level 2 SA, it is important to assess comprehension of Level 1 elements while requiring no projection of the future status (a Level 3 factor). In the LFT setting, it was also considered necessary to clearly distinguish between letters already in boxes in the loading areas, compared with letters that examinees could place there. To make this distinction clear, short (i.e., 30-second) scenarios where no letters ever crossed any availability lines were developed. That way, no letters were in boxes at the end of these short scenarios. In these scenarios, examinees could only place boxes and maintain their situational awareness.

The following are examples of Level 2 item stems (i.e., the questions) that item analysis supported:

· What color was the LAST box you should have placed in the loading area in order to correctly place all the letters into boxes?
· Consider the LAST box you should have placed in the loading area. Which letter caused you to need to place this last box?
· How many (or which) boxes should be in the loading area in order to correctly place all the letters?
· Consider all the letters on the belts. What color were the letters that, when combined, could fill at least one box?

Listed below are examples of item stems that item analysis did NOT support:

· If all the letters were correctly placed in boxes, how many empty spaces for Ds would be in the boxes in the loading area?
· If all letters were correctly placed in boxes, how many more (or which) letters would be needed to completely fill the GREEN box?

Item analysis, therefore, suggested that Level 2 SA is more reliably measured in the LFT by asking about the last box an examinee should have placed in the loading area, the number or color of boxes that should be in the loading area, and what color of letters could completely fill a box. These questions are consistent with targeting items toward the more immediate LFT concerns. Generally, the items that did not contribute well to Level 2 scale reliability included those about the number of empty spaces for a particular letter, and how many more or which letters were required to fill a particular color of box. Those items are more consistent with a more fine-grained level of situational awareness. Therefore, development of new Level 2 SA items was focused on the types of areas listed above that required only an awareness of the more immediate LFT concerns.

Level 3 Items. Level 3 SA items assess an examinee's ability to project the future status or actions of the elements in the environment. This requires a knowledge of the status and dynamics of the elements, as well as a comprehension of the situation—both Level 1 and Level 2 SA. In asking about Level 3 SA for LFT sequences, it is important to ensure that all examinees are responding to the same situation. Although the software stops at exactly the same place for all examinees, the quicker examinees may have placed more boxes in the loading area or placed more letters in boxes. For this reason, all Level 3 SA items were started with the following two sentences separated as a paragraph before the Level 3 question:

"Assume that you correctly placed all required boxes in the loading area. Also assume that you correctly placed all the letters that remained on the belts."

The intent of this introduction was to allow all examinees to (at least mentally) begin from the same situation. Item analysis showed that Level 3 SA is more reliably measured in the LFT by asking about simple situations. In the case of the LFT, that seemed to be a situation having only two or three boxes in the loading area. The following are examples of Level 3 item stems (i.e., the questions) that item analysis supported for those simple situations:

· After the full boxes are removed, which boxes would remain in the loading area?
· Which letters would you need to complete an ORANGE box?
· How many more letters would you need to fill all the boxes remaining in the loading area?
· If the next letter was a PURPLE A, which of the following would be true?

Therefore, development of new Level 3 SA items was focused on simple situations and used the types of questions listed above.

Summary and Recommendations

The original plan was to measure three worker requirements using the LFT. Because the measure of Recall From Interruption showed ceiling effects and unreliable difference scores, it was recommended that attempts to measure that worker requirement with this test be abandoned. To more adequately measure the worker requirements of Planning and Thinking Ahead and Situational Awareness, lengthening the test to 93 minutes was recommended. This longer version includes doubling the number of practice sequences that participants complete before they begin the test. It was estimated that this extra practice would reduce the practice effect observed between the LFT and the retest LFT on a small (N = 184) subsample. This would help ensure that participants perform at or near their ability prior to beginning the test portion of the LFT.

Scan Test

Data Collection/Software Problems

As the data collection proceeded on the Scan test, it became clear that the software was not writing data for change items nor was it recording item latencies. A correction to the software was implemented on February 26. Of the 429 cases on which data were collected, 151 cases had complete data.

Case Elimination

Because all participants proceed at the same rate during practice and test sequences, test completion time could not be used to assess participants' test-taking motivation. Likewise, because the test software automatically writes out data for each item indicating whether the participant correctly selected the item, no cases should have missing data.

Unmotivated Participants. It was believed that unmotivated participants would respond to very few or none of the items, or respond with irrelevant answers. The number-correct scores were used to identify unmotivated participants. The distribution of the 429 participants is provided in Table 3.3.24. An examination of the data showed that no participant simply sat at the computer and allowed the software to progress on its own. Each participant entered some appropriate responses, and each got at least a few items correct. The lowest score shown was 22 out of the 162 questions correct (13.6%). While there may have been participants who were not trying their best, this screening algorithm was unable to identify participants who blatantly did nothing at all. Therefore, all cases were kept for the analyses.

Item Analyses

Table 3.3.25 presents findings from the reliability analysis on the four test sequences (i.e., T1 to T4). The three parts of the table show how the sequence reliabilities measured by alpha differed as different groups of items were deleted. The first part ("With Change Items") presents results that include all the items in each sequence. Each change item may be considered as two items; the item is what was presented originally, and the second is the item with the change in the bottom or three-digit number. The middle columns include the pre-change items and exclude the post-change items, and the third part of the table removes both versions of the change items (i.e., the original and the change part). Notice, too, that the second and third parts of the table show "Actual" and "Expected" alphas. The actual alphas are the results provided by the data. The expected alphas are the ones estimated by the Spearman-Brown formula if "like" items were deleted. In every case, the alphas from the data are higher than the expected alphas. This finding supports the notion that the change items

differ from the other items in the test. Not including them in the scoring, therefore, should increase the alpha test reliability.

