DOT/FAA/AM-01/6

Office of Aviation Medicine
Washington, D.C. 20591

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

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

N O T I C E

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

Technical Report Documentation Page

1. Report No. DOT/FAA/AM-01/6	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Documentation of Validity for the AT-SAT Computerized Test Battery, Volume II		5. Report Date March 2001 6. Performing Organization Code
7. Author(s) Ramos, R.A.[1], Heil, M.C.[2], and Manning, C.A.[2]		8. Performing Organization Report No.
9. Performing Organization Name and Address [1]Human Resources Research Organization 68 Canal Center Plaza, Suite 400 Alexandria, VA 22314-1591	[2]FAA Civil Aeromedical Institute P. O. Box 25082 Oklahoma City, OK 73125	10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency name and Address Office of Aviation Medicine Federal Aviation Administration 800 Independence Ave., S. W. Washington, D.C. 20591		13. Type of Report and Period Covered 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 Air Traffic Controllers, Selection, Assessment, Job Analyses		18. Distribution Statement Document is available to the public through the National Technical Information Service, Springfield, Virginia 22161	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 179	22. Price

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

TABLE OF CONTENTS
VOLUME II

Page

CHAPTER 4 - Development of Criterion Measures of Air Traffic Controller Performance 1
Cbpm 1

CHAPTER 5.1 - Field Procedures for Concurrent Validation Study 13

CHAPTER 5.2 - Development of Pseudo-Applicant Sample 17

CHAPTER 5.3 - Development of Data Base 21

CHAPTER 5.4 - Biographical and Computer Experience Information: Demographics for the Validation Study 31
Total Sample 31
Controller Sample 31
Pseudo-Applicant Sample 32
Computer Use and Experience Questionnaire 32
Performance Differences 33
Relationship Between Cue-Plus and Predictor Scores 33
Summary 35

CHAPTER 5.5 - Predictor-Criterion Analyses 37

CHAPTER 5.6 - Analyses of Group Differences and Fairness 43

CHAPTER 6 - The Relationship of FAA Archival Data to AT-SAT Predictor and Criterion Measures .. 49
Previous ATC Selection Tests 49
Other Archival Data Obtained for ATC Candidates 51
Archival Criterion Measures 52
Historical Studies of Validity of Archival Measures 52
Relationships Between Archival Data and AT-SAT Measures 54

REFERENCES 61

List of Figures and Tables

Figures

Figure 4.1. Map of CBPM Airspace 67
Figure 4.2. Airspace Summary: Sector 05 in Hub Center 68
Figure 4.3. Example CBPM Item 69
Figure 4.4. Aero Center Airspace 70
Figure 5.2.1. Sample Classified Newspaper Advertisement for Soliciting Civilian Pseudo-Applicants 70
Figure 5.2.2. Sample flyer advertisement for soliciting civilian pseudo-applicants 71
Figure 5.3.1. AT-SAT Data Base (*) 72
Figure 5.3.2. CD-ROM Directory Structure of AT-SAT Data Base 73
Figure 5.5.1. Expected Performance: OPM vs. AT-SAT 74
Figure 5.5.2. Percentage of Selected Applicants whose Expected Performance is in the Top Third of Current Controllers: OPM vs. AT-SAT 75
Figure 5.6.1. Fairness Regression for Blacks Using AT-SAT Battery Score and Composite Criterion 75

Figure 5.6.2. Fairness Regression for Hispanics Using AT-SAT Battery Score and Composite Criterion 76
Figure 5.6.3. Fairness Regression for Females Using AT-SAT Battery Score and Composite Criterion 77
Figure 5.6.4. Confidence Intervals for the Slopes in the Fairness Regressions 78
Figure 5.6.5. Expected Score Frequency by Applicant Group 79
Figure 5.6.6. Percent Passing by Recruitment Strategy 80

Tables

Table 4.1. CBPM Development and Scaling Participants: Biographical Information 81
Table 4.2. CBPM Scaling Workshops: Interrater Reliability Results 82
Table 4.3. Performance Categories for Behavior Summary Scales 83
Table 4.4. Pilot Test Results: Computer-Based Performance Measure (CBPM) Distribution of Scores 84
Table 4.5. Pilot Test Results: Means and Standard Deviations for Ratings on Each Dimension 85
Table 4.6. Pilot Test Results: Interrater Reliabilities for Ratings 85
Table 4.7. HFPM Pilot Test Results - Correlations Between Ratings for Rater Pairs (Collapsed Across Ratee) Both Across All Scenarios and Within Each Scenario 87
Table 4.8. Rater-Ratee Assignments 88
Table 4.9. Computer-Based Performance Measure (CBPM): Distribution of Scores in Validation Sample 88
Table 4.10. Number and Percentage of Supervisor Ratings at Each Scale Point in the Validation Sample 88
Table 4.11. Number and Percentage of Peer Ratings at Each Scale Point in the Validation Sample 89
Table 4.12. Interrater Reliabilities for Peer, Supervisor and Combined Ratings 89
Table 4.13. Means and Standard Deviations for Mean Ratings on Each Dimension 90
Table 4.14. Correlations Between Rating Dimensions for Peers and Supervisors 91
Table 4.15. Factor Analysis Results for Performance Ratings 92
Table 4.16. Descriptive Statistics of High Fidelity Performance Measure Criterion Variables 92
Table 4.17. Interrater Reliabilities for OTS Ratings (N=24) 93
Table 4.18. Principal Components Analysis of the High Fidelity Criterion Space 93
Table 4.19. Intercorrelations Between Proposed Criterion Scores 95
Table 4.20. Job Analysis-Item Linkage Task Results for CBPM and HFPM 96
Table 5.2.1. 1990-1992 Profile Analysis of Actual FAA ATCS Applicants 97
Table 5.2.2. Bureau of Census Data for Race/Ethnicity 98
Table 5.2.3. Background Characteristics by Testing Samples 98
Table 5.4.1. Ethnicity and Gender of all Participants 99
Table 5.4.2. Educational Background of All Participants 99
Table 5.4.3. Data Collection Locations for All Participants 99
Table 5.4.4. Ethnicity and Gender of Air Traffic Controllers 101
Table 5.4.5. Air Traffic Controller Sample Educational Background 102
Table 5.4.6. Air Traffic Controller Sample from Participating Locations 102
Table 5.4.7. Air Traffic Controller Sample Time in Current Position 102
Table 5.4.8 Air Traffic Controller Sample Job Experience at any Facility 103
Table 5.4.9. Ethnicity and Gender of Pseudo-Applicant Sample 103
Table 5.4.10. CUE-Plus Scale Item Means and Frequencies 105
Table 5.4.11. CUE-Plus Means and Standard Deviations by Sample 106
Table 5.4.12. Inter-Correlations of CUE-Plus Items 107
Table 5.4.13. Item-Total Statistics for CUE-Plus: All Respondents 109
Table 5.4.14. Varimax and Oblique Rotated Factor Patterns (CUE-Plus) 110
Table 5.4.15. Eigenvalues and Variance (CUE-Plus) 110
Table 5.4.16. CUE-Plus Means, S.D. and d-Score for Gender 111
Table 5.4.17. Means, S.D. and d-Score for Ethnicity 111
Table 5.4.18. Correlations between CUE-Plus and Predictor Battery: Controllers 112
Table 5.4.19. Correlations between CUE-Plus and Predictor Battery: Controllers 113
Table 5.4.20. Correlations between CUE-Plus and Predictor Battery: Pseudo Applicants 114
Table 5.4.21. Correlations between CUE-Plus and Predictor Battery: Pseudo Applicants 115

Table 5.4.22. Determinants of Applied Math Test:: No. of Items Correct 116
Table 5.4.23. Determinants of Angles Test: No. of Items Correct 116
Table 5.4.24. Determinants of Air Traffic Scenarios: Efficiency 116
Table 5.4.25. Determinants of Air Traffic Scenarios: Safety 117
Table 5.4.26. Determinants of Air Traffic Scenarios: Procedural Accuracy 117
Table 5.4.27. Determinants of Analogy: Information Processing 117
Table 5.4.28. Determinants of Analogy Test: Reasoning 118
Table 5.4.29. Determinants of Dials Test: No. of Items Correct 118
Table 5.4.30. Determinants of Letter Factory Test: Situational Awareness 118
Table 5.4.31. Determinants of Letter Factory Test: Planning & Thinking Ahead 119
Table 5.4.32. Determinants of Scan Test: Total Score 119
Table 5.5.1. Simple Validities: Correlations Between Predictor Scores and Criteria 120
Table 5.5.2. Incremental Validities: Increases in Validities when Adding a Scale or Test 122
Table 5.5.3. Comparison of Five Predictor Weighting Methods 123
Table 5.5.4. Validity Coefficients for the Predictor Composite 124
Table 5.5.5. Effect of Cut Score on Predicted Controller Performance 125
Table 5.5.6. Expected Performance by Validity and Selectivity 126
Table 5.6.1. Means for All Scales by Sample, Gender, and Race 127
Table 5.6.2. Standard Deviations for All Scales by Sample, Gender, and Race 128
Table 5.6.3. Sample Sizes for All Scales by Sample, Gender, and Race 129
Table 5.6.4. Frequency Table for Chi-Square Test of Association for Predictor Composite 131
Table 5.6.5. Group Differences in Means and Passing Rates for the Pseudo-Applicants 131
Table 5.6.6. Fairness Analysis Results 133
Table 5.6.7. Criterion *d*-Scores Analyses for Controllers 135
Table 5.6.8. Power Analysis of Fairness Regressions 136
Table 5.6.9. Potential Impact of Targeted Recruitment 136
Table 6.1. Correlations Between Archival and AT-SAT Criterion Measures (N=669) 137
Table 6.2. Correlations of Archival Selection Procedures with Archival and AT-SAT Criterion Measures 138
Table 6.3. Correlations of Archival Selection Procedure Components with Archival and AT-SAT Criterion Measures (N=212) 139
Table 6.4. Correlations of Criterion Measures from High Fidelity Simulation with Archival Performance-Based Predictors and Criterion Measures. 141
Table 6.5. Correlations Between OPM Selection Tests and AT-SAT PredictorTests (N=561). 142
Table 6.6. Correlations of AT-SAT Applied Math, Angles, and Dials tests with Archival Dial Reading, Directional Headings, Math Aptitude Tests, & H.S. Math Grades Biographical Item. 143
Table 6.7. Correlation of the Version of Air Traffic Scenarios Test Used in Pre-Training Screen Validation with the Version of Air Traffic Scenarios Test Used in AT-SAT Validation 143
Table 6.8. Oblique Principal Components Analysis of EQ Scales 144
Table 6.9. Description of 16PF Scales. 145
Table 6.10. Correlation of EQ and 16PF Scales 147
Table 6.11. Results of Multiple linear Regression of OPM Rating, Final Score in Nonradar Screen Program, and AT-SAT Predictor Tests on AT-SAT Composite Criterion Measure (N=586) 148

Appendices:

Appendix C - Criterion Assessment Scales C1
Appendix D - Rater Training Script D1
Appendix E - AT-SAT High Fidelity Simulation Over the Shoulder (OTS) Rating Form E1
Appendix F - Behavioral and Event Checklist F1
Appendix G - AT-SAT High Fidelity Standardization Guide G1
Appendix H - Pilot Test Rater Comparisons H1
Appendix I - Sample Cover Letter and Table to Assess the Completeness of Data Transmissions I1

CHAPTER 4

Development of Criterion Measures of Air Traffic Controller Performance

Walter C. Borman, Jerry W. Hedge, Mary Ann Hanson, Kenneth T. Bruskiewicz
Personnel Decisions Research Institutes, Inc.
Henry Mogilka and Carol Manning
Federal Aviation Administration
Laura B. Bunch and Kristen E. Horgen
University of South Florida and
Personnel Decisions Research Institutes, Inc.

INTRODUCTION

An important element of the AT-SAT predictor development and validation project is criterion performance measurement. To obtain an accurate picture of the experimental predictor tests' validity for predicting controller performance, it is important to have reliable and valid measures of controller job performance. That is, a concurrent validation study involves correlating predictor scores for controllers in the validation sample with criterion performance scores. If these performance scores are not reliable and valid, our inferences about predictor test validities are likely to be incorrect.

The job of air traffic controller is very complex and potentially difficult to capture in a criterion development effort. Yet, the goal here was to develop criterion measures that would provide a comprehensive picture of controller job performance.

Initial job analysis work suggested a model of performance that included both maximum and typical performance (Bobko, Nickels, Blair & Tartak, 1994; Nickels, Bobko, Blair, Sands, & Tartak, 1995). More so than with many jobs, maximum "can-do" performance is very important in controlling air traffic. There are times on this job when *the* most important consideration is maximum performance - does the controller have the technical skill to keep aircraft separated under very difficult conditions? Nonetheless, typical performance over time is also important for this job.

Based on a task-based job analysis (Nickels et al., 1995), a critical incidents study (Hedge, Borman, Hanson, Carter & Nelson, 1993), and past research on controller performance (e.g., Buckley, O'Connor, & Beebe, 1969; Cobb, 1967), we began to formulate ideas for the criterion measures. Hedge et al. (1993) discuss literature that was reviewed in formulating this plan, and summarize an earlier version of the criterion plan. Basically, this plan was to develop multiple measures of controller performance. Each of these measures has strengths for measuring performance, as well as certain limitations. However, taken together, we believe the measures will provide a valid depiction of each controller's job performance. The plan involved developing a special situational judgment test (called the Computer-Based Performance Measure, or CBPM) to represent the maximum performance/technical proficiency part of the job and behavior-based rating scales to reflect typical performance. A high-fidelity air traffic control test (the High Fidelity Performance Measure, HFPM) was also to be developed to investigate the construct validity of the lower fidelity CBPM with a subset of the controllers who were administered the HFPM.

The Computer Based Performance Measure (CBPM)

The goal in developing the CBPM was to provide a relatively practical, economical measure of technical proficiency that could be administered to the entire concurrent validation sample. Practical constraints limited the administration of the higher fidelity measure (HFPM) to a subset of the validation sample.