Of the 166 remaining items in sequences T1 to T4, only four items (i.e., items 359, 373, 376, and 410) had item-total correlations less than .10. The alpha computed on the 162 items remaining was .959. This supported computing a number correct score using these 162 items for the scanning worker requirement.

Time Limit Analyses

Table 3.3.26 shows the distribution of test times for participants in this sample; 95% completed the Scan Test in 19.87 minutes or less. If we take 1.96 times the standard deviation of test completion times (1.75) and add that product to the mean test completion time (16.92), we find that a 95th percentile participant might take 20.35 minutes to complete the Scan Test. Due to the obtained test reliability, it was recommended that no change be made to the test time for the Scan test, with 21 minutes allocated for this test.

Summary and Recommendations

Items in the Scan test that change during their screen presentation did not behave the same as other items in the test. Eliminating those items improved estimates of internal consistency reliability. After eliminating four items that had poor item-total correlations, the 162 remaining items in the test (i.e., non-practice) portion of the Scan test produced an alpha of .96. Therefore, we recommend keeping the Scan test at its current length and allocating 21 minutes for test completion.

Participant performance on the Scan test items is affected by the configuration of other items presented on the same screen so any change must be considered carefully. As a class of items, the change items tended to reduce the Scan test reliability. By eliminating the changing nature of the change items, the test instructions could be simplified. However, eliminating those items might make the test easier or change the test in some other way. Therefore, it was recommended to keep the items as they are presented initially (i.e., without the changing feature) but not count them. A similar recommendation was made for the four items that showed poor item-total correlations.

Planes Test

Case Elimination

The Planes test consisted of three parts and cases were eliminated from each part independently. The screening algorithms for each part were based on similar premises.

Part 1 consisted of 48 items. Participants were eliminated from further analyses if any of three screening criteria were satisfied. The first screen for this part was a total latency less than 48 seconds. The second screen was percent correct less than or equal to 40%. The final screen was the skipping of six or more items. These screening algorithms reduced the sample from 450 to 429 for Part 1.

The screening for Part 2 was similar. Participants were eliminated from further analyses on these criteria: (1). Part 2 total latency less than 1.2 minutes, (2). 40% correct or less, or (3). missing data for six or more items. These screening algorithms reduced the available sample from 450 to 398 for Part 2.

Part 3, participants were eliminated on these criteria: (1) Part 3 total latency less than 2.4 minutes, (2) 40% correct or less, or (3). missing data for 12 or more items. These screening algorithms reduced the available sample from 450 to 366 for Part 3.

Participant elimination across all three test parts left a final sample of 343 having data on all three parts.

Item Analyses

Scale Reliabilities and Item Analyses. Reliability analyses were conducted to identify items within each part of the Planes test that contribute to internal consistency. The corrected item-total correlation was computed for each item within each part as was the overall alpha for that part. Table 3.3.27 presents an overview of the results of these reliability analyses.

The Planes test is not a new test, having been developed previously as the Ships test (Schemmer et al., 1996). In its alpha text form, the number of items was cut in half to meet the time allowed for it in the pretest. In reducing the number of items, the same proportion was kept for all item types. However, there are many parallels between the items in each of the three parts of the test; a particular item that may not work well in Part 1 might work very well in Parts 2 or 3. For these reasons and because data from all three parts were to be used to develop a residual score for the coordinating ability component of multitasking, eliminating items based on poor item-total correlations alone was not considered desirable.

Restoring the Planes test to its original length would require doubling the number of items. Using the Spearman-Brown formula, the new reliabilities are estimated at .86 for Part 1, .91 for Part 2, and .89 for Part 3.

Computing Residual Scores. Using number correct scores from Planes Part 1 and Part 2, the regression procedure outlined in Yee, Hunt, and Pellegrino (1991) was followed to compute an estimate of the coordinating ability component of multitasking. First, the regression equation for predicting the Part 3 score was computed. Then, the difference between the actual and predicted scores for Part 3 was computed by subtracting the predicted from the actual score. This residual score estimates the coordinating ability aspect of multitasking.

Yee et al. argue that a necessary but not sufficient condition for the residual scores to be useful is that they must be reliable. As they indicate, the quantity ($1-R^2$) must be greater than zero *after* allowing for unreliability in the performance measures involved. To show the residual score as reliable, analysts corrected the test scores for each of the three parts of the Planes test for unreliability and created a new correlation matrix. Using this corrected correlation matrix, the multiple correlation was computed to predict Part 3 from Parts 1 and 2 ($R = .506$, $R^2 = .256$). To the extent that this multiple R^2 is less than unity after correcting all performance measures for unreliability, the residual scores may be considered reliable.

In addition, analysts followed the procedure of Yee et al. and compared the multiple R^2 (computed on observed scores, $R^2 = .164$) to the reliability of Part 3 (alpha = .804). Both analyses supported the inference of residual score reliability. Finally, we used the reliabilities of the predicted and actual scores to estimate the reliability of the residual score ($r = .613$). The reliability of the coordinating ability score for a Planes test of twice the length was estimated to be .65.

Time Limit Analyses

Table 3.3.28 shows the distribution of test completion times for the Planes test. Ninety-five per cent of participants completed the Planes test in 34.6 minutes or less. When we take 1.96 times the standard deviation (4.42) and add that product to the mean, we estimate a slightly higher amount of time (36.4 minutes). A test completion time of 37 minutes, then, seems appropriate for the test at its current length. Of this 37 minutes, 95% of participants completed instructions and practice in 21.6 minutes. This leaves 15.4 minutes for completing the 192 items in the three test parts. Doubling the number of items, then, would increase the test time by 15.4 minutes, from 37 to 52 minutes.

Test Revisions

Following the alpha testing of the Air Traffic Controller Test, the Planes test was revised in several ways, including test and practice length, test instructions, response mode, and content.