Previous research conducted by Buckley and Beebe (1972) suggested that scores on a lower fidelity simulation are likely to correlate with scores on a real time,

hands-on simulation and also with performance ratings provided by peers and supervisors. Their motion picture or "CODE" test, presented controllers with a motion picture of a radar screen and asked them to note when there were potential conflictions. Buckley and Beebe reported significant correlations between CODE scores and for-research-only ratings provided by the controllers' peers, but the sample size in this research was only 19. Buckley, O'Connor, and Beebe (1969) also reported that correlations between CODE scores and scores on a higher-fidelity simulation were substantial, the highest correlation was .73, but, again, the sample size was very small. Finally, Milne and Colmen (1972) found a substantial correlation between the CODE test and for-research-only job performance ratings. In general, results for the CODE test suggest that a lower-fidelity simulation can capture important air traffic controller judgment and decision-making skills.

Again, the intention in the present effort was to develop a computerized performance test that as closely as possible assessed the critical technical proficiency, separating-aircraft part of the controller job. Thus, the target performance constructs included judgment and decision making in handling air traffic scenarios, procedural knowledge about how to do technical tasks, and "confliction prediction"; i.e., the ability to know when a confliction is likely to occur sometime in the near future if nothing is done to address the situation.

The CBPM was patterned after the situational judgment test method. The basic idea was to have an air traffic scenario appear on the computer screen, allow a little time for the problem to evolve, and then freeze the screen and ask the examinee a multiple choice question about how to respond to the problem. To develop this test, we trained three experienced controllers on the situational judgment test method and elicited initial ideas about applying the method to the air traffic context.

The first issue in developing this test was the airspace in which the test would be staged. There is a great deal of controller job knowledge that is unique to controlling traffic in a specific airspace (e.g., the map, local obstructions). Each controller is trained and certified on the sectors of airspace where he or she works. Our goal in designing the CBPM airspace was to include a set of airspace features (e.g., flight paths, airports, special use airspace) sufficiently complicated to allow for development of difficult, realistic situations or problems, but to also keep the airspace relatively simple because it is important that controllers who take the CBPM can learn these features very quickly. Figure 4.1 shows the map of the CBPM airspace, and Figure 4.2 is a summary of important features of this airspace that do not appear on the map.

After the airspace was designed, the three air traffic controller subject matter experts (SMEs) were provided with detailed instructions concerning the types of scenarios and questions appropriate for this type of test. These SMEs then developed several air traffic scenarios on paper and multiple choice items for each scenario. The plan was to generate many more items than were needed on the final test, and then select a subset of the best items later in the test development process. Also, based on the job analysis (Nickels et al., 1995) a list of the 40 most critical en route controller tasks was available, and one primary goal in item development was to measure performance in as many of these tasks as possible, especially those that were rated most critical.

At this stage, each scenario included a map depicting the position of each aircraft at the beginning of the scenario, flight strips that provided detailed information about each aircraft (e.g., the intended route of flight), a status information area (describing weather and other pertinent background information), and a script describing how the scenario would unfold. This script included the timing and content of voice communications from pilots and/or controllers, radar screen updates (which occur every 10 seconds in the en route environment), other events (e.g., hand-offs, the appearance of unidentified radar targets, emergencies), and the exact timing and wording of each multiple choice question (along with possible responses).

After the controllers had independently generated a large number of scenarios and items, we conducted discussion sessions in which each SME presented his scenarios and items, and then the SMEs and researchers discussed and evaluated these items. Discussion included topics such as whether all necessary information was included, whether the distractors were plausible, whether or not there were "correct" or at least better responses, whether the item was too tricky (i.e., choosing the most effective response did not reflect an important skill), or too easy (i.e., the correct response was obvious), and whether the item was fair for all facilities (e.g., might the item be answered differently at different facilities because of different policies or procedures?). As mentioned previously, the CBPM was patterned after the situational judgment test approach. Unlike other multiple choice tests, there was not necessarily only one correct answer, with all the others being wrong. Some

items had, for example, one best answer and one or two others that represented fairly effective responses. These test development sessions resulted in 30 scenarios and 99 items, with between 2 and 6 items per scenario.

An initial version of the test was then programmed to run on a standard personal computer with a 17-inch high-resolution monitor. This large monitor was needed to realistically depict the display as it would appear on an en route radar screen. The scenarios were initially programmed using a "radar engine," which had previously been developed for the FAA for training purposes. This program was designed to realistically display airspace features and the movement of aircraft. After the scenarios were programmed into the radar engine, the SMEs watched the scenarios evolve and made modifications as necessary to meet the measurement goals. Once realistic positioning and movement of the aircraft had been achieved, the test itself was programmed using Authorware. This program presented the radar screens, voice communications, and multiple choice questions, and also it collected the multiple choice responses.

Thus, the CBPM is essentially self-administering and runs off a CD-ROM. The flight strips and status information areas are compiled into a booklet, with one page per scenario, and the airspace summary and sector map (see Figures 4.1 and 4.2) are displayed near the computer when the test is administered. During test administration, controllers are given 60 seconds to review each scenario before it begins. During this time, the frozen radar display appears on the screen, and examinees are allowed to review the flight strips and any other information they believe is relevant to that particular scenario (e.g., the map or airspace summary). Once the test items have been presented, they are given 25 seconds to answer the question. This is analogous to the controller job, where they are expected to "get the picture" concerning what is going on in their sector of airspace, and then are sometimes required to react quickly to evolving situations. We also prepared a training module to familiarize examinees with the airspace and instructions concerning how to take the test.

After preparing these materials, we gathered a panel of four experienced controllers who were teaching at the FAA Academy and another panel of five experienced controllers from the field to review the scenarios and items. Specifically, each of these groups was briefed regarding the project, trained on the airspace, and then shown each of the scenarios and items. Their task was to rate the effectiveness level of each response option. Ratings were made independently on a 1-7 scale. Table 4.1 describes the controllers who participated in this initial scaling workshop, and Table 4.2 summarizes the intraclass correlation, interrater agreement across items for the two groups. After this initial rating session with each of the groups, the panel members compared their independent ratings and discussed discrepancies. In general, two different outcomes occurred as a result of these discussions. In some cases, one or two SMEs failed to notice or misinterpreted part of the item (e.g., did not examine an important flight strip). For these cases, no changes were generally made to the item. In other cases, there was a legitimate disagreement about the effectiveness of one or more response options. Here, we typically discussed revisions to the item or the scenario itself that would lead to agreement between panel members (without making the item overly transparent). In addition, discussions with the first group indicated that several items were too easy (i.e., the answer was obvious). These items were revised to be less obvious. Five items were dropped because they could not be satisfactorily revised.

These ratings and subsequent discussions resulted in substantial revisions to the CBPM. The revisions were accomplished in preparation for a final review of the CBPM by a panel of expert SMEs. For this final review session, 12 controllers from the field were identified who had extensive experience as controllers and had spent time as either trainers or supervisors. Characteristics of this final scaling panel group are shown in Table 4.1.

The final panel was also briefed on the project and the CBPM and then reviewed each item. To ensure that they used all of the important information in making their ratings, short briefings were prepared for each item, highlighting the most important pieces of information that affected the effectiveness of the various responses. Each member of the panel then independently rated the effectiveness level of each response option. This group did not review each other's ratings or discuss the items.

Interrater agreement data appear in Table 4.2. These results show great improvement because in the final scaling of the CBPM, 80 of the 94 items have interrater reliability. As a result of the review, 5 items were dropped because there was considerable disagreement among raters. These final scaling data were used to score the CBPM. For each item, examinees were assigned the mean effectiveness of the response option they chose, with a few exceptions. First, for the knowledge items, there was only one correct response. Similarly, for the "confliction prediction" items, there was one correct response. In addition, it is more effective to predict a

confliction when there is not one (i.e., be conservative) than to fail to predict a confliction when there is one. Thus, a higher score was assigned for an incorrect conservative response than an incorrect response that predicted no confliction when one would have occurred. The controller SMEs generated rational keys for 23 knowledge and confliction prediction type items. Figure 4.3 shows an example of a CBPM item. One final revision of the CBPM was made based on pilot test data. The pilot test will be discussed in a later section.

The Behavior Summary Scales

The intention here was to develop behavior-based rating scales that would encourage raters to make evaluations as objectively as possible. An approach to accomplish this is to prepare scales with behavioral statements anchoring different effectiveness levels on each dimension so that the rating task is to compare observed ratee behavior with behavior on the scale. This matching process should be more objective than, for example, using a 1 = very ineffective to 7 = very effective scale. A second part of this approach is to orient and train raters to use the behavioral statements in the manner intended.

The first step in scale development was to conduct workshops to gather examples of effective, mid-range, and ineffective controller performance. Four such workshops proceeded with controllers teaching at the FAA academy and with controllers at the Minneapolis Center. A total of 73 controllers participated in the workshops; they generated 708 performance examples.

We then analyzed these performance examples and tentatively identified eight relevant performance categories: (1) Teamwork, (2) Coordinating, (3) Communicating, (4) Monitoring, (5) Planning/Prioritizing, (6) Separation, (7) Sequencing/Preventing Delays, and (8) Reacting to Emergencies. Preliminary definitions were developed for these categories. A series of five "mini-workshops" were subsequently held with controllers to review the categories and definitions. This iterative process, involving 24 controllers, refined our set of performance categories and definitions. The end result was a set of ten performance categories. These final categories and their definitions are shown in Table 4.3.

Interestingly, scale development work to this point resulted in the conclusion that these ten dimensions were relevant for all three controller options: tower cab, TRACON, and en route. However, subsequent work with tower cab controllers resulted in scales with somewhat different behavioral content. Because AT-SAT focused on en route controllers, we limit our discussion to scale development for that group.

The next step was to "retranslate" the performance examples. This required controller SMEs to make two judgments for each example. First, they assigned each performance example to one (and only one) performance category. Second, the controllers rated the level of effectiveness (from 1 = very ineffective to 7 = very effective) of each performance example.

Thus, we assembled the ten performance categories and 708 performance examples into four separate booklets that were used to collect the SME judgments just discussed. In all, booklets were administered to 47 en route controllers at three sites within the continental United States. Because each booklet required 2-3 hours to complete, each of the SMEs was asked to complete only one booklet. As a result, each performance example or "item" was evaluated by 9 to 20 controllers.

Results of the retranslation showed that 261 examples were relevant to the en route option, were sorted into a single dimension more than 60% of the time, *and* had standard deviations of less than 1.50 for the effectiveness ratings. These examples were judged as providing unambiguous behavioral performance information with respect to both dimension and effectiveness level.

Then for each of the ten dimensions, the performance examples belonging to that dimension were further divided into high effectiveness (retranslated at 5 to 7), middle effectiveness (3 to 5), and low effectiveness (1-3). Behavior summary statements were written to summarize all of the behavioral information reflected in the individual examples. In particular, two or occasionally three behavior statements for each dimension and effectiveness level (i.e., high, medium, or low) were generated from the examples. Additional rationale for this behavior summary scale method can be found in Borman (1979).

As a final check on the behavior summary statements, we conducted a retranslation of the statements using the same procedure as was used with the individual examples. Seventeen en route controllers sorted each of the 87 statements into one of the dimensions and rated the effectiveness level reflected on a 1-7 scale. Results of this retranslation can be found in Pulakos, Keichel, Plamondon, Hanson, Hedge, and Borman (1996). Finally, for those statements either sorted into the wrong dimensions by 40% or more of the controllers *or* retranslated at an overly high or low effectiveness level, we

made revisions based on our analysis of the likely reason for the retranslation problem. The final behavior summary scales appear in Appendix C.

Regarding the rater orientation and training program, our experience and previous research has shown that the quality of performance ratings can be improved with appropriate rater training (e.g., Pulakos, 1984, 1986; Pulakos & Borman, 1986). Over the past several years, we have been refining a training strategy that (1) orients raters to the rating task and why the project requires accurate evaluations; (2) familiarizes raters with the rating dimensions and how each is defined; (3) teaches raters how to most effectively use the behavior summary statements to make objective ratings; (4) describes certain rater errors (e.g., halo) in simple, common-sense terms and asks raters to avoid them; and finally (5) encourages raters to be as accurate as possible in their evaluations.

For this application, we revised the orientation and training program to encourage accurate ratings in this setting. In particular, a script was prepared to be used by persons administering the rating scales in the field. Appendix D contains the script. In addition, a plan for gathering rating data was created. Discussions with controllers in the workshops described earlier suggested that both supervisors and peers (i.e., fellow controllers) would be appropriate rating sources. Because gathering ratings from relatively large numbers of raters per ratee is advantageous to increase levels of interrater reliability, we requested that two supervisor and two peer raters be asked to contribute ratings for each controller ratee in the study. Supervisor and peer raters were identified who had worked in the same area as a controller for at least 6 months and were very familiar with their job performance. For practical reasons we set a limit of 5-6 controllers to be rated by any individual rater in the research. The rater orientation and training program and the plan for administering the ratings in the field were incorporated into a training module for those professionals selected to conduct the data collection. That training session is described in a subsequent section.

The High-Fidelity Performance Measure (HFPM)

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

However, there have been some research and development efforts aimed at capturing the performance of air traffic controllers, including Buckley, O'Connor, Beebe, Adams, and MacDonald (1969), Buckley, DeBaryshe, Hitchner, and Kohn (1983), and Sollenberger, Stein, and Gromelski (1997). For example, in the Buckley et al. (1983) study, trained observers' ratings of simulator performance were found highly related to various aircraft safety and expeditiousness measures. Full-scale dynamic simulation allows the controller to direct the activities of a sample of simulated air traffic, performing characteristic functions such as ordering changes in aircraft speed or flight path, but within a relatively standardized work sample framework.

The intention of the HFPM was to provide an environment that would, as nearly as possible, simulate actual conditions existing in the controller's job. One possibility considered was to test each controller working in his or her own facility's airspace. This approach was eventually rejected, however, because of the problem of unequal difficulty levels across facilities and even across sectors within a facility (Borman, Hedge, & Hanson, 1992; Hanson, Hedge, Borman, & Nelson, 1993; Hedge, Borman, Hanson, Carter, & Nelson, 1993). Comparing the performance of controllers working in environments with unequal (and even unknown) difficulty levels is extremely problematic. Therefore, we envisioned that performance could be assessed using a "simulated" air traffic environment. This approach was feasible because of the availability at the FAA Academy of several training laboratories equipped with radar stations similar to those found in the field. In addition, they use a generic airspace (Aero Center) designed to allow presentation of typical air traffic scenarios that must be controlled by the trainee (or in our case, the ratee). Use of a generic airspace also allowed for standardization of assessment. See Figure 4.4 for a visual depiction of the Aero Center airspace.

Thus, through use of the Academy's radar training facility (RTF) equipment, in conjunction with the Aero Center generic airspace, we were able to provide a test environment affording the potential for both high stimulus and response fidelity. Our developmental efforts

focused, then, on: (1) designing and programming specific scenarios in which the controllers would control air traffic; and (2) developing measurement tools for evaluating controller performance.

Scenario Development

The air traffic scenarios were designed to incorporate performance constructs central to the controller's job, such as maintaining aircraft separation, coordinating, communicating, and maintaining situation awareness. Also, attention was paid to representing in the scenarios the most important tasks from the task-based job analysis. Finally, it was decided that, to obtain variability in controller performance, scenarios should be developed with either moderate or quite busy traffic conditions. Thus, to develop our HFPM scenarios, we started with a number of pre-existing Aero Center training scenarios, and revised and reprogrammed to the extent necessary to include relevant tasks and performance requirements with moderate- to high-intensity traffic scenarios. In all, 16 scenarios were developed, each designed to run no more than 60 minutes, inclusive of start-up, position relief briefing, active air traffic control, debrief, and performance evaluation. Consequently, active manipulation of air traffic was limited to approximately 30 minutes.

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

Development of Measurement Instruments

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

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

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

A second instrument required the raters to focus on more detailed behaviors and activities, and note whether and how often each occurred. A "Behavioral and Events Checklist" (BEC) was developed for use with each scenario. The BEC required raters to actively observe the ratees controlling traffic during each scenario and note behaviors such as: (1) failure to accept hand-offs, coordinate pilot requests, etc.; (2) letters of agreement (LOA)/directive violations; (3) readback/hearback errors; (4) unnecessary delays; (5) incorrect information input into the computer; and, (6) late frequency changes. Raters also noted operational errors and deviations. The BEC form can be found in Appendix F.

Rater Training

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

During their first week at the Academy, raters were exposed to (1) a general orientation to the AT-SAT project, its purposes and objectives, and the importance of the high-fidelity component; (2) airspace training; (3) the HFPM instruments; (4) all supporting materials (such as Letters of Agreement, etc.); (5) training and evaluation scenarios; and (6) rating processes and procedures. The training program was an extremely hands-on, feedback intensive process. During this first week raters served as both raters and ratees, controlling traffic in each scenario multiple times, as well as serving as raters of their associates who took turns as ratees. This process allowed raters to become extremely familiar with both the scenarios and evaluation of performance in these scenarios. With multiple raters evaluating performance in each scenario, project personnel were able to provide immediate critique and feedback to raters, aimed at improving accuracy and consistency of rater observation and evaluation.

In addition, prior to rater training, we "scripted" performances on several scenarios, such that deliberate errors were made at various points by the individual controlling traffic. Raters were exposed to these "scripted" scenarios early in the training so as to more easily facilitate discussion of specific types of controlling errors. A standardization guide was developed with the cooperation of the raters, such that rules for how observed behaviors were to be evaluated could be referred to during data collection if any questions arose (see Appendix G). All of these activities contributed to enhanced rater calibration.

Pilot Tests of the Performance Measures

The plan was to pilot test the CBPM and the performance rating program at two Air Route Traffic Control Centers (ARTCCs), Seattle and Salt Lake City. The HFPM was to be pilot tested in Oklahoma City. All materials were prepared for administration of the CBPM and ratings, and two criterion research teams proceeded to the pilot test sites. In general, procedures for administering these two assessment measures proved to be effective. Data were gathered on a total of 77 controllers at the two locations. Test administrators asked pilot test participants for their reactions to the CBPM, and many of them reported that the situations were realistic and like those that occurred on their jobs.

Results for the CBPM are presented in Table 4.4. The distribution of total scores was promising in the sense that there was variability in the scores. The coefficient alpha was moderate, as we might expect from a test that is likely mutidimensional. Results for the ratings are shown in Tables 4.5 and 4.6. First, we were able to approach our target of two supervisors and two peers for each ratee. A mean of 1.24 supervisors and 1.30 peers per ratee participated in the rating program. In addition, both the supervisor and peer ratings had reasonable degrees of variability. Also, the interrater reliabilities (intraclass correlations) were, in general, acceptable. The Coordinating dimension is an exception. When interrater reliabilities were computed across the supervisor and peer sources, they ranged from .37 to .62 with a median of .54. Thus, reliability improves when both sources' data are used.

In reaction to the pilot test experience, we modified the script for the rater orientation and training program. We decided to retain the Coordinating dimension for the main study, with the plan that if reliability continued to be low we might not use the data for that dimension. With the CBPM, one item was dropped because it had a negative item-total score correlation. That is, controllers who answered this item correctly tended to have low total CBPM scores.

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

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

In addition to this logistical, administration focus of the pilot test, we also examined the consistency of ratings by our HFPM raters. Two raters were assigned to each ratee, and the collection of HFPM data by two raters for each ratee across each of the seven scenarios allowed us to check for rater or scenario peculiarities.

Table 4.7 presents correlations between ratings for rater pairs both across scenarios and within each scenario, and suggested that Scenarios 2 and 7 should be examined more closely, as well as three OTS dimensions (Communicating Clearly, Accurately, and Efficiently; Facilitating Information Flow; and Coordination). To provide additional detail, we also generated a table showing magnitude of effectiveness level differences between each rater pair for each dimension on each scenario (see Appendix H).

Examination of these data and discussion with our raters helped us to focus on behaviors or activities in the two scenarios that led to ambiguous ratings and to subsequently clarify these situations. Discussions concerning these details with the raters also allowed us to identify specific raters in need of more training. Finally, extensive discussion surrounding the reasons for lower than expected correlations on the three dimensions generated the conclusion that excessive overlap between the three dimensions generated confusion as to where to represent the observed performance. As a result, the "Facilitating Information Flow" dimension was dropped from the OTS form.

Training the Test Site Managers

Our staff prepared a manual describing data collection procedures for the criterion measures during the concurrent validation and conducted a half-day training session on how to collect criterion data in the main sample. We reviewed the CBPM, discussed administration issues, and described procedures for handling problems (e.g., what to do when a computer malfunctions in mid-scenario). Test site managers had an opportunity to practice setting up the testing stations and review the beginning portion of the test. They were also briefed on the performance rating program. We described procedures for obtaining raters and training them. The script for training raters was thoroughly reviewed and rationale for each element of the training was provided. Finally, we answered all of the test site managers' questions. These test site managers hired and trained data collection staff at their individual testing locations. There were a total of 20 ARTCCs that participated in the concurrent validation study (both Phase 1 and Phase 2).

Data Collection

CBPM data were collected for 1046 controllers. Performance ratings for 1227 controllers were provided by 535 supervisor and 1420 peer raters. Table 4.8 below shows the number of supervisors and peers rating each controller. CBPM *and* rating data were available for 1043 controllers.

HFPM data were collected for 107 controllers. This sample was a subset of the main sample so 107 controllers had data for the CBPM, the ratings, *and* the HFPM. In particular, controllers from the main sample arrived in Oklahoma City from 12 different air traffic facilities throughout the U.S. to participate in the two-and one-half day HFPM process. The one-and one-half days of training consisted of four primary activities: orientation, airspace familiarization and review, airspace certification testing, and scenarios practice. To accelerate learning time, a hard copy and computer disk describing the airspace had been developed and sent to controllers at their home facility for "preread" prior to arrival in Oklahoma City.

Each controller was then introduced to the Radar Training Facility (RTF) and subsequently completed two practice scenarios. After completion of the second scenario and follow-up discussions about the experience, the controllers were required to take an airspace certification test. The certification consisted of 70 recall and recognition items designed to test knowledge of Aero Center. Those individuals not receiving a passing grade (at least 70% correct) were required to retest on that portion of the test they did not pass. The 107 controllers scored an average of 94% on the test, with only 7 failures (6.5%) on the first try. All controllers subsequently passed the retest and were certified by the trainers to advance to the remaining day of formal evaluation.

After successful completion of the air traffic test, each controller received training on six additional air traffic scenarios. During this time, the raters acted as trainers and facilitated the ratee's learning of the airspace. While questions pertaining to knowledge of airspace and related regulations were answered by the raters, coaching ratees on how to more effectively and efficiently control traffic was prohibited.

After the eight training scenarios were completed, all ratees' performance was evaluated on each of seven scenarios that together required approximately 8 hours to complete. The seven scenarios consisted of four moderately busy and three very busy air traffic conditions, increasing in complexity from Scenario 1 to Scenario 7. During this 8 hour period of evaluation, raters were randomly assigned to ratees before each scenario, with the restriction that a rater should not be assigned to a ratee (1) from the rater's home facility; or (2) if he/she was the rater's training scenario assignment.

While ratees were controlling traffic in a particular scenario, raters continually observed and noted performance using the BEC. After the scenario ended, each rater completed the OTS ratings. In all, 11 training/evaluation sessions were conducted within a 7-week period. During four of these sessions, a total of 24 ratees were evaluated by two raters at a time, while a single rater evaluated ratee performance during the other seven sessions.

Results

CBPM

Table 4.9 shows the distribution of CBPM scores. As with the pilot sample, there is a reasonable amount of variability. Also, item-total score correlations range from .01 to .27 (mean = .11). The coefficient alpha was .63 for this 84-item test. The relatively low item-total correlations and the modest coefficient alpha suggest that the CBPM is measuring more than a single construct.

Supervisor and Peer Ratings

In Tables 4.10 and 4.11, the number and percent of ratings at each scale point are depicted for supervisors and peers separately. A low but significant percentage of ratings are at the 1, 2, or 3 level for both supervisor and peer ratings. Most of the ratings fall at the 4-7 level, but overall, the variability is reasonable for both sets of ratings.

Table 4.12 contains the interrater reliabilities for the supervisor and peer ratings separately and for the two sets of ratings combined. In general, the reliabilities are quite high. The supervisor reliabilities are higher than the peer reliabilities, but the differences are for the most part very small. Importantly, the combined supervisor/peer ratings reliabilities are substantially higher than the reliabilities for either source alone. Conceptually, it seems appropriate to get both rating sources' perspectives on controller performance. Supervisors typically have more experience evaluating performance and have seen more incumbents perform in the job; peers often work side-by-side with the controllers they are rating, and thus have good first-hand knowledge of their performance. The result of higher reliabilities for the combined ratings makes an even more convincing argument for using both rating sources.

Scores for each ratee were created by computing the mean peer and mean supervisor rating for each dimension. Scores across peer and supervisor ratings were also computed for each ratee on each dimension by taking the mean of the peer and supervisor scores. Table 4.13 presents the means and standard deviations for these rating scores on each dimension, supervisors and peers separately, and the two sources together. The means are higher for the peers (range = 5.03-5.46), but the standard deviations for that rating source are generally almost as high as those for the supervisor raters.

Table 4.14 presents the intercorrelations between supervisor and peer ratings on all of the dimensions. First, within rating source, the between-dimension correlations are large. This is common with rating data. And second, the supervisor-peer correlations for the same dimensions (e.g., Communicating = .39) are at least moderate in size, again showing reasonable agreement across-source regarding the relative levels of effectiveness for the different controllers rated.

The combined supervisor/peer ratings were factor analyzed to explore the dimensionality of the ratings. This analysis addresses the question, is there a reasonable way of summarizing the 10 dimensions with a smaller number of composite categories? The 3-factor solution, shown in Table 4.15, proved to be the most interpretable. The first factor was called Technical Performance, with Dimensions, 1, 3, 6, 7, and 8 primarily defining the factor. Technical Effort was the label for Factor 2, with Dimensions 2, 4, 5, and 9 as the defining dimensions. Finally, Factor 3 was defined by a single dimension and was called Teamwork.

Although the 3-factor solution was interpretable, keeping the three criterion variables separate for the validation analyses seemed problematic. This is because (1) the variance accounted for by the factors is very uneven (82% of the common variance is accounted for

by the first factor); (2) the correlations between unit-weighted composites representing the first two factors is .78; correlations between each of these composites and Teamwork are high as well (.60 and .63 respectively); and (3) all but one of the 10 dimensions loads on a technical performance factor, so it seemed somewhat inappropriate to have the one-dimension Teamwork variable representing 1/3 of the rating performance domain.

Accordingly, we formed a single rating variable represented by a unit-weighted composite of ratings on the 10 dimensions. The interrater reliability of this composite is .71 for the combined supervisor and peer rating data. This is higher than the reliabilities for individual dimensions. This would be expected, but it is another advantage of using this summary rating composite to represent the rating data.

HFPM

Table 4.16 contains descriptive statistics for the variables included in both of the rating instruments used during the HFPM graded scenarios. For the OTS dimensions and the BEC, the scores represent averages across each of the seven graded scenarios.

The means of the individual performance dimensions from the 7-point OTS rating scale are in the first section of Table 4.16 (Variables 1 through 7). They range from a low of 3.66 for *Maintaining Attention and Situation Awareness* to a high of 4.61 for *Communicating Clearly, Accurately and Efficiently*. The scores from each of the performance dimensions are slightly negatively skewed, but are for the most part, normally distributed.

Variables 8 through 16 in Table 4.16 were collected using the BEC. To reiterate, these scores represent instances where the controllers had either made a mistake or engaged in some activity that caused a dangerous situation, a delay, or in some other way impeded the flow of air traffic through their sector. For example, a *Letter of Agreement (LOA)/Directive Violation* was judged to have occurred if a jet was not established at 250 knots prior to crossing the appropriate arrival fix or if a frequency change was issued prior to completion of a handoff for the appropriate aircraft. On average, each participant had 2.42 *LOA/Directive Violations* in each scenario.

Table 4.17 contains interrater reliabilities for the OTS Ratings for those 24 ratees for whom multiple rater information was available. Overall, the interrater reliabilities were quite high for the OTS ratings, with median interrater reliabilities ranging from a low of .83 for *Maintaining Attention and Situation Awareness* to a high of .95 for *Maintaining Separation*. In addition, these OTS dimensions were found to be highly intercorrelated (median r = .91). Because of the high levels of dimension intercorrelation, an overall composite will be used in future analyses.