Test Length. Part 3 was reduced to 48 from 96 questions, the one-minute breaks were cut to 30 seconds, and practice sessions were reduced from 24 to 12 questions.

Mode of Response. The mode of response was changed for all three subtests. Parts 1 and 2 were changed to keys labeled **R** for the red plane and **W** for the white plane instead of the numeric keypad "1" key to represent the red plane and "3" key on the numeric keypad to represent the white plane. Part 3 changed to keys labeled **T** for true and **F** for false, instead of using "1" and "3" of the numeric keypad to represent false and true, respectively.

Test Content. The content of Part 3 was changed so that all questions used "double-negative" statements (e.g., "It is not true that the white plane will not arrive after the red plane."), thereby making Part 3 distinct from Part 2. Previously, some questions in Part 3 were like those in Part 2.

Instructions. The test instructions were simplified in several places. Also, the instructions in the "Answering Test Items" section were revised to correspond to the changes made in mode of response (noted above).

Summary and Recommendations

The project team cut the number of items in each part of the original Planes test in half for the alpha data collection effort. This was done to meet project time constraints. After completing reliability analyses, it was clear that the test would benefit from restoring it to its original length. Available test time in the beta version was limited, however. As a result, the number of items in Part 3 and in the practice sessions was cut in half. The time allotted for breaks between the three test parts was also halved.

Experiences Questionnaire

The following Experiences Questionnaire analyses were performed on data from the first 9 of the 12 days of pilot testing at Pensacola in February, 1997. The total

N in this data set is 330. The last 2 days of pilot testing included a large number of the ATCS students; performance of the ATCS students and performance of Non-ATCS students on the EQ have not been compared.

EQ Format

The pilot test version of the EQ contained 201 items representing 17 scales, including a Random Response Scale. All items used the same set of five response options: Definitely True, Somewhat True, Neither True Nor False, Somewhat False, and Definitely False.

Data Screening

Three primary data quality screens are typically performed on questionnaires like the EQ: (a) a missing data screen, (b) an unlikely virtues screen, and (c) a random response screen. The missing data rule used was that if more than 10% of the items on a particular scale were missing (blank), that scale score was not computed. No missing data rule was invoked for across-scale missing data, so there could be a data file with, for example, all scale scores missing. No one was excluded based on responses to the unlikely virtues items, that is, those items with only one "likely" response (Example: "You have never hurt someone else's feelings," where the only "likely" response is "Definitely False").

A new type of random response item was tried out in the pilot test, replacing the more traditional, right/wrong-answer type, such as "Running requires more energy than sitting still." There were four random response items, using the following format: "This item is a computer check to verify keyboard entries. Please select the Somewhat True response and go on to the next item." The response that individuals were instructed to select varied across the four items. A frequency distribution of the number of random responses (responses other than the correct one) follows:

Number of Random Responses	N	Percent
0	222	67.3
1	52	15.8
2	34	10.3
3	18	5.5
4	4	1.2
	330	100.0

As can be seen, a large number of students gave one or more random responses (108, or 32.8%). Whether this indicates that the new random response items are too difficult, or that a large number of students were not attending very closely to the EQ (or other tests?) is unclear. Students with two or more random responses were removed from the data set, resulting in a screened sample of 274 EQs available for further analysis.

Time to Complete EQ

The mean amount of time required to complete the EQ for the screened data set was 29.75 minutes (SD = 9.53, Range = 10-109). A few individuals finished in approximately 10 minutes, which translates into roughly 3 seconds per response. The records of the fastest finishers were checked for unusual response patterns such as repeating response patterns or patterns of all the same response (which would yield a high random response score anyway), and none were found. Thus, no one was deleted from the data set due solely to time taken to complete the test. It is not surprising to note that the fastest finishers in the entire, unscreened sample of 330 were deleted based on their scores on the random response scale.

Scale Scoring

EQ items were keyed 1 - 5, the appropriate items were reversed (5 - 1), and the scale scores were computed as (the mean item response) x 20, yielding scores ranging from 20 to 100. The higher the score, the higher the standing on the characteristic.

Descriptive Statistics and Reliability Estimates

Appendix B contains the descriptive statistics and internal consistency reliabilities for 16 scales (Random Response Scale excluded). The scale means were comfortingly low and the standard deviations were comfortingly high, relieving concerns about too little variance and/or a ceiling effect. The Unlikely Virtues scale had the lowest mean of all (51.85), as it should.

The scale reliabilities were within an acceptable range for scales of this length and type. Most were in the .70s and .80s. The two exceptions were Self Awareness (.55) and Self-Monitoring/Evaluating (.54).

Four items had very low item-scale correlations, so they were removed from their respective scales: Items 21 and 53 from the Decisiveness scale (item-scale correlations of -.02 and -.05 respectively), item 144 from the Self-Monitoring/Evaluating scale (correlation of .04), and item 163 from the Interpersonal Tolerance scale

(correlation of -.19). Each of these four items was correlated with all of the other scales to see if they might be better suited to another scale. Item 144 correlates .23 with the Interpersonal Tolerance scale, and its content is consistent with that scale, so it has been moved. The remaining three items either did not correlate high enough with other scales, or the item content was not sufficiently related to the other scales to warrant moving them. These three items were deleted, and the descriptive statistics and internal consistency reliabilities were rerun for the three scales affected by the item deletions/moves. Appendix B contains the revised descriptive statistics and internal consistency reliabilities for the three scales affected.

At the item level, the means and standard deviations were satisfactory. (Item means and SDs can be found in the reliability output in Appendix B, The only items with extreme values and/or low standard deviations were on the Unlikely Virtues scale, which is as it should be.

Scale Intercorrelations and Factor Analysis

Appendix B also contains EQ scale intercorrelations and factor analysis output. Principal axis factor analysis was used, and the 2-, 3-, and 4-factor solutions were examined, with solutions rotated to an oblimin criterion. As can be seen in Appendix B, there is a large positive manifold. Consequently, there is a large general factor, and it is most likely that any other factors that emerge will be at least moderately correlated.

In the 2-factor solution, the two factors correlate .75. Factor 1 consists of Decisiveness, Concentration, Self-Confidence, Task Closure/Thoroughness, Taking Charge, Execution, Composure, Tolerance for High Intensity, Sustained Attention, and Flexibility. Factor 2 consists of Interpersonal Tolerance, Working Cooperatively, Consistency of Work Behaviors, Self-Awareness, and Self-Monitoring/Evaluating. Although the high correlation between these two factors indicates that a general factor accounts for much of the variance in these two factors, there is some unique variance. Factor 1 appears to reflect a cool, confident, decisive character; Factor 2 appears to reflect a character that is self-aware and works well with others.

In the 3-factor solution, the third factor does not appear to add any useful information. The 4-factor solution appears to be interpretable. In this solution, the factors are comprised of the following scales:

Factor 1: Concentration, Tolerance for High Intensity, Composure, Decisiveness, Sustained Attention, and Flexibility.
Factor 2: Consistency of Work Behaviors, Interpersonal Tolerance, and Self- Awareness.
Factor 3: Self-Monitoring/Evaluating and Working Cooperatively.
Factor 4: Taking Charge, Self-Confidence, Task Closure/Thoroughness, and Execution.

In the 4-factor solution, the first factor of the 2-factor solution is split into two parts. One part (Factor 1) contains scales related to maintaining attentional focus and the ability to remain composed and flexible. The other part (Factor 4) contains scales related to taking charge of situations and following through. The second factor in the 2-factor solution also split into two parts in the 4-factor solution, although not quite so tidily. Actually, Working Cooperatively correlates just about equally with Factors 2 and 3 of the 4-factor solution. If EQ predictor composites were to be created at this point, the tendency would be toward three composites, drawn from the 4-factor solution: Factor 1, Factor 4, and the combination of Factors 2 and 3.

Summary and Recommendations

The EQ results in the pilot test were promising. Most of the scales looked good in terms of their means, variances, and reliabilities. The two scales that were weakest, psychometrically, were Self-Awareness and Self-Monitoring/Evaluating.

Item analysis suggested that items 21, 53, and 163 should be deleted, and item 144 moved to a different scale. If the EQ must be shortened, deletion of scales rather than individual items seemed preferable, given the high correlations between scales. However, even the scales most highly correlated (e.g., Decisiveness and Sustained Attention, r = .80, and Decisiveness and Composure, r = .81) appear to be measuring somewhat different constructs. Based on considerations including the desired length of the beta version of the AT-SAT test battery, a final decision was made to decrease the EQ to 175 items. The Self-Monitoring scale was deleted in its entirety, and several scales were shortened slightly.

The issue of how to use the Unlikely Virtues scale remained unresolved. Although the mean and standard deviation for this scale appeared just as they should in

the pilot test, this sample did not provide any information about how much "faking good" would actually occur in an applicant population.

Air Traffic Scenarios

The Air Traffic (AT) Scenarios test consisted of two brief practice scenarios of 4 to 5 minutes each, and four test scenarios of 15 to 20 minutes each. One-fourth of the examinees that completed the AT test were also given a seventh (retest) test scenario at the end of the day. Two types of scores were recorded for each trial. First, there were counts of different types of errors, including crashes and separation errors (plane-to-plane and plane-to-boundary) and procedural errors (wrong destination, landing/exit speed, or landing/exit altitude). Second, there were efficiency measures expressed in terms of percentage of aircraft reaching their target destination and delays in getting the aircraft to their destination and in accepting handoffs.

Scoring

Initial inspection of the results suggested that crashes and separation errors (safety) were relatively distinct from (uncorrelated with) procedural errors. Consequently, four separate scores were generated to account for the data. Initial scores were:

CRASHSEP = crashes + separation errors
PROCERR = total number of procedural errors of all kinds
PCTDEST = percent reaching target destination
TOTDELAY = total delay (handoff and enroute)

In computing safety errors, crashes were initially given a weight of 4.0 to equalize the variance of crashes and separation errors. Since crashes are relatively rare events, overweighting crashes led to reduced consistency across trials (reliability). Alternative weightings might be explored at a later date, but would be expected to make little difference. Consequently, it was decided to simply count crashes as an additional separation error.

One other note about the initial computation of scores is that airport flyovers were initially listed with separation errors but appeared to behave more like procedural errors. Examinees are not given the same type of warning signal when an aircraft approaches an airport as when it approaches another plane or a boundary, so avoiding airport flyovers was more a matter of knowing and following the rules.

For all measures except PCTDEST, the next step was to define a new scaling of each of these variables so that higher scores indicated better performance and so that the scale would be most sensitive to differences at higher levels of performance. In the initial scaling, the difference between 0 and 1 error was treated the same as the difference between 50 and 51 errors, even though the former is a much more important distinction. The transformations used were of the form:

$$\text{New Scale} = 1 / (a + b*\text{Old Scale})$$

where a and b were chosen so that optimal performance would be around 100 and performance at the average of the old scale would map onto 50. For the AT Test, optimal performance was indicated by 0 on each of the original measures so that the transformation could be rewritten as:

$$\text{New Scale} = 100 / (1 + \text{Old Scale}/\text{Old Mean}).$$

It was also decided to scale each trial separately. The last two trials were considerably more difficult than the preceding ones, so variance in performance was much higher for these trials. If the data from each trial were not rescaled separately, the last trials would receive most of the effective weight when averages were computed. Consequently, the means referred to in the above formula were trial-specific means. The new scale variables for each trial had roughly equivalent means and variances which facilitated cross-trial comparisons and averaging.