All relevant variables for the OTS and BEC measures were combined and subjected to an overall principal components analysis to represent a final high-fidelity performance criterion space. The resulting two- factor solution is presented in Table 4.18. The first component, *Overall Technical Proficiency*, consists of the OTS rating scales, plus the operational error, operational deviation, and LOA/Directive violation variables from the BEC. The second component is defined by six additional BEC variables and represent a *sector management* component of controller performance. More specifically, this factor represents *Poor Sector Management*, whereby the controllers more consistently make late frequency changes, fail to accept hand-offs, commit readback/hearback errors, fail to accommodate pilot requests, delay aircraft unnecessarily, and enter incorrect information in the computer. This interpretation is reinforced by the strong negative correlation (-.72) found between *Overall Technical Proficiency* and *Poor Sector Management.*

Correlations Between the Criterion Measures: Construct Validity Evidence

Table 4.19 depicts the relationships between scores on the 84-item CBPM, the two HFPM factors, and the combined supervisor/peer ratings. First, the correlation between the CBPM total scores and the HFPM Factor 1, arguably our purest measure of technical proficiency, is .54. This provides strong evidence for the construct validity of the CBPM. Apparently, this lower fidelity measure of technical proficiency is tapping much the same technical skills as the HFPM, which had controllers working in an environment highly similar to their actual job setting. In addition, a significant negative correlation exists between the CBPM and the second HFPM factor, *Poor Sector Management*.

Considerable evidence for the construct validity of the ratings is also evident. Correlations between the ratings and the first HFPM factor is .40. Thus, the ratings, containing primarily technical proficiency-oriented content, correlate substantially with our highest fidelity measure of technical proficiency. The ratings

also correlate significantly with the second HFPM factor (r = -.28), suggesting the broad-based coverage of the criterion space toward which the ratings were targeted. Finally, the ratings-CBPM correlation is .22, suggesting that the ratings also share variance associated with the judgment, decision-making, and procedural knowledge constructs we believe the CBPM is measuring. This suggests that, as intended, the ratings on the first two categories are measuring the typical performance component of technical proficiency.

Overall, there is impressive evidence that the CBPM and the ratings are measuring the criterion domains they were targeted to measure. At this point, and as planned, we examined individual CBPM items and their relations to the other criteria, with the intention of dropping items that were not contributing to the desired relationships. For this step, we reviewed the item-total score correlations, and CBPM item correlations with HFPM scores and the rating categories. Items with very low or negative correlations with: (1) total CBPM scores; (2) the HFPM scores, especially for the first factor; and (3) the rating composite were considered for exclusion from the final CBPM scoring system. Also considered were the links to important tasks. The linkage analysis is described in a later section. Items representing one or more highly important tasks were given additional consideration for inclusion in the final composite. These criteria were applied concurrently and in a compensatory manner. Thus, for example, a quite low item-total score correlation might be offset by a high correlation with HFPM scores.

This item review process resulted in 38 items being retained for the final CBPM scoring system. The resulting CBPM composite has a coefficient alpha of .61 and correlates .61 and -.42 with the two HFPM factors, and .24 with the rating composite. Further, coverage of the 40 most important tasks is at approximately the same level, with all but one covered by at least one CBPM item. Thus, the final composite is related more strongly to the first HFPM factor, and correlates a bit more highly with the technically-oriented rating composite. We believe this final CBPM composite has even better construct validity in relation to the other criterion measures than did the total test.

Additional Construct Validity Evidence

Hedge et al. (1993) discuss controller performance measures that are currently collected and maintained by the FAA and the issues in using these measures as criteria in the validation of controller predictor measures. Some of the more promising archival measures are those related to training performance, especially the time to complete various phases of training and ratings of performance in these training phases. However, there are some serious problems even with these most promising measures (e.g., standardization across facilities, measures are not available for all controllers). Thus, our approach in the present effort was to use these measures to further evaluate the construct validity of the AT-SAT criterion measures.

In general, training performance has been shown to be a good predictor of job performance, so measures of training performance should correlate with the AT-SAT measures of job performance. Training performance data were available for 809 of the 1227 controllers in the concurrent validation sample. Two of the on-the-job training phases (Phase 6 and Phase 9) are reasonably standardized across facilities, so performance measures from these two phases are good candidates for use as performance measures. We examined the correlation between ratings of performance across these two phases and the correlations between five variables measuring training time (hours and days to complete training at each phase). The rating measures did not even correlate significantly with each other, and thus were not included in further analyses. Correlations between the training time variables were higher. Because the time variables appeared to be tapping similar performance dimensions, we standardized and added these measures to create a "training time" scale. Controllers with less than four out of the five variables measuring training time were removed from further analyses (N=751). Correlations between training time and ratings of performance are moderate (r = .23). The correlation with CBPM scores is small but also significant (.08; $p < .05$). Thus, the correlations with training time support the construct validity of the AT-SAT field criterion measures. (Sample sizes for the HFPM were too small to conduct these analyses.)

Linkage Analysis

A panel of 10 controller SMEs performed a judgment task with the CBPM items. These controllers were divided into three groups, and each group was responsible for approximately one third of the 40 critical tasks that were targeted by the CBPM. They reviewed each CBPM scenario and the items, and indicated which of these important tasks from the job

analysis were involved in each item. These ratings were then discussed by the entire group until a consensus was reached. Results of that judgment task appear in Table 4.20. For each task, the table shows the number of CBPM items that this panel agreed measured that task.

Similarly, 10 controller SMEs performed a judgment task with the seven HFPM scenarios. These controllers were divided into two groups, and each group was responsible for half of the scenarios. Each scenario was viewed in three 10-minute segments, and group members noted if a critical subactivity was performed. After the three 10-minute segments for a given scenario were completed, the group discussed their ratings and arrived at a consensus before proceeding to the next scenario. Results of these judgments can also be found in Table 4.20. In summary, 38 of the 40 critical subactivities were covered by at least a subset of the seven scenarios. On average, almost 25 subactivities appeared in each scenario.

Conclusions

The 38-item CBPM composite provides a very good measure of the technical skills necessary to separate aircraft effectively and efficiently on the "real job." The .61 correlation with the highly realistic HFPM (Factor 1) is especially supportive of its construct validity for measuring performance in the very important technical proficiency-related part of the job. Additional ties to the actual controller job are provided by the links of CBPM items to the most important controller tasks identified in the job analysis.

The performance ratings provide a good picture of the typical performance over time elements of the job. Obtaining both a supervisor and a peer perspective on controller performance provides a relatively comprehensive view of day-to-day performance. High interrater agreement across the two rating sources further strengthens the argument that the ratings are valid evaluations of controller performance.

Thus, impressive construct validity evidence is demonstrated for both the CBPM and the rating composite. Overall, we believe the 38-item CBPM and the rating composite represent a comprehensive and valid set of criterion measures.

CHAPTER 5.1

Field Procedures for Concurrent Validation Study

Lucy B. Wilson, Christopher J. Zamberlan, and James H. Harris
Caliber Associates

The concurrent validation data collection was carried out in 12 locations from May to July, 1997. Additional data were collected in 4 locations from March to May, 1998 to increase the sample size. Data collection activities involved two days of computer-aided test administration with air traffic controllers and the collection of controller performance assessments from supervisory personnel and peers. Each site was managed by a trained Test Site Manager (TSM) who supervised trained on-site data collectors, also known as Test Administrators (TAs). A subset of 100 air traffic controllers from the May-July sample (who completed both the predictor and criterion battery of testing and for whom complete sets of performance assessment information were available), was selected to complete the high fidelity criterion test at the Academy in Oklahoma City. See Chapter 4 for a description of this activity.

Criterion Measure Pretest

An in-field pretest of the computerized criterion measure and the general protocol to be used in the concurrent validation test was conducted in April, 1997. The en-route air traffic control centers of Salt Lake City, UT and Seattle, WA served as pretest sites. A trained TSM was on site and conducted the pretest in each location.

Field Site Locations

In 1997, the concurrent validation testing was conducted in 12 en-route air traffic control centers across the country. The test center sites were:

- Atlanta, GA
- Albuquerque, NM
- Boston, MA
- Denver, CO
- Ft. Worth, TX
- Houston, TX
- Jacksonville, FL
- Kansas City, MO
- Los Angeles, CA
- Memphis, TN
- Miami, FL
- Minneapolis, MN

The additional testing in 1998 ran in Chicago, Cleveland, Washington, DC, and Oklahoma City. The en-route centers of Chicago and Cleveland performed like the original AT-SAT sites, testing their own controllers. The en-route center at Leesburg, Virginia, which serves the Washington, DC area, tested their controllers as well as some from New York. At the Mike Monroney Aeronautical Center in Oklahoma City, the Civil Aeromedical Institute (CAMI), with the help of Omni personnel, tested controllers from Albuquerque, Atlanta, Houston, Miami, and Oakland. All traveling controllers were scheduled by Caliber with the help of Arnold Trevette in Leesburg and Shirley Hoffpauir in Oklahoma City.

Field Period

Data collection activities began early in the Ft. Worth and Denver Centers in May, 1997. The remaining nine centers came on line two weeks later. To ensure adequate sample size and diversity of participants, one additional field site — Atlanta — was included beginning in June 1997. The concurrent data collection activities continued in all locations until mid-July.

Of the four sites in 1998, Chicago started the earliest and ran the longest, for a little over two months beginning in early March. Washington, DC began simultaneously, testing and rating for just under two months. Cleveland and Oklahoma City began a couple of weeks into March and ended after about four and five weeks, respectively.

Selection and Training of Data Collectors

A total of 13 experienced data collection personnel were selected to serve as TSMs during the first data collection. One manager was assigned to each of the test centers and one TSM remained on call in case an emergency replacement was needed in the field.

All TSMs underwent an intensive 3-day training in Fairfax, VA from April 22 to 24, 1997. The training was led by the team of designers of the concurrent validation tests. The objective of the training session was three-fold:

• To acquaint TSMs with the FAA and the en route air traffic control environment in which the testing was to be conducted
• To familiarize TSMs with the key elements of the concurrent validation study and their roles in it
• To ground TSMs in the AT-SAT test administration protocol and field procedures.

A copy of the TSM training agenda is attached.

Each TSM was responsible for recruiting and training his or her on-site data collectors who administered the actual test battery. The TSM training agenda was adapted for use in training on-site data collectors. In addition to didactic instruction and role-playing, the initial test administrations of all on-site data collectors were observed and critiqued by the TSMs.

Three TSMs repeated their role in the second data collection. Because of the unique role of the fourth site in the second data collection (e.g., a lack of previous experience from the first data collection and three times as many computers, or "testing capability," as any other testing site), Caliber conducted a special, lengthier training for CAMI personnel in Oklahoma City before the second data collection began.

Site Set Up

TSMs traveled to their sites a week in advance of the onset of data collection activities. During this week they met with the en-route center personnel and the "Partner Pairs" assigned to work with them. The Partner Pairs were composed of a member of ATC management and the union representative responsible for coordinating the center's resources and scheduling the air traffic controllers for testing. Their assistance was invaluable to the success of the data collection effort.

TSMs set up and secured their testing rooms on site during this initial week and programmed five computers newly acquired for use in the concurrent validation. They trained their local data collectors and observed their first day's work.

Air Traffic Controller Testing

Up to five controllers could, and frequently were, tested on an 8-hour shift. Testing was scheduled at the convenience of the center, with most of the testing occurring during the day and evening shifts, although weekend shifts were included at the discretion of the site. Controllers were scheduled to begin testing at the same time. While Oklahoma City had the capacity to test 15 controllers at a time, it did not use its expanded capability and operated like every other five-computer site, for all intents and purposes.

At the beginning of the first day of the 2-day testing effort, the data collector reviewed the Consent Form with each participating controller and had it signed and witnessed. (See the appendix for a copy of the Consent Form.) Each controller was assigned a unique identification number through which all parts of the concurrent validation tests were linked.

The predictor battery usually was administered on the first day of controller testing. The predictor battery was divided into four blocks with breaks permitted between each block and lunch generally taken after completion of the second block.

The second day of testing could occur as early as the day immediately following the first day of testing or could be scheduled up to several weeks later. The second day of concurrent validation testing involved completion of the computerized criterion test, that is, the Computer Based Performance Measure (CBPM), and the Biographical Information Form. (See appendix for a copy of the Biographical Information Form.) At the end of the second day of testing, participating controllers were asked to give their social security numbers so that archival information (e.g., scores on Office of Personnel Management employment tests) could be retrieved and linked to their concurrent validation test results.

Supervisory Assessments

Participating controllers nominated two supervisory personnel and two peers to complete assessments of them as part of the criterion measurement. While the selection of the peer assessors was totally at the discretion of the controller, supervisory and administrative staff had more leeway in selecting the supervisory assessors (although not one's "supervisor of record") from the much smaller pool of supervisors in order to complete the ratings. Throughout the data collection period, supervisors and peers assembled in small groups and were given standardized instructions by on-site data collectors in the completion of the controller assessments. To the extent feasible, supervisors and peers completed assessments in a single session on all the controllers who designated them as their assessor. When the assessment form was completed, controller names were removed and replaced

by their unique identification numbers. The assessment forms were placed in sealed envelopes as a further means of protecting confidentiality.

During the second data collection, assessors sometimes viewed PDRI's "How To" video in lieu of verbal instruction. This was especially important at the five non-testing sites that had no TSMs or on-site data collectors (Albuquerque, Atlanta, Houston, Miami, and Oakland). The four testing sites employed the video much less frequently, if at all.

Record Keeping and Data Transmission

On-site data collectors maintained records of which controllers had participated and which tests had been completed. This information was reported on a daily basis to TSMs. Several times a week on-site data collectors transmitted completed test information (on diskettes) and hard copies of the Biographical Information and performance assessment forms to the data processing center in Alexandria, VA.

Site Shut Down

At the end of the data collection period, each site was systematically shut down. The predictor and criterion test programs were removed from the computers, as were any data files. Record logs, signed consent forms, unused test materials, training manuals and other validation materials were returned to Caliber Associates. Chicago, the last site of the second data collection effort, shut down on Monday, May 11, 1998.

CHAPTER 5.2

Development Of Pseudo-applicant Sample

Anthony Bayless, Caliber Associates

RATIONALE FOR PSEUDO-APPLICANT SAMPLE

Prior to becoming a Full Performance Level (FPL) controller, ATCSs have been previously screened on their entry-level OPM selection test scores, performance in one of the academy screening programs, and on-the-job training performance. Because of these multiple screens and stringent cutoffs, only the better performing ATCSs are retained within the air traffic workforce. For these reasons, the concurrent validation of the AT-SAT battery using a sample of ATCSs is likely to result in an underestimate of the actual validity because of restriction in range in the predictors. The goal of this part of the project, then, was to administer the AT-SAT predictor battery to a sample that more closely resembled the likely applicant pool than would a sample of ATCS job incumbents.

The purpose of including a pseudo-applicant (PA) sample in the validation study was to obtain variance estimates from an unrestricted sample (i.e., not explicitly screened on any prior selection criteria). Data collected from the PA study were used to statistically "correct" predictor scores obtained from the restricted, concurrent validation sample of ATCS job incumbents. This statistical correction was necessary because the validity of predictors is based on the strength of the relationship between the predictors and job performance criteria. If this relationship was assessed using only the restricted sample (i.e., FAA job incumbents who have already been screened and selected) without any statistical correction, the strength of the relationships between the predictors and job performance criteria would be underestimated.[1] This underestimation of the validity of the predictors might lead to an omission of an important predictor based on an inaccurate estimation of its validity. By using the PA data to obtain variance/covariance estimates from an unrestricted sample (i.e., a pool of subjects that more closely represents the potential range of applicants), the underestimation of predictor validity computed from the restricted sample can be corrected.

ATCS Applicant Pool

The administration of the AT-SAT predictor battery to a sample closely resembling the applicant pool required an analysis of the recent ATCS applicant pool. Therefore, the project team requested from the FAA data about recent applicants for the ATCS job. Because of a recent hiring freeze on ATCS positions, the latest background data available for ATCS applicants was from 1990 through part of 1992. Although the data were somewhat dated (i.e., 1990-1992), it did provide some indication of the characteristics that should be emulated in the PA sample. Based on a profile analysis provided by the FAA, relevant background characteristics of 36,024 actual applicants for FAA ATCS positions were made available. Table 5.2.1 provides a breakout of some pertinent variables from that analysis.

The data indicated that about 81% of applicants were male, 50% had some college education but no degree, and 26% had a bachelor's degree. A disconcerting fact from the OPM records was the large percentage of missing cases (51.3%) for the race/ethnicity variable. Information available for the race/ethnicity variable represented data from 17,560 out of 36,024 cases. Another issue of some concern was the age of the data provided. The latest data were at least four years old. Although it seems unlikely that the educational profile of applicants would have changed much over four years, it was more likely that the gender and the race/ethnicity profiles may have changed to some extent over the same period of time (i.e., more female and ethnic minority applicants).

[1] This underestimate is the result of decreased variation in the predictor scores of job incumbents; they would all be expected to score relatively the same on these predictors. When there is very little variation in a variable, the strength of its association with another variable will be weaker than when there is considerable variation. In the case of these predictors, the underestimated relationships are a statistical artifact resulting from the sample selection.

Because of the concern about the age of the applicant pool data and the amount of missing data for the race/ethnicity variable, a profile of national background characteristics was obtained from the U.S. Bureau of the Census. As shown in Table 5.2.2, 1990 data from the U.S. Bureau of the Census indicated the following national breakout for race/ethnicity:

Without more up-to-date and accurate data about the applicant pool, the national data were used to inform sampling decisions. Using the percentages provided above for race/ethnicity upon which to base preliminary sampling plans, we recommended a total sample size of at least 300 PAs be obtained assuming it followed the same distributional characteristics as the national race/ethnicity data.

Pseudo-Applicant Sample Composition and Characteristics

Again, the impetus for generating a PA sample was to administer the AT-SAT predictor battery to a sample that more closely resembled the likely applicant pool than would a sample of ATCS job incumbents. The project team decided to collect data from two different pools of PAs: one civilian and the other military. The civilian PA sample was generated using public advertisement and comprised the volunteers obtained from such advertisement. Because the sample size of the civilian PA sample was dependent on an unpredictable number of volunteers, a decision was made to also collect data from a military PA sample. The military PA sample afforded a known and large sample size and access to scores on their Armed Services Vocational Aptitude Battery (ASVAB) with their granted permission. Each of these two pools of PAs are described in the following two subsections.

Civilian Pseudo-Applicant Sample

Because the computer equipment with the predictor and criterion software was already set up at each of the 12 CV testing sites, public advertisements were placed locally around the CV testing sites to generate volunteers for the civilian PA sample. The goal for each testing site was to test 40 PAs to help ensure an adequate civilian PA sample size.

Public advertisement for the civilian PA sample was accomplished via several different methods. One method was to place classified advertisements in the largest local, metropolitan newspapers (and some smaller newspapers for those CV sites located away from major metropolitan areas). An example classified newspaper advertisement is shown in Figure 5.2.1. Another means of advertising the testing opportunity was to place flyers at locations in proximity to the testing site. For example, flyers were placed at local vocational technical schools and colleges/universities. An example flyer advertisement is shown in Figure 5.2.2. A third means of advertising the testing to civilian PAs was to publicize the effort via ATCS to their family, friends, and acquaintances.

When responding to any form of advertisement, potential civilian PAs were requested to call a toll-free number where a central scheduler/coordinator would screen the caller on minimum qualifications (i.e., US citizenship, ages between 17 and 30, AND at least 3 years of general work experience) and provide the individual with background about the project and the possible testing dates and arrival time(s). After a PA had been scheduled for testing, the scheduler/coordinator would contact the testing site manager for the relevant testing location and notify him/her so that the testing time slot could be reserved for a PA instead of an ATCS (for those sites testing PAs and ATCSs concurrently). The scheduler/coordinator would also mail a form letter to the newly scheduled PA indicating the agreed upon testing time and date, directions to the testing facility, and things to bring with them (i.e., driver's license and birth certificate or passport) for verification of age and citizenship.

Military Pseudo-Applicant Sample

Because of the uncertainty about being able to generate a sufficient PA sample from the civilian volunteers, it was decided to collect additional data from a military PA sample. Again, the military PA sample would afford a known sample size and access to their ASVAB scores which would prove useful for validation purposes. For these reasons, the FAA negotiated with the U.S. Air Force to test participants at Keesler A.F.B., Biloxi, Mississippi. The military PAs were students and instructors stationed at Keesler A.F.B. Predictor data were collected from approximately 262 military PAs of which 132 (50.4%) were currently enrolled in the Air Traffic Control School; 106 (40.5%) were students in other fields such as Weather Apprentice, Ground Radar Maintenance, and Operations Resource Management; and 24 (9.2%) were Air Traffic Control School instructors. Table 5.2.3 provides a breakout of gender and race/ethnicity by type of sample.

The data in 5.2.1 indicate that the civilian and military PA samples were very similar with respect to their gender and race/ethnicity profiles. In addition, both of the PA samples were more diverse than the ATCS sample and fairly similar to the 1990 U.S. Bureau of Census national breakdown (compare data of Table 5.2.1 to data of Table 5.2.2).

On-Site Data Collection

Pseudo-applicants were administered the predictor battery using the same testing procedures as followed for the ATCS CV sample. The only differences between the civilian and military PA sample data collection procedures were that:

1. civilians were tested with no more than four other testing participants at a time (due to the limited number of computers available at any one of the testing sites), whereas military PAs at Keesler A.F.B. were tested in large groups of up to 50 participants per session.
2. the replacement caps for select keyboard keys were not compatible with the rental computer keyboards and were unusable. Because of this problem, index cards were placed adjacent to each of the computer test stations informing the test taker of the proper keys to use for particular predictor tests. The use of the index cards instead of the replacement keys did not appear to cause any confusion for the test takers.

Test site administrators provided the PAs with a standardized introduction and set of instructions about the testing procedures to be followed during the computer-administered battery. During the introduction the administrators informed the PAs of the purpose of the study and any risks and benefits associated with participation in the study. The confidentiality of each participants' results were emphasized. In addition, participants were asked to sign a consent agreement attesting to their voluntary participation in the study, their understanding of the purpose of the study, the risks/benefits of participation, and the confidentiality of their results. For the military PAs, those who signed a Privacy Act Statement gave their permission to link their predictor test results with their ASVAB scores.

The testing volunteers were required to sacrifice one eight-hour day to complete the predictor battery. Although testing volunteers were not compensated for their time due to project budget constraints, they were provided with compensation for their lunch.

Correction for Range Restriction

As mentioned previously, the reason for collecting predictor data from PAs was to obtain variance estimates from individuals more similar to actual applicants for use in correcting validity coefficients for tests derived from a restricted sample (i.e., job incumbents). A description of the results of the range restriction corrections is contained in Chapter 5.5.

CHAPTER 5.3

Development of Database

Ani S. DiFazio
HumRRO

The soundness of the validity and fairness analyses conducted on the beta test data, and of the recommendations based on those results, was predicated on reliable and complete data. Therefore, database design, implementation, and management were of critical importance in validating the predictor tests and selecting tests for inclusion in Version 1 of the Test Battery. The Validation Analysis Plan required many diverse types of data from a number of different sources. This section describes the procedures used in processing these data and integrating them into a cohesive and reliable analysis database.

Data Collection Instruments

As described in section 5.1, data from computerized predictor and criterion tests were automatically written as ASCII files by the test software at the test sites. Depending on the test, the data were written either as the examinee was taking the test or upon completion of the test. The data file structure written by each test program was unique to that test. Each file represented an individual test taken by a single examinee. A complete battery of tests consisted of 13 computerized predictor tests as well as one computerized criterion test. For the first AT-SAT data collection (AT-SAT 1), high-fidelity criterion measures were also obtained on a subset of the controller participants.

In addition to the automated test data, several different types of data were collected by hard copy data collection instruments. These include three biographical information forms for controller participants, pseudo-applicant participants, and assessors, a Request of SSN for Retrieval of the Historical Archival Data form, and a Criterion Assessment Rating Assessment Sheet. The Validation Analysis Plan also called for the integration of historical archival data from the FAA.

Initial Data Processing

Automated Test Files

Data Transmittals. The automated test data collected at the 17 test sites were initially sent to HumRRO via Federal Express on a daily basis. This was done so that analysts could monitor test sites closely in the beginning of the test period and solve problems immediately as they arose. Once confident that a test site was following the procedures outlined in *the AT-SAT Concurrent Validation Test Administration Manual* and was not having difficulty in collecting and transmitting data, it was put on a weekly data transmittal schedule. Out of approximately seven and a half weeks of testing, the typical site followed a daily transmittal schedule for the first two weeks and then sent data on a weekly schedule for the remainder of the testing period. In total, HumRRO received and processed 297 Federal Express packages containing data transmittals from the 17 test sites.

The sites were provided detailed instructions on the materials to be included in a data transmittal packet. First, packets contained a diskette of automated test files for each day of testing.[2] Sites were asked to include a Daily Activity Log (DAL) if any problems or situations arose that might affect examinee test performance. Along with each diskette, the sites were required to submit a Data Transmittal Form (DTF)[3] which provided an inventory of the pieces of data contained in the transmittal packet. During the testing period, HumRRO received and processed 622 hard copy DTFs.

Data Processing Strategy. Because of the magnitude of data and the very limited time allocated for its processing, a detailed data processing plan was essential. The three main objectives in developing a strategy for processing the automated test data from the test sites were to —

[2] Some sites wrote the transmittal diskette at the end of the test day, while others cut the data at the end of a shift. In these cases, more than one diskette would be produced for each test day.

[3] While a DTF was supposed to be produced for each diskette transmitted, some sites sent one DTF covering a number of test days, and, conversely, more than one DTF describing a single diskette.

- Ensure that the test sites were transmitting all the data they were collecting and that no data were inadvertently falling through the cracks in the field.
- Closely monitor the writing and transmittal of data by the sites, so that problems would be quickly addressed before large amounts of data were affected.
- Identify *and* resolve problematic or anomalous files.

To accomplish these objectives, the test data were initially passed through two stages of data processing as testing was in progress. A third processing stage, described in the later subsection "Integration of AT-SAT Data," occurred after testing was completed and served to integrate the diverse data collected for this effort into a reliable and cohesive database.

During the testing period, up to four work stations were dedicated to processing data transmittal packets sent by the sites. One work station was reserved almost exclusively for preliminary processing of the packets. This "stage one" processing involved unpacking Federal Express transmittals, identifying obvious problems, date stamping and transcribing the DTF number on all hard copy data collection forms, summarizing AT-SAT 1 examinee demographic information for weekly reports, and ensuring that the data were passed on to the next stage of data processing.

The "stage two" data processors were responsible for the initial computer processing of the test data. Their work began by running a Master Login procedure that copied the contents of each diskette transmitted by the test sites onto the work station's hard drive. This procedure produced a hard copy list of the contents of the diskette and provided a baseline record of all the data received from the sites.[4] Next, using a key entry screen developed solely for this application, information on participant data from each DTF was automated and Statistical Analysis System (SAS) DTF files were created.[5]

This "stage two" automation of DTF hard copy forms served both record keeping and quality assurance functions. To gauge whether the sites were transmitting all the data they collected, the inventory of participant predictor and CBPM test data listed on the DTF was compared electronically to the files contained on the diskette being processed.[6] Whenever there was a discrepancy, the data processing software developed for this application automatically printed a report listing the names of the discrepant files. Discrepancies involving both in fewer and more files recorded on the diskettes than expected from the DTF were reported. Test site managers/administrators were then contacted by the data processors to resolve the discrepancies. This procedure identified files that test sites inadvertently omitted in the data transmittal package.[7]

As helpful as this procedure was in catching data that may have been overlooked at sites, it was able to identify missing files only if the DTF indicated that they should not be missing. The procedure would not catch files that were never listed on the DTF. It was clear that this sort of error of omission was more likely to occur when large amounts of data were being collected at sites. While the second AT-SAT data collection (AT-SAT 2) tested just over 300 participants, AT-SAT 1 included over four and a half times that number. Therefore, if this type of error of omission was going to occur, it would likely occur during the first AT-SAT data collection rather than the second. To avoid this error, the AT-SAT 1 test site managers needed to assess the completeness of the data sent for processing against other records maintained at

---

[4] The Master Login software did not copy certain files, such as those with zero bytes.

[5] In automating the DTF, we wanted one DTF record for each diskette transmitted. Because sites sometimes included the information from more than one diskette on a hard copy DTF, more than one automated record was created for those DTFs. Conversely, if more than one hard copy DTF was transmitted for a single diskette, they were combined to form one automated DTF record.