Case Elimination

During the initial analyses, prior to rescaling, there were several cases with very high error rates or long delay times that appeared to be outliers. The concern was that these individuals did not understand the instructions and so were not responding appropriately. (In one case, it was suspected that the examinee was crashing planes on purpose.) The rescaling, however, shrunk the high end (high errors or long times) of the original scales relative to the lower end, and after rescaling these cases were not clearly identifiable as outliers. Inspection of the data revealed that all of the cases of exceptionally poor performance occurred on the last test trial. The fact that the last trial was exceptionally difficult and that similar problems were not noted on the earlier trials, suggested that most of these apparent outliers were simply instances of low ability and not random or inappropriate

responding. In the end, cases with more than 125 crash/separation errors or more than 400 minutes of total delay time were flagged as probable "random" (inappropriate) responders. A total of 16 cases were so flagged.

There were a number of instances of incomplete data. The alpha pilot version of the software was not completely shock-proofed, and some examinees managed to "skip out" of a trial without completing it. This rarely happened on either the first or the last (fourth test) trial. Where there was only one missing trial, averages were computed across the remaining trials. Where more than one trial was missing, the overall scores were set to missing as well. In the end, we also flagged cases missing either of the last two test trials. A total of 38 cases were so flagged, leaving 386 cases with no flags for use in analyses.

Reliability

After revised scale scores were computed for each trial, reliability analyses were performed. In this case, an ANOVA (generalizability) model was used to examine the variance in scores across trials, examinee groups (test orders), and examinees (nested within groups). The analyses were conducted for varying numbers of trials, from all six (two practice and four test) down to the last two (test) trials. Table 3.3.29 shows variance component estimates for each of the sources of variation. Notwithstanding modest efforts to standardize across trials, there was still significant variation due to Trial main effects in many cases. These were ignored in computing reliabilities (using relative rather than absolute measures of reliability) since the trials would be constant for all examinees and would not contribute to individual variation in total scores. Similarly, Group and Group by Trial effects were minimal and were not included in the error term used for computing reliabilities. Group effects are associated with different positions in the overall battery. There will be no variation of test position in the final version of the battery.

Single trial reliabilities were computed as the ratio of the valid variance due to subjects nested within groups, SSN(Group) to the total variance, expressed as the sum of SSN(Group) and SSN*Trial. For each variable, the single trial reliability based on the last two trials was identical to the correlation between the scores for those two trials. Reliabilities for means across higher numbers of trials were computed by dividing the SSN*T error component by the number of trials. This is exactly the Spearman-Brown adjustment expressed in generalizability terms.

Another measure of reliability was the correlation between the overall scores generated during the regular testing and the "retest" scores for those examinees who completed an additional trial at the end of the day. Table 3.3.30 shows the correlation between alternative composite scores and the retest score. The alternative composites included means across trials, possibly leaving out the first few trials, and a weighted composite giving increasing weight to the later composites. (For AT, the weights were 0 for the practice trials and 1, 2, 3, and 4 for the test trials. For the TW test, the weights were 1, 2, and 3 for the three regular trials.) The row labeled SEPSK1-6, for example, corresponds to the simple mean across all six (two practice and four test) trials. Since the retest was a single trial and, in most cases, the composite score from regular testing encompassed more than one trial, the two measures being correlated do not have equal reliability. In general, as expected, the correlations ranged between the values estimated for single trial reliabilities and reliability estimates based on the number of trials included in the composite scores. In some cases, these "test-retest" reliabilities were lower than the internal consistency estimates, indicating some individual differences in the retention of skill over the course of the testing day.

Based on analyses of the reliability data, it was concluded that the most appropriate scores for use with the pilot data were averages of the last two test trials. The 2-trial reliability for these scores was higher than the three-trial reliability for the last 3 scores, the 4-trial reliability for the last four scores, and so on. The composite based on the last two trials also had the highest correlation with the retest scores in most cases or was at least a close second.

Summary and Recommendations

It was felt if separate scores were to be used in the concurrent validation, additional practice and test trials would be needed to achieve a high level of reliability for the "Separation Skill" variable. It was recommended that **three practice trials be used with each trial targeted to test understanding of specific rules and more tailored feedback after each trial**. For example, the first trial might include four planes, two headed for their designated airport runways and two headed for their designated exit gates. One of the two exiting planes would be at the wrong level and the other at the wrong speed. Similarly, one of the landing planes would be at the wrong level and the other at the wrong speed. No changes in direction would be required. At the end of a

very brief time, it could be determined whether the examinee changed level and speed appropriately for each aircraft, with feedback if they did not.

A second example might involve turning planes to get to their destinations. Specific feedback on changing directions would be given if the planes failed to reach their assigned destination. Further testing of speed and level rules would also be included. The final practice scenario would involve airport landing directions and flyovers.

Following the three practice scenarios (which might take a total of 10 minutes to run with another 10 minutes for feedback), five test scenarios with increasing difficulties were proposed. The alpha fourth test scenario may be a bit too difficult and might be toned down a little. However, controller performance is expected to be at a much higher level, so at least two relatively difficult scenarios should be included. After three practice and three easier test scenarios, performance on the last two more difficult scenarios should be quite reliable.

Software Changes

After the Alpha Version pilot test, the AT test was changed to have more extensive and more highly edited instructions and was converted to a 32-bit version to run under Windows 95. The practice scenarios were modified to "teach" specific aspects of the exercises (changing speed and level in practice 1, changing directions in practice 2, noticing airport landing directions, and coping with pilot readback errors in practice 3). Specific feedback was provided after each practice session keyed to aspects of the examinee's performance on the practice trial.

The "new" version of the scenario player provided slightly different score information. In particular the "en route delay" variable was computed as the total en route time for planes that landed correctly. We modified the shell program to read the "replay" file and copy information from the "exit" records (type XT) into the examinee's data file. This allowed us to record which planes either crashed or were still flying at the end of the scenario. We computed a "total en route" time to replace the "delay" time provided by the Alpha version.

Time Wall/Pattern Recognition Test

The analyses for the Time Wall (TW) test were very similar to the analyses performed for the Air Traffic Scenarios test. One difference was that TW had three exactly parallel trials instead of two practice and four test scenarios that differed in difficulty. Each TW trial had a brief "practice" trial where no results were recorded.

The three scores analyzed for the TW test were (a) Pattern Recognition Accuracy (PRACCY), defined as the percent of correct pattern matching responses out of all correct and incorrect responses (e.g., excluding timeouts); (b) Pattern Recognition Speed (PRSPD), a transformation of the average time, in milliseconds, for correct responses; and (c) Time Wall Accuracy (TWACCY), a transformation of the mean absolute time error, in milliseconds. The transformations used for Pattern Recognition Speed and Time Wall Accuracy were identical in form to those used with the AT test. In this case, however, the transformations mapped the maximum value to about 100 and the mean value to about 50 across all trials, rather than using a separate transformation for each trial. This was done because the trials did not vary in difficulty for TW as they did for AT.

Case Elimination

Figure 3.3.1 shows a plot of the Pattern Recognition Accuracy and Speed variables. A number of cases had relatively high speed scores and lower than chance (50%) accuracy scores. In subsequent analyses, all cases with an accuracy score less than 40 on any of the operational trials were deleted. This resulted in a deletion of 12 participants.

Reliability

Tables 3.3.29 and 3.3.30 show internal consistency and test-retest reliability estimates for TW as well as for AT. Analyses of these data suggested that averaging across all three trials led to the most reliable composite for use in analyses of the pilot data.

Summary and Recommendations

Time Wall Accuracy reliability estimates were modest, although the test-retest correlations held up fairly well. Preliminary results suggested that five or six trials may be needed to get highly reliable results on all three measures.

Software Changes

The trial administration program was changed to allow us to specify the number of Time Wall items administered and to shut off the "warm up" trials for each administration. The main program then called the trial administration program 6 times. The first three trials had 5 Time Wall items each and were considered test trials. The next three trials had 25 Time Wall items each and were considered test trials. After the practice trials, the examinee's performance was analyzed and specific feedback was given on how to improve their score.

Testing Times

Table 3.3.31 shows distributional statistics for instruction time and total time for the AT and TW tests in their current form. While there was some variation in instruction time, the total times were quite close to the original targets (90 and 25 minutes, respectively).

Conclusions

The purpose of the pilot study was to determine if the predictor battery required revisions prior to its use in the proposed concurrent validation study. A thorough analysis of the various tests was performed. A number of recommendations related to software presentation - item changes, and predictor construct revisions - were outcomes of the pilot study. The project team believed that the changes made to the test battery represented a substantial improvement over initial test development. The beta battery, used in the concurrent validation study, was a professionally developed set of tests that benefited greatly from the pilot study.

REFERENCES

Aerospace Sciences, Inc. (1991). Air traffic control specialist pre-training screen preliminary validation. Fairfax, VA: Aerospace Sciences, Inc.

Alexander, J., Alley, V., Ammerman, H., Fairhurst, W., Hostetler, C., Jones, G., & Rainey, C. (1989, April). FAA air traffic control operation concepts: Volume VII, ATCT tower controllers (DOT/FAA/AP-87/01, Vol. 7). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration.

Alexander, J., Alley, V., Ammerman, H., Hostetler, C., & Jones, G. (1988, July). FAA air traffic control operation concepts: Volume II, ACF/ACCC terminal and en route controllers (DOT/FAA/AP-87/01, Vol. 2, CHG 1). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration.

Alexander, J., Ammerman, H., Fairhurst, W., Hostetler, C., & Jones, G. (1989, September). FAA air traffic control operation concepts: Volume VIII, TRACON controllers (DOT/FAA/AP-87/01, Vol. 8). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration.

Alley, V., Ammerman, H., Fairhurst, W., Hostetler, C., & Jones, G. (1988, July). FAA air traffic control operation concepts: Volume V, ATCT/TCCC tower controllers (DOT/FAA/AP-87/01, Vol. 5, CHG 1). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration.

Ammerman, H., Bergen, L., Davies, D., Hostetler, C., Inman, E., & Jones, G. (1987, November). FAA air traffic control operation concepts: Volume VI, ARTCC/HOST en route controllers (DOT/FAA/AP-87/01, Vol. 6). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration.

Ammerman, H., Fairhurst, W., Hostetler, C., & Jones, G. (1989, May). FAA air traffic control task knowledge requirements: Volume I, ATCT tower controllers (DOT/FAA/ATC-TKR, Vol. 1). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration.

Ammerman, H., Fligg, C., Pieser, W., Jones, G., Tischer, K., Kloster, G. (1983, October). Enroute/terminal ATC operations concept (DOT/FAA/AP-83/16) (CDRL-AOO1 under FAA contract DTFA01-83-Y-10554). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Advanced Automation Program Office.

Bobko, P., Nickels, B. J., Blair, M. D., & Tartak, E. L. (1994). Preliminary internal report on the current status of the SACHA model and task interconnections: Volume I.

Boone, J. O. (1979). Toward the development of a new selection battery for air traffic control specialists. (DOT/FAA/AM-79/21). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

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

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.

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

Boone, J. O (1979). Toward the development of a new selection battery for air traffic control specialists (DOT/FAA/AM-79/21). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Broach, D. & Brecht-Clark, J. (1994). Validation of the Federal Aviation Administration air traffic control specialist pre-training screen (DOT/FAA/AM-94/4). Oklahoma City, OK: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Broach, D. (1996, November). User's Guide for v4.0 of the Air Traffic Scenarios Test for Windows((WinATST). Oklahoma City, OK: Federal Aviation Administration Civil Aeromedical Institute, Human Resources Research Division.