[6] This computerized comparison was made between the automated DTF and an ASCII capture of the DOS directory of the diskette from the test site. The units of analysis in these two datasets were originally different. Since a record in the directory capture data was a file (i.e., an examinee/test combination), there was more than one record per examinee. An observation in the original DTF file was an examinee, with variables indicating the presence (or absence) of specific tests. In addition, the DTF inventoried predictor tests in four testing blocks rather than as individual tests. Examinee/test-level data were generated from the DTF by producing dummy electronic DTF records for each predictor test that was included in a test block that the examinee took. Dummy CBPM DTF records were also generated in this manner. By this procedure, the unit of analysis in the automated DTF and DOS directory datasets was made identical and a one-to-one computerized comparison could be made between the DTF and the data actually received.

[7] Conversely, this procedure was also used to identify and resolve with the sites those files that appeared on the diskette, but not on the DTF.

the site, such as the Individual Control Forms. Approximately three quarters into the AT-SAT 1 testing period, the data processors developed a table for each site that listed examinees by the types of data[8] that had been received for them. A sample of this table and the cover letter to test site managers is provided in Appendix I. The site managers were asked to compare the information on this table to their Individual Control Forms and any other records maintained at the site. The timing of this exercise was important because, while we wanted to include as many examinees as possible, the test sites still had to be operational and able to resolve any discrepancies discovered. The result of this diagnostic exercise was very encouraging. The only type of discrepancy uncovered was in cases where the site had just sent data that had not yet been processed. Because no real errors of omission were detected and since AT-SAT 2 involved fewer cases that AT-SAT 1, this diagnostic exercise was not undertaken for AT-SAT 2.

Further quality assurance measures were taken to identify and resolve any systematic problems in data collection and transmission. Under the premise that correctly functioning test software would produce files that fall within a certain byte size range and that malfunctioning software would not, a diagnostic program was developed to identify files that were too small or too big, based on "normal" ranges for each test. The objective was to avoid pervasive problems in the way that the test software wrote the data by reviewing files with suspicious byte sizes as they were received. To accomplish this, files with anomalous byte sizes and the pertinent DALs were passed on to a research analyst for review. A few problems were identified in this way. Most notably, we discovered that the software in the Scan predictor test stopped writing data when the examinee did not respond to test items. Also, under some conditions, the Air Traffic Scenarios test software did not write data as expected; investigation indicated that the condition was rare and that the improperly written data could, in fact, be read and used, so the software was not revised. No other systematic problems in the way the test software wrote data were identified.

This procedure was also one way to identify files with problems of a more idiosyncratic nature. The identification of file problems by the data processors was typically based on improper file name and size attributes. In some cases, the sites themselves called attention to problems with files whose attributes were otherwise normal. In most cases, the problem described by the site involved the use of an incorrect identification number for an examinee in the test start-up software. A number of other situations at the test sites led to problematic files, such as when a test administrator renamed or copied a file when trying to save an examinee's test data in the event of a system crash. Very small files or files containing zero bytes would sometimes be written when an administrator logged a participant onto a test session and the examinee never showed up for the test. In the first few weeks of testing, a number of files used by test site managers to train administrators had then been erroneously transmitted to the data processors. It is important to note that the contents of the test files were not scrutinized at this stage of processing.

The "stage two" processors recorded each problem encountered in a Problem Log developed for this purpose. The test site manager or administrator was then contacted and the test site and data processor worked together to identify the source of the problem. This approach was very important because neglected systematic data collection and transmittal issues could have had far-reaching negative consequences. Resolution of the problem typically meant that the test site would retransmit the data, the file name would be changed according to specific manager/administrator instructions, or the file would be excluded from further processing. For each problem identified, stage two data processors reached a resolution with the test sites, and recorded that resolution in the processor's Problem Log.

Once all of these checks were made, data from the test sites were copied onto a ZIP[9] disk. Weekly directories on each ZIP disk contained the test files processed during a given week for each stage two work station. The data in the weekly directories were then passed on for "stage three" processing. To ensure that only non-problematic files were retained on the ZIP disks and that none were inadvertently omitted from further processing, a weekly reconciliation was performed that compared all the test files processed during the week (i.e., those copied to the work station's hard drive by the Master Login procedure) to the files written on the week's ZIP disk. A computer

[8] This table reported whether predictor and CBPM test data, participant biographical information forms, and SSN Request Forms had been received.

[9] ZIP disks are a virtually incorruptible data storage medium that hold up to 100 megabytes of data.

application was written that automatically generated the names of all the discrepant files between these two sources.

Every week, each stage two data processor met with the database manager to discuss these discrepancies. The data processor had to provide either a rationale for the discrepancy or a resolution. The most typical rationale was that the data processor was "holding out" a file or waiting for the re-issuance of a problem file from the test site. Meticulous records were kept of these "hold-out" files and all were accounted for before the testing periods were completed. Resolutions of discrepancies typically included deletion or addition of files or changes to file names. In these cases, the database manager handled resolutions and the reconciliation program was re-executed to ensure accuracy. These procedures resulted in a total of 23,107 files[10] written onto ZIP disk at the conclusion of stage two processing for AT-SAT 1 and 2 combined.

So as not to waste analysis time during AT-SAT 1, raw CBPM test files contained on weekly ZIP disks were sent to PDRI on a weekly basis during the testing period, along with the DALs and lists of files with size problems. During AT-SAT 2, CBPM files were sent to PDRI at the end of the testing period; DALs and DTFs were sent to PDRI directly from the sites. Similarly, Analogies (AN), Planes (PL), Letter Factory (LA), and Scan (SC) raw test files were sent to RGI on a weekly basis during AT-SAT 1 and at the end of the testing period for AT-SAT 2. At the end of the AT-SAT 1 testing period, all the collected data for each of these tests were re-transmitted to the appropriate organization, so that the completeness of the cumulative weekly transmittals could be assessed against the final complete transmittal.

HumRRO wrote computer applications that read the raw files for a number of predictor tests. These tests, which contained multiple records per examinee, were reconfigured into ASCII files with a single record for each participant for each test. SAS files were then created for each test from these reconfigured files. This work was performed for the following tests: Applied Math (AM), Dials (DI), Memory 1 (ME), Memory 2 (MR), Sound (SN), Angles (AN), Air Traffic Scenarios (AT), Time Wall (TW), and the Experience Questionnaire (EQ). At the conclusion of testing, the reconfigured EQ data were sent to PDRI for scoring and analysis.

Hard Copy Data

Data Handling of Participant Biographical Data and Request for SSN Forms. As mentioned above, stage one processors handled the data transmittal packages from the test sites. Once each hard copy form had been date stamped, these processors passed the participant biographical forms and SSN Request Forms to stage two processors. Here, as in the processing of automated test data, to ensure that all the data indicated on the DTF had been sent, a report printed by the DTF automation program listed all the hard copy participant forms that the DTF indicated should be present for an examinee. The stage two data processors were then required to find the hard copy form and place a check mark in the space provided by the reporting program. As with the automated test data, all problems were recorded in the data processor's Problem Log and the test sites were contacted for problem resolution.

Data Handling of Assessor Biographical Data and Criterion Assessment Rating Sheets*:* As discussed earlier, the automated DTF file contained information recorded on the first page of the DTF form describing the participant data transmitted from the site. The second page of the hard copy DTF contained information on assessor data—specifically, whether a Confidential Envelope, which contained the Criterion Rating Assessment Sheet(s) (CARS), and an Assessor Biographical Form were present in the data transmittal package. HumRRO handled assessor biographical data and the Criterion Rating Assessment Sheets during AT-SAT 1; these hard copy instruments were processed by PDRI during AT-SAT 2. As with other types of data, to ensure that all collected assessor information was actually transmitted, stage one processors compared the assessor data contained in each data transmittal package to the information contained on the DTF. Test sites were informed of all discrepancies by e-mailed memoranda or telephone communication and were asked to provide a resolution for each discrepancy. Because the assessors were often asked to provide CARS ratings and complete the Assessor Biographical Data Form at the same time, they often included the biographical form in the Confidential Envelope along with the CARS. As a consequence, the test site administrator did not have first-hand knowledge of which forms were contained in the envelopes. In processing the hard copy assessor data, there were a total

10 The 23,107 files were comprised of the CBPM test, the 13 predictor tests, and one start-up (ST) file for controller examinees and 13 predictor tests, and one start-up (ST) file for pseudo-applicants.

of 29[11] assessor discrepancies between the data actually received and the data the DTF indicated should have been received. Of these 29, only four discrepancies could not be resolved. In these instances the assessor simply may not have included in the Confidential Envelope the forms that the administrator thought were included.

Data Automation. Hard copy forms that passed through to stage two processing were photocopied and the originals filed awaiting automation. Since there were no other copies of these data, photocopies insured against their irrevocable loss, particularly once they were sent to key-punch. All original and photocopied Request for SSN Forms were stored in a locked cabinet. Five separate ASCII key entry specifications were developed by the AT-SAT database manager: for the three biographical data instruments, the CARS form, and the Request for SSN Form. The database manager worked closely with the data automation company chosen to key enter the data. The data were double-keyed to ensure accuracy. Once the data were keyed and returned, the total number of cases key entered were verified against the total number of hard copy forms sent to key-punch. Data were sent to key-punch in three installments during the course of AT-SAT 1 testing; a small fourth installment comprised of last minute "stragglers" was keyed in-house. CAR and assessor biographical AT-SAT 2 data were sent to key-punch in two installments during testing and a small third installment of "stragglers" was keyed in-house by PDRI. In AT-SAT 1, automated files containing assessor and participant biographical data and criterion ratings data were sent to PDRI a few times during the course of testing; complete datasets were transmitted when testing was concluded.

Historical Data

Confidentiality of test participants was a primary concern in developing a strategy for obtaining historical data from the FAA computer archives and linking that data to other AT-SAT datasets. Specifically, the objective was to ensure that the link between test examinees and controllers was not revealed to the FAA, so that test results could never be associated with a particular employee. Also, although the FAA needed participant controller Social Security Numbers (SSN) to identify and extract cases from their historical archives, these SSNs could not be returned once the historical information had been extracted. Therefore, examinee number or SSN could not be used as the link between records in the historical data and the other AT-SAT data collected. To overcome this problem, a unique random identification number was generated for each controller examinee who submitted a Request for SSN form in AT-SAT 1 and 2. Electronic files containing the SSN, this random identification number, and site number were sent to the FAA. Of the 986 controllers who submitted a Request for SSN Form, 967 had non-missing SSNs that could be linked to the FAA archival data. In addition to these 967 SSNs, the FAA received 4 SSN Forms during the high fidelity testing in Oklahoma City, which increased the number of cases with historical data to 971.

Pseudo-Applicant ASVAB Data

AFQT scores and composite measures of ASVAB subtests G (General), A (Administrative), M (Mechanical), and E (Electronic) were obtained for Kessler pseudo-applicants and merged with test and biographical data during stage three data processing.

Integration of AT-SAT Data

The goal in designing the final AT-SAT database was to create a main dataset that could be used to address most analytic needs, with satellite datasets providing more detailed information in specific areas. Before the database could be created, data processors needed to perform diagnostic assessments of the accuracy of the data and edit the data on the basis of those assessments. "Stage three" data processing activities included these diagnostic data checks and edits, as well as data merging and archive.

Data Diagnostics and Edits

Since the data contained on the test files were written by test software that was generally performing as expected, there were no errors in data recordation, and therefore no need for large-scale data editing. There were two types of diagnostic checks to which the test files were subjected, however. First, a check was made to see whether an examinee had taken the same test more than once. It is a testament to the diligent work of the test sites and the data processors that this anomaly was not evident

[11] The total number of assessor discrepancies e-mailed to sites was 41. For 12 participant assessors, the test administrator indicated the presence of an assessor biographical form on the DTF when a participant biographical form had actually been completed. Therefore, the number of true assessor discrepancies was 29.

in the data. Second, the test analysts performed diagnostics to identify observations that might be excluded from further analysis, such as those examinees exhibiting motivational problems. Obviously, historical data from the FAA archives were not edited. Data collected on hard copy instruments were subjected to numerous internal and external diagnostic and consistency checks and programmatic data editing. A primary goal in data editing was to salvage as much of the data as possible without jeopardizing accuracy.

Participant Biographical Data. Several different types of problems were encountered with the participant biographical data:

- More than one biographical information form completed by the same participant
- Missing or out-of-range examinee identification number
- Out-of-range date values

First, to correct the problem of duplicate[12] biographical forms for the same examinee, all forms completed after the first were deleted. Second, information from the DTF sent with the biographical form often made it possible to identify missing examinee numbers through a process of elimination. Investigation of some out-of-range examinee numbers revealed that the digits had been transposed at the test site. Third, out-of-range date values were either edited to the known correct value or set to missing when the correct value was unknown.

Other data edits were performed on the controller and pseudo-applicant participant biographical data. A number of examinees addressed the question of racial/ethnic background by responding "Other" and provided open-ended information in the space allowed. In many cases, the group affiliation specified in the open-ended response could be re-coded to one of the five specific alternatives provided by the item (i.e., Native American/Alaskan Native, Asian/Pacific Islander, African American, Hispanic, or Non-Minority). In these cases, the open-ended responses were recoded to one of the close-ended item alternatives. In other cases, a sixth racial category, mixed race, was created and applicable open-ended responses were coded as such.

Two types of edits were applicable only to the controller sample. First, in biographical items that dealt with the length of time (months and years) that the controller had been performing various duties, when only the month or year component was missing, the missing item was coded as zero. Also, for consistency, year was always made to be included in the year, rather than month (e.g., 24 months), field. When year was reported in the month field, the year field was incremented by the appropriate amount and the month field re-coded to reflect any remaining time less than that year(s).

Second, a suspiciously large group of controller participants reported their race as "Native American/Alaskan Native" on the biographical form. To check the accuracy of self-reported race, the responses were compared to the race/ethnic variable on the historical FAA archive data. For those controllers with historical data, racial affiliation from the FAA archives was used rather than self-reported race as a final indication of controller race. The following frequencies of race from these two sources of information show some of the discrepancies (Source 1 represents self-reported race from biographical form only, and Source 2 represents race based on archival race when available and self reported race, when it was not). Using Source 1, there were 77 Native American/Alaskan, compared to 23 using Source 2. Similarly there were 9 and 7 Asian/Pacific Islander respectively (Source 1 is always given first), 95 and 98 African Americans, 64 and 61 Hispanic, 804 and 890 Non-Minority, 20 and 8 Other , and 4 and 1 Mixed Race. This gives a total of 1073 participants by Source 1 and 1088 by Source 2, with 159 Source 1 and 144 missing Source 2 data. (Counts for Other were produced after "Other" was re-coded into one of the five close-ended specified item alternatives whenever possible.)