Brokaw, L. D. (1957, July). Selection measures for air traffic control training. (Technical Memorandum PL-TM-57-14). Lackland Air Force Base, TX: Personnel Laboratory, Air Force Personnel and Training Research Center.

Brokaw, L. D. (1959). School and job validation of selection measures for air traffic control training. (WADC-TN-59-39). Lackland Air Force Base, TX: Wright Air Development Center, United States Air Force.

Brokaw, L. D. (1984). Early research on controller selection: 1941-1963. In S. B. Sells, . T. Dailey, E. W. Pickrel (Eds.) Selection of Air Traffic Controllers. (DOT/FAA/AM-84/2). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Buckley, E. P., & Beebe, T. (1972). The development of a motion picture measurement instrument for aptitude for air traffic control (DOT/FAA/RD-71/106). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Systems Research and Development Service.

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

Buckley, E. P., House, K., & Rood, R. (1978). Development of a performance criterion for air traffic control personnel research through air traffic control simulation. (DOT/FAA/RD-78/71). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Systems Research and Development Service.

Buckley, E. P., O'Connor, W. F., & Beebe, T. (1969). A comparative analysis of individual and system performance indices for the air traffic control system (Final report) (DOT/FAA/NA-69/40; DOT/FAA/RD-69/50; Government accession #710795). Atlantic City, NJ: U.S. Department of Transportation, Federal Aviation Administration, National Aviation Facilities Experimental Center, Systems Research and Development Service.

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

Buckley, E. P., O'Connor, W. F., & Beebe, T. (1970). A comparative analysis of individual and system performance indices for the air traffic control system (DOT/FAA/NA-69/40). Atlantic City, N.J: U.S. Department of Transportation, Federal Aviation Administration, National Aviation Facilities Experimental Center.

Cattell, R. B., & Eber, H. W. (1962). The sixteen personality factor questionnaire. Champaign, IL: Institute for Personality and Ability Testing.

Carter, D. S. (1979). Comparison of different shrinkage formulas in estimating population umultiple correlation coefficients. Educational and Psychological Measurement, 39, 261-266.

Cobb, B. B. (1967). The relationships between chronological age, length of experience, and job performance ratings of air route traffic control specialists (DOT/FAA/AM-67/1). Oklahoma City, OK: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Cobb, B. B. & Mathews, J. J. (1972). A proposed new test for aptitude screening of air traffic controller applicants. (DOT/FAA/AM-72/18). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Collins, W. E., Manning, C. A., & Taylor, D. K. (1984). A comparison of prestrike and poststrike ATCS trainees: Biographic factors associated with Academy training success. In A. VanDeventer,, W. Collins, C. Manning, D. Taylor, & N. Baxter (Eds.) Studies of poststrike air traffic control specialist trainees: I. Age, biographical actors, and selection test performance. (DOT/FAA/AM-84/18). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Collins, W. E., Nye, L. G., & Manning, C. A.. (1990). Studies of poststrike air traffic control specialist trainees: III. Changes in demographic characteristics of Academy entrants and bio-demographic predictors of success in air traffic control selection and Academy screening. (DOT/FAA/AM-90/4). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Convey, J. J. (1984). Personality assessment of ATC applicants. In S. B. Sells, J. T. Dailey, E. W. Pickrel (Eds.) Selection of Air Traffic Controllers. (DOT/FAA/AM-84/2). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Cooper, M., Blair, M. D., & Schemmer, F.M. (1994). Separation and Control Hiring Assessment (SACHA) Draft Preliminary Approach Predictors Vol 1: Technical Report . Bethesda, MD: University Research Corporation.

Costa, P.T., Jr., & McCrae, R.R. (1988). Personality in Adulthood: A six-year longitudinal study of self-reports and spouse ratings on the NEO personality inventory. Journal of Personality and Social Psychology, 54, 853-863.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). Manual for Kit of Factor-Referenced Cognitive Tests. Princeton, NJ: Educational Testing Service.

Fleishman, E.A., & Quaintance, M.K. (1984). Taxonomies of human performance. Orlando, FL: Academic Press.

Gibb, G.D., Smith, M.L., Swindells, N., Tyson, D., Gieraltowski, M.J., Petschauser, K.J., & Haney, D.U. (1991). The development of an experimental selection test battery for air traffic control specialists. Daytona Beach, FL.

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

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

Hogan, R. (1996). Personality Assessment. In R.S. Barrett (Ed.), Fair Employment in Human Resource Management (pp.144-152). Westport, Connecticut: Quorum Books.

Houston, J.S., & Schneider, R.J. (1997). Analysis of Experience Questionnaire (EQ) Beta Test Data. Unpublished manuscript.

Human Technology, Inc. (1991). Cognitive task analysis of en route air traffic controller: Model extension and validation (Report No. OPM-87-9041). McLean, VA: Author.

Human Technology, Inc. (1993). Summary Job Analysis. Report to the Federal Aviation Administration Office of Personnel, Staffing Policy Division. Contract #OPM-91-2958, McLean, VA: Author.

Landon, T.E. (1991). Job performance for the en-route ATCS: A review with applications for ATCS selection. Paper submitted to Minnesota Air Traffic Controller Training Center.

Manning, C. A. (1991). Individual differences in air traffic control specialist training performance. Journal of Washington Academy of Sciences, 11, 101-109.

Manning, C. A. (1991). Procedures for selection of air traffic control specialists. In H. Wing & C. Manning (Eds.) Selection of air traffic controllers: Complexity, requirements and public interest. (DOT/FAA/AM-91/9). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Manning, C.A., Della Rocco, P. S., & Bryant, K. D. (1989). Prediction of success in air traffic control field training as a function of selection and screening test performance . (DOT/FAA/AM-89/6). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Mecham, R.C., & McCormick, E.J. (1969). The rated attribute requirements of job elements in the position analysis questionnaire. Office of Naval Research Contract Nonr-1100 (28), Report No. 1. Lafayette, Ind.: Occupational Research Center, Purdue University.

Mies, J., Coleman, J. G., & Domenech, O. (1977). Predicting success of applicants for positions as air traffic control specialists in the Air Traffic Service (Contract No. DOT FA-75WA-3646). Washington, DC: Education and Public Affairs, Inc.

Milne, A. M. & Colmen, J. (1972). Selection of air traffic controllers for FAA. Washington, DC: Education and Public Affairs, Inc. (Contract No. DOT=FA7OWA-2371).

Myers, J., & Manning, C. (1988). A task analysis of the Automated Flight Service Station Specialist job and its application to the development of the Screen and Training program (Unpublished manuscript). Oklahoma City, OK: Civil Aeromedical Institute, Human Resources Research Division.

Nickels, B.J., Bobko, P., Blair, M.D., Sands, W.A., & Tartak, E.L. (1995). Separation and control hiring assessment (SACHA) final job analysis report (Deliverable Item 007A under FAA contract DFTA01-91-C-00032). Washington, DC: Federal Aviation Administration, Office of Personnel.

Potosky, D. , & Bobko, P. (1997). Assessing computer experience: The Computer Understanding and Experience (CUE) Scale. Poster presented at the Society for Industrial and Organizational Psychology (SIOP), April 12, St. Louis, MO.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

Pulakos, E. D. (1986). The development of a training program to increase accuracy with different rating formats. Organizational Behavior and Human Decision Processes, 38, 76-91.

Pulakos, E. D., & Borman, W. C. (1986). Rater orientation and training. In E. D. Pulakos & W. C. Borman (Eds.), Development and field test report for the Army-wide rating scales and the rater orientation and training program (Technical Report #716). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Pulakos, E. D, Keichel, K. L., Plamondon, K., Hanson, M. A., Hedge, J. W., & Borman, W. C. (1996). SACHA task 3 final report. (Institute Report #286). Minneapolis, MN: Personnel Decisions Research Institutes.

Rock, D. B., Dailey, J. T., Ozur, H., Boone, J. O., & Pickerel, E. W. (1978). Study of the ATC job applicants 1976-1977 (Technical Memorandum PL-TM-57-14). In S. B. Sells, J.T. Dailey, & E. W. Pickrel (Eds.), Selection of air traffic controllers (pp. 397-410). (DOT/FAA/AM-84/2).Oklahoma City, OK: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Schemmer, F.M., Cooper, M.A., Blair, M.D., Barton, M.A., Kieckhaefer, W.F., Porter, D.L., Abrahams, N. Huston, J. Paullin, C., & Bobko, P. (1996). Separation and Control Hiring Assessment (SACHA) Interim Approach Predictors Volume 1: Technical Report. Bethesda, MD: University Research Corporation.

Schroeder, D. J., & Dollar, C. S. (1997). Personality characteristics of pre/post-strike air traffic control applicants. (DOT/FAA/AM-97/17). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Schroeder, D. J., Dollar, C. S., & Nye, L. G. (1990). Correlates of two experimental tests with performance in the FAA Academy Air Traffic Control Nonradar Screen Program. (DOT/FAA/AM-90/8). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

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

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

Stein, E. S. (1992). Simulation variables. Unpublished manuscript.

Taylor, M.V., Jr. (1952). The development and validation of a series of aptitude tests for the selection of personnel for positions in the field of Air Traffic Control. Pittsburgh, PA: American Institutes for Research.

Taylor, D. K., VanDeventer, A. D., Collins, W. E., & Boone, J. O. (1983). Some biographical factors associated with success of air traffic control specialist trainees at the FAA Academy during 1980. In A. VanDeventer, D. Taylor, W. Collins, & J. Boone (Eds.) Three studies of biographical factors associated with success in air traffic control specialist screening/training at the FAA Academy. (DOT/FAA/AM-83/6). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Trites, D. K. (1961). Problems in air traffic management: I. Longitudinal prediction of effectiveness of air traffic controllers. (DOT/FAA/AM-61/1). Oklahoma City, OK: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Trites & Cobb (1963.) Problems in air traffic management: IV. Comparison of pre-employment job-related experience with aptitude test predictors of training and job performance of air traffic control specialists. (DOT/FAA/AM-63/31). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Tucker, J. A. (1984). Development of dynamic paper-and-pencil simulations for measurement of air traffic controller proficiency (pp. 215-241). In S. B. Sells, J. T. Dailey & E. W. Pickrel (Eds.), Selection of air traffic controllers (DOT/FAA/AM-84/2). Oklahoma City, OK: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

VanDeventer, A. D. (1983). Biographical profiles of successful and unsuccessful air traffic control specialist trainees. In A. VanDeventer, D. Taylor, W. Collins, & J. Boone (Eds.) Three studies of biographical factors associated with success in air traffic control specialist screening/training at the FAA Academy. (DOT/FAA/AM-83/6). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Weltin, M., Broach, D., Goldbach, K., & O'Donnell, R. (1992). Concurrent criterion related validation of air traffic control specialist pre-training screen. Fairfax, VA: Author.

Wherry, R.J. (1940). Appendix A. In W.H.Stead, & Sharyle (Eds.), C.P. Occupational Counseling Techniques.

Yee, P. L., Hunt, E., & Pellegrino, J. W. (1991). Coordinating cognitive information: Task effects and individual differences in integrating information from several sources. Cognitive Psychology, 23, 615-680.

Figures & Tables

Appendix A and B