All edits were performed programmatically, with hard copy documentation supporting each edit maintained in a separate log. In 33 cases, participant assessors completed only assessor rather than participant biographical forms. In these cases, biographical information from the assessor form was used for participants.

Assessor Biographical Data. Like the participant data, the assessor biographical data required substantial data cleaning. The problems encountered were as follows:

- More than one biographical information form completed by the same assessor
- Incorrect assessor identification numbers
- Out-of-range date values

[12] The word "duplicate" here does not necessarily mean identical, but simply that more than one form was completed by a single participant. More often than not, the "duplicate" forms completed by the same participant were not identical.

First, the same rule formulated for participants, deleting all duplicate biographical records completed after the first, was applied. Second, by consulting the site Master Rosters and other materials, misassigned or miskeyed[13] rater identification numbers could be corrected. Third, out-of-range date values were either edited to the known correct value (i.e., the year that all biographical forms were completed was 1997) or set to missing when the correct value was unknown.

In addition to data corrections, the race and time fields in the assessor data were edited following the procedures established in the participant biographical data. Open-ended responses to the racial/ethnic background item were re-coded to a close-ended alternative whenever possible. In addition, when only the month or year component in the "time" fields was missing, the missing item was coded as zero. When full years were reported in the month field (e.g., 24 months), the year field was incremented by the appropriate amount and the month field re-coded to reflect any remaining time less than a year.

Since the test sites were instructed to give participants who were also assessors a participant, rather than assessor, biographical form, data processors also looked for biographical information on raters among the participant data. Specifically, if an assessor who provided a CARS for at least one participant did not have an assessor biographical form, participant biographical data for that assessor were used, when available

Criterion Ratings Data. Of all the hard copy data collected, the CARS data required the most extensive data checking and editing. Numerous consistency checks were performed within the CARS dataset itself (e.g., duplicate rater/ratee combinations), as well as assessing its consistency with other datasets (e.g., assessor biographical data). All edits were performed programmatically, with hard copy documentation supporting each edit maintained in a separate log. The following types of problems were encountered:

- Missing or incorrect examinee/rater numbers
- Missing rater/ratee relationship
- Duplicate rater/ratee combinations
- Rater/ratee pairs with missing or outlier ratings or involved in severe DAL entries
- Out-of-range date values

First, the vast majority of missing or incorrect identification numbers and/or rater/ratee relationships were corrected by referring back to the hard copy source and/or other records. In some cases the test site manager was contacted for assistance. Since the goal was to salvage as much data as possible, examinee/rater numbers were filled in or corrected whenever possible by using records maintained at the sites, such as the Master Roster. Problems with identification numbers often originated in the field, although some key-punch errors occurred despite the double-key procedure. Since examinee number on a CARS record was essential for analytic purposes, six cases were deleted where examinee number was still unknown after all avenues of information had been exhausted.

Second, some raters provided ratings for the same examinee more than once, producing records with duplicate rater/ratee combinations. In these cases, hard copy sources were reviewed to determine which rating sheet the rater had completed first; all ratings produced subsequently for that particular rater/ratee combination were deleted.

Third, some cases were deleted based on specific direction from data analysts once the data had been scrutinized. These included rater/ratee combinations with more than 3 of the 11 rating dimensions missing, outlier ratings, ratings dropped due to information in the Problem Logs, or incorrect assignment of raters to ratees (e.g., raters who had not observed ratees controlling traffic). Fourth, CARS items that dealt with the length of time (months and years) that the rater had worked with the ratee were edited, so that when only the month or year component was missing, the missing item was coded as zero. Where full years were reported in the month field, the year field was incremented and the month field re-coded to reflect any remaining time.

AT-SAT Database

As stated above, the database management plan called for a main AT-SAT dataset that could address most analytic needs, with satellite datasets that could provide detailed information in specific areas. The AT-SAT Database, containing data from the alpha and beta tests, is presented in Figure 5.3.1. To avoid redundancy, datasets that are completely contained within other datasets are not presented separately in the AT-SAT

[13] The miskeying was often the result of illegible handwriting on the hard copy forms.

Database. For example, since participant biographical data is completely contained in the final summary dataset, it is not provided as a separate satellite dataset in the AT-SAT Database. Similarly, since the rater biographical data contains all the data recorded on the assessor biographical form, as well as some participant forms, the assessor biographical form is not listed as a separate dataset in the AT-SAT Database. All data processing for the AT-SAT Database was done in the Statistical Analysis System (SAS). The datasets contained in the archived AT-SAT Database were stored as portable Statistical Package for the Social Sciences (SPSS) files.

Alpha Data. The Alpha data consist of a summary dataset as well as scored item level test data from the Pensacola study conducted in the spring of 1997. Scored test data and biographical information are stored in the summary dataset called "SUMMARY.POR". Item level scored test data are contained in 14 individual files named "xx_ITEMS.POR", where xx is the predictor test acronym; an additional 15th file called AS_ITEMS.POR contains ASVAB test scores.

Beta Test Data. The Final Analytic Summary Data file in the AT-SAT database is comprised of a number of different types of data:

- Subset of scored test variables
- Complete historical FAA archive data
- Participant biographical information
- ASVAB data for Keesler participants
- Information on rater identification numbers

As stated previously, HumRRO, RGI, and PDRI were each responsible for developing and analyzing specific tests in the beta test battery. The results of these analyses are presented in detail elsewhere in this report. Once the tests had been scored, each organization returned the scored item-level data to the AT-SAT database manager. Salient scored variables were extracted from each of these files and were linked together by examinee number. This created an examinee-level dataset with a single record containing test information for each examinee. Participant biographical data and historical FAA archive data were merged to this record, also by examinee number. For Keesler pseudo-applicants, ASVAB data were added. Participants for whom at least one CARS had been completed also had variable(s) appended to their main record containing the identification number of their assessor(s), so that examinee-level and assessor-level data can be easily linked. Test variable names always begin with the two letter test acronym; the names of biographical items in this data file begin with "BI".

This main analysis dataset is called XFINDAT5.POR and contains 1,752 cases with 1,466 variables.[14]

The satellite test and rating data in the AT-SAT Database are comprised of three types of files. The first group consists of the 23,107 raw ASCII examinee test (predictor and CBPM) files stored in weekly data processing directories. The processing of these data is described in the subsection, Initial Data Processing, Automated Test Files. These raw files are included in the AT-SAT Database primarily for archival purposes. Second, there is the electronic edited version of the CARS hard copy data, called CAR.POR, which is described in the subsection, Initial Data Processing, Hard Copy Data. This file is also included in the AT-SAT Database mainly for purposes of data archive. The third group of files contains complete scored item-level test data for examinees, derived from the first two types of data files listed above. The predictor scored item-level files (e.g., EQ_ITEM.POR, AM_ITEMS.POR) were derived from the raw ASCII predictor test files; the criterion file (CR_ITEMS.POR) was derived from raw CBPM test files and the CAR data.[15] Salient variables from these scored item-level test files constitute the test data in the analytic summary file XFINDAT5.POR.

Biographical Data were also included in the beta test datasets. Complete examinee biographical data are contained in the analytic summary file XFINDAT5.POR and are, therefore, not provided as a separate file in the database. Biographical information on assessors only and participant assessors is contained in the dataset called XBRATER.POR and is described in the subsection, Initial Data Processing, Hard Copy Data.

Data Archive. The AT-SAT database described above is archived on CD-ROM. Figure 5.3.2 outlines the directory structure for the AT-SAT CD-ROM data archive. The root directory contains a README.TXT

[14] The following FAA-applied alphanumeric variables were assigned an SPSS system missing value when the original value consisted of a blank string: CFAC, FAC, FORM, IOPT, OPT, ROPT, STATSPEC, TTYPE , and @DATE. The following FAA-supplied variables were dropped since they contained missing values for all cases: REG, DATECLRD, EOD, FAIL16PF, P_P, and YR.

[15] This file also contains scored High Fidelity test data.

file that provides a brief description of the t; it also contains two subdirectories. The first subdirectory contains Alpha data, while the second contains data for the Beta analysis. Within the Alpha subdirectory, there are two subdirectories, "Final Summary Data" and "Examinee Item Level Scored Data", each of which contain data files. The Beta subdirectory contains the following subdirectories:

- Edited Criterion Assessment Rating Sheets
- Edited Rater Biodata Forms
- Examinee Item Level Scored Test Data
- Final Analytic Summary Data
- Raw Examinee Test Data in Weekly Subdirectories
- Scaled, Imputed, and Standardized Test Scores

Each Beta subdirectory contains data files. In addition, the "Final Analytic Summary Data" subdirectory contains a codebook for XFINDAT5.POR. The codebook consists of two volumes that are stored as Microsoft Word files CBK1.DOC and CBK2.DOC. The CBK1.DOC file contains variable information generated from an SPSS SYSFILE INFO. It also contains a Table of Contents to the SYSFILE INFO for ease of reference. The CBK2.DOC file contains frequency distributions for discrete variables, means for continuous data elements, and a Table of Contents to these descriptive statistics.[16]

[16] Means were generated on numeric FAA-generated historical variables unless they were clearly discrete.

CHAPTER 5.4

Biographical and Computer Experience Information: Demographics for the Validation Study

Patricia A. Keenan, HumRRO

This chapter presents first, the demographic characteristics of the participants in both the concurrent validation and the pseudo-applicant samples. The data on the controller sample are presented first, followed by the pseudo-applicant information. The latter data divided between civilian and military participants. It should be noted that not all participants answered each question in the biographical information form, so at times the numbers will vary or cumulative counts may not total 100%.

TOTAL SAMPLE

Participant Demographics

A total of 1,752 individuals took part in the study (incumbents and pseudo-applicants); 1,265 of the participants were male (72.2%) and 342 were female (19.5%). 145 participants did not indicate their gender; 149 did not identify their ethnicity. The cross-tabulation of ethnicity and gender, presented in Table 5.4.1, represents only those individuals who provided complete information about both their race and gender.

The sample included incumbent FAA controllers, supervisors and staff (Controller sample) as well as pseudo–applicants from Keesler Air Force base (Military PA sample) and civilian volunteers from across the country (Civilian PA sample). The pseudo-applicants were selected based on demographic similarity to expected applicants to the controller position. The estimated average age of the total sample was 33.14 years (SD = 8.43). Ages ranged from 18 to 60 years. This number was calculated based on the information from 1,583 participants; 169 people did not provide information about their date of birth and were not included in this average.

Participants were asked to identify the highest level of education they had received. Table 5.4.2 presents a breakdown of the educational experience for all participants. (151 people did not provide information about their educational background.) The data were collected at 18 locations around the U.S. Table 5.4.3 shows the number of participants who tested at each facility.

CONTROLLER SAMPLE

Participant Demographics

A total of 1,232 FAA air traffic controllers took part in the concurrent validation study. 912 controllers were male (83.7%), 177 controllers were female (16.3%). 143 participants did not specify their gender so their participation is not reflected in analyses. The majority of the data was collected in 1997. A supplementary data collection was conducted in 1998 to increase the minority representation in the sample. A total of 1,081 controllers participated in the 1997 data collection; 151 additional controllers participated in 1998. Table 5.4.4 shows the cross-tabulation of race and gender distribution for the 1997 and 1998 samples, as well as the combined numbers across both years. 143 individuals did not report their gender and 144 did not report their race. These individuals are not reflected in Table 5.4.4. The average age of the controllers was 37.47 (SD = 5.98), with ages ranging from 25 to 60 years. The mean was based on information provided by 1,079 of the participants; age could not be calculated for 153 participants.

Also of interest was the educational background of the controllers. Table 5.4.5 shows the highest level of education achieved by the respondents. No information on education was provided by 145 controllers.

Professional Experience

The controllers represented 17 enroute facilities. The locations of the facilities and the number of controller participants at each one are shown in Table 5.4.6. A total of 1,218 controllers identified the facility at which they are assigned; 14 did not identify their facility.

One goal of the study was to have a sample composed of a large majority of individuals with air traffic experience, as opposed to supervisors or staff personnel. For this reason, participants were asked to identify both their current and previous positions. This would allow us to identify everyone who had current or previous experience in air traffic control. Table 5.4.7 indicates the average number of years the incumbents in each job

category had been in their current position. 142 controllers did not indicate their current position. The air traffic controller participant sample included journeyman controllers, developmental controllers, staff and supervisors, as well as holding several "other" positions. These "other" positions included jobs described as Traffic Management Coordinator.

Overall, the participants indicated they had spent an average of 4.15 years in their previous position. These positions included time as journeyman controller, developmental controller, staff, supervisor or other position. Those responding "Other" included cooperative education students, Academy instructors, and former Air Force air traffic controllers.

One goal of the biographical information form was to get a clear picture of the range and length of experience of the participants in the study. To this end they were asked the number of years and months as FPL, staff, or supervisor in their current facility and in any facility. The results are summarized in Table 5.4.8. Few of the respondents had been in staff or supervisory capacity for more than a few months. Half of the respondents had never acted in a staff position and almost two-thirds had never held a supervisory position. The amount of staff experience ranged from 0 to 10 years, with 97.6% of the participants having less than four years of experience. The findings are similar for supervisory positions; 99% of the respondents had seven or fewer years of experience. This indicates that our controller sample was indeed largely composed of individuals with current or previous controller experience.

Also of interest was the amount of time the incumbents (both controllers and supervisors) spent actually controlling air traffic. Respondents were asked how they had spent their work time over the past six months and then to indicate the percentage of their work time they spent controlling traffic (i.e., "plugged-in time") and the percentage they spent in other job-related activities (e.g., crew briefings, CIC duties, staff work, supervisory duties). The respondents indicated that they spent an average of 72.41% of their time controlling traffic and 23.33% of their time on other activities.

PSEUDO-APPLICANT SAMPLE

A total of 518 individuals served as pseudo-applicants in the validation study; 258 individuals from Keesler Air Force Base and 256 civilians took part in the study. The racial and gender breakdown of these samples is shown in Table 5.4.9.

COMPUTER USE AND EXPERIENCE QUESTIONNAIRE

To determine if individual familiarity with computers could influence their scores on several of the tests in the predictor battery, a measure of computer familiarity and skill was included as part of the background items. The Computer Use and Experience (CUE) Scale, developed by Potosky and Bobko (1997), consists of 12 5-point Likert-type items (1= Strongly Disagree, 2 = Disagree, 3 = Neither Agree nor Disagree, 4 = Agree, 5 = Strongly Agree), which asked participants to rate their knowledge of various uses for computers and the extent to which they used computers for various reasons. In addition, 5 more items were written to ask participants about actual use of the computer for such purposes as playing games, word processing and using e-mail. The resulting 17-item instrument is referred to in this report as the CUE-Plus.

Item Statistics

The means and standard deviations for each item are presented in Table 5.4.10. The information reported in the table includes both the Air Traffic Controller participants and the pseudo-applicants. Overall, the respondents show familiarity with computers and use them to different degrees. Given the age range of our sample, this is to be expected. As might be expected, they are fairly familiar with the day-to-day uses of computers, such as doing word processing or sending email. Table 5.4.11 shows the item means and standard deviations for each sample, breaking out the civilian and military pseudo-applicant samples and the controller participants. The means for the samples appear to be fairly similar. Table 5.4.12 shows the inter-item correlations of the CUE-Plus items. All the items were significantly correlated with each other.

Reliability of Cue-Plus

Using data from 1,541 respondents, the original 12-item CUE Scale yielded a reliability coefficient (alpha) of .92. The scale mean was 36.58 (SD = 11.34). The CUE-Plus, with 17 items and 1,533 respondents, had a reliability coefficient (alpha) of .94. The scale mean was 51.47 (SD = 16.11). Given the high intercorrelation between the items, this is not surprising. The item-total statistics are shown in Table 5.4.13. There is a high degree of redundancy among the items. The reliability coefficient for the samples are as follows: controllers, .93,

civilian pseudo-applicants, .91, and military pseudo applicants, .93, indicating that there were no large differences between sub-groups in responding to the CUE-Plus items.

Factor Analysis

Principal components analysis indicated that CUE-Plus had two factors, but examination of the second factor showed that it made no logical sense. Varimax and oblique rotations yielded the same overall results. The item "I often use a mainframe computer system" did not load strongly on either factor, probably because few individuals use mainframe computers. The varimax rotation showed an inter-factor correlation of .75. Table 5.4.14 shows the eigenvalues and percentages of variance accounted for by the factors. The eigenvalues and variance accounted for by the two-factor solution are shown in Table 5.4.15. The first factor accounts for over half of the variance in the responses, with the second factor accounting for only 6%. The last column in Table 5.4.16 shows the component matrix when only one factor was specified. Taken together, the data suggests that one factor would be the simplest explanation for the data structure.

PERFORMANCE DIFFERENCES

Gender Differences

The overall mean for the CUE-Plus was 51.31 (SD = 16.09). To see whether males performed significantly different than females on the CUE-Plus, difference scores were computed for the different samples. The difference score (d) is the standardized mean difference between males and females. A positive value indicates superior performance by males. The results are reported in Table 5.4.16. For all samples, males scored higher on the CUE (i.e., were more familiar with or used computers for a wider range of activities), but at most, these differences were only moderate (.04 to .42).

Ethnic Differences

Performance differences on the CUE-Plus between ethnic groups were also investigated. The means, standard deviations and difference scores (d) for each group is presented in Table 5.4.17. The table is split out by sample type (e.g., Controller, Military PA, Civilian PA). Comparisons were conducted between Caucasians and three comparison groups: African-Americans, Hispanics, and all non-Caucasian participants. A positive value indicates superior performance by Caucasians; a negative value indicates superior performance by the comparison group. The differences were very low to moderate, with the absolute value of the range from .04 to .31. The highest d scores were in the Military PA sample. Caucasians scored higher than the comparison groups in all cases except for the Civilian PA, in which African-Americans scored higher than Caucasians.

Summary

All in all, these results show the CUE-Plus to have very small differences for both gender and race. To the extent that the instrument predicts scores on the test battery, test differences are not likely to be attributable to computer familiarity.

RELATIONSHIP BETWEEN CUE-PLUS AND PREDICTOR SCORES

Correlations

An argument could be made that one's familiarity with and use of computers could influence scores on the computerized predictor battery. To address that question, correlations between the individual CUE-Plus items and the CUE-Plus total score with the AT-SAT predictor scores were computed. One area of interest is to what extent computer familiarity will affect the scores of applicants. To better examine the data in this light, the sample was separated into controllers and pseudo-applicants and separate correlations performed for the two groups. The correlations for the controller sample are shown in Tables 5.4.18 and 5.4.19. Table 5.4.18 shows the correlations between the CUE items and Applied Math, Angles, Air Traffic Scenarios, Analogy, Dials, and Scan scores. Table 5.4.19 shows the correlations between CUE-Plus and Letter Factory, Memory, Memory Recall, Planes, Sounds and Time-Wall (TW) scores. Tables 5.4.20 and 5.4.21 contain the same information for the pseudo-applicant sample. In general, the CUE-Plus scores were more highly correlated with performance on the AT-SAT battery for the pseudo-applicants than for the controllers.

The CUE-Plus total score was correlated ($p < .05$ or $p < .01$) with all predictor scores with the exception of those for Analogy: Latency and Time-Wall: Perceptual Speed for the pseudo-applicants. The same was true for the controller sample with regard to Air Traffic Scenarios: Accuracy, Memory: Number Correct, Recall: Number Correct, Planes: Projection and Planes: Time Sharing. Given the widespread use of computers at work and school and the use of Internet services this rate of correlation is not surprising.

The Letter Factory test scores on Situational Awareness and Planning and Thinking Ahead are highly correlated with the individual CUE-Plus items for the pseudo-applicants, while the controllers' Planning and Thinking Ahead scores were more often correlated with the CUE-Plus items than were their Awareness scores. One explanation for these high correlations is that the more comfortable one is with various aspects of using a computer, the more cognitive resources can allocated for planning. When the use of the computer is automatic, more concentration can be focused on the specific task.

The Time-Wall perception scores (Time Estimate Accuracy and Perceptual Accuracy) are highly correlated with the individual CUE items for the pseudo-applicants and correlated to a lesser extent for the controllers. The reverse is true for the Perceptual Speed variable: the controller scores are almost all highly correlated with CUE-Plus items, while only two of the items are correlated for the pseudo-applicants. The Time-Wall test will not be included in the final test battery, so this is not a consideration as far as fairness is concerned.

Using a mainframe computer correlated with only one of the test battery scores for the controller sample, but correlated highly with several test scores for the pseudo-applicants. The fact that controllers use mainframes in their work probably had an effect on their correlations.

Regression Analyses

Regression analyses were conducted to investigate the extent to which the CUE-Plus and four demographic variables predict test performance. The dependent variables predicted were the measures that are used in the test battery. Dummy variables for race were calculated, one to compare Caucasians and African-Americans, one to compare Hispanics to Caucasians, and the third to compare all minorities to Caucasians. Those identified as Caucasian were coded as 1, members of the comparison groups were coded as 0. 1,497 cases were analyzed. Thus, five variables were used in the regression analyses: three "race" variables, education, age, gender and score on CUE-Plus.

Applied Math

The variables described above were entered as predictors for the total number of items correct. For all three comparisons, all variables were included in the final model. That model accounted for approximately 20% of the variance for all three comparisons. Gender was the best predictor of performance. Negative b weights for gender indicate that males performed better than females. The positive weights for age indicate that the older the individual, the higher their score on the Applied Math test. Education and CUE-Plus score were also positively weighted, indicating that the more education one received and the more familiar one is with computers, the better one is likely to do on the Applied Math test. Caucasian participants scored higher than did their comparison groups. The statistics for each variable entered are shown in Table 5.4.22.

Angles Test

The same general pattern of results holds true for the Angles test. Table 5.4.23 shows the statistics for each variable. Age was not a predictor of performance for this test in any of the comparisons. The other variables were predictive for the Caucasian/African-American and the Caucasian/Minority models. Race was not a predictor for the Caucasian/Hispanic model. In all cases, females performed less well than males. Amount of education and CUE-Plus were positive indicators of performance. The predictor sets accounted for about 10% of the variance in Angles test scores;, the CUE-Plus score contributed little to explaining the variance in scores.

Air Traffic Scenarios

The predictor variables accounted for between 15% and 20% of the variance in the Efficiency scores (see Table 5.4.24), but only about 3% for Safety (Table 5.4.25) and 7% for Procedural Accuracy (Table 5.4.26). CUE-Plus scores were predictive of performance for all three variables, but not particularly strongly. Age was a positive predictor of performance for only the Procedural Accuracy variable. Gender was a predictor for Efficiency in all three models, but not consistently for the other two variables. Education predicted only Procedural Accuracy. Race was not a predictor for the Caucasian/Hispanic models, although it was for the other models.

Analogy Test

Age was a fairly consistent predictor for the Information Processing (see Table 5.4.27) and Reasoning variables (see Table 5.4.28), although it did not predict Reasoning performance in the Caucasian/Minority and Caucasian/African-American equations. Education was a negative predictor for Information Processing, but was positively related to Reasoning. CUE-Plus was a predic-

tor for Reasoning, but not for Information Processing. Together, the independent variables accounted for about 11% of the variance in the Information Processing scores and about 16% of the Reasoning scores.

Dials Test

The number of items correct on the Dials test was predicted by gender, education, race and CUE-Plus. Table 5.4.29 shows the statistics associated with the analysis. Males are predicted to score higher than females; those with higher education are predicted to perform better on the test than those with less education. Race was positively related with Dials scores, indicating that Caucasians tended to score higher than their comparison groups. CUE-Plus was a significant, but weak predictor for the Caucasian/Minority and Caucasian/African-American models. It did not predict performance in the Caucasian/Hispanic model. The four variables accounted for between 8% and 10% of the variance in Dials test performance.

Letter Factory Test

The Letter Factory test had two scores of interest: Situational Awareness and Planning and Thinking Ahead. Age and gender did not predict for either score. Race and CUE-Plus score were predictors for both variables; education was a predictor for Situational Awareness. These variables accounted for between 7% and 12% of the variance in the Situational Awareness score (see Table 5.4.30) and 11% to 15% of the variance in the Planning and Thinking Ahead score (see Table 5.4.31).

Scan Test

The variables in the regression equation accounted for only 1% to 3% of the variance in the Scan score (see Table 5.4.32). Education was a positive predictor for all three equations. Race was a predictor for the Caucasian/African-American model. CUE-Plus score positively predicted performance in the Caucasian/Hispanic equation.

Summary

The question of interest in this section has been the extent to which computer familiarity, as measured by CUE-Plus, influences performance on the AT-SAT test battery. The correlation matrices indicated a low to moderate level of relationship between CUE-Plus and many of the variables in the pilot test battery for the controller sample. The correlations were higher for the pseudo-applicant sample. To further investigate these relationships, regression analyses were conducted to see how well Cue-Plus and other relevant demographic variables predicted performance on the variables that were used in the V 1.0 test battery.

The results showed that overall, the demographic variables were not strong predictors of test performance. The variables accounted for relatively little of the variance in the test scores. CUE-Plus was identified as a predictor for nine of the eleven test scores. However, even for the scores where CUE-Plus was the strongest predictor of the variables entered, it accounted for no more than 8% of the variance in the score. In most of the scores, the effect, although statistically significant, was realistically negligible.

SUMMARY

This chapter described the participants in the AT-SAT validation study. The participants represented both genders and the U.S. ethnicities likely to form the pool of applicants for the Air Traffic Controller position.

In addition to describing the demographic characteristics of the sample on which the test battery was validated, this chapter also described a measure of computer familiarity, CUE. CUE was developed by Potosky and Bobko (1997) and revised for this effort (CUE-Plus). The CUE-Plus is a highly reliable scale (alpha = .92); factor analysis indicated that there was only one interpretable factor. Analysis of the effect of gender on CUE-Plus scored showed moderate differences for the controller sample, none for the pseudo-applicant sample; males scored higher on the CUE-Plus than did females. There were also small to moderate differences in CUE-Plus for ethnicity. The strongest differences were found in the military pseudo-applicant sample.

CUE-Plus items showed a moderate to high correlation with the variables assessed in the validation study. The CUE-Plus was also shown to be a fairly weak but consistent predictor of performance on the variables that were included in V 1.0 test battery. Although there were some performance differences attributable to gender, race and computer experience none of these were extremely strong. The effects of computer skill would be washed out by recruiting individuals who have strong computer skills.

CHAPTER 5.5

Predictor-Criterion Analyses

Gordon Waugh, HumRRO

Overview of the Predictor-Criterion Validity Analyses

The main purpose of the validity analyses was to determine the relationship of AT-SAT test scores to air traffic controller job performance. Additional goals of the project included selecting tests for the final AT-SAT battery, identifying a reasonable cut score, and the development of an approach to combine the various AT-SAT scores into a single final score. Several steps were performed during the validity analyses:

- Select the criteria for validation analyses
- Compute zero-order validities for each predictor score and test
- Compute incremental validities for each test
- Determine the best combination of tests to include in the final battery
- Determine how to weight the test scores and compute the predictor composite score
- Compute the validity coefficients for the predictor composite
- Correct the validity coefficient for statistical artifacts

Many criterion scores were computed during the project. It was impractical to use all of these scores during the validation analyses. Therefore, a few of these scores had to be selected to use for validation purposes. The three types of criterion measures used in the project were the CBPM (Computer-Based Performance Measure), the Behavior Summary Scales (which are also called *Ratings* in this chapter), and the HiFi (High Fidelity Performance Measure). The development, dimensionality, and construct validity of the criteria are discussed at length in Chapter 4 of this report.

The CBPM was a medium fidelity simulation. A computer displayed a simulated air space sector while the examinee answered questions based on the air traffic scenario shown. The Behavior Summary Scales were performance ratings completed by the examinee's peers and supervisors. The HiFi scores were based upon observers' comprehensive ratings of the examinee's two-day performance on a high-fidelity air traffic control simulator.

Based on the analyses of the dimensions underlying the criteria, it was concluded that the criteria space could be summarized with four scores: (a) the CBPM score, (b) a single composite score of the 10 Behavior Summary Scales (computed as the mean of the 10 scales), (c) HiFi 1: Core Technical score (a composite of several scores) and (d) HiFi 2: Controlling Traffic Safely and Efficiently (a composite of several scores). The small sample size for the HiFi measures precluded their use in the selection of a final predictor battery and computation of the predictor composite. They were used, however, in some of the final validity analyses as a comparison standard for the other criteria.

A single, composite criterion was computed using the CBPM score and the composite Ratings score. Thus, the following three criteria were used for the validity analyses: (a) the CBPM score, (b) the composite Ratings score, and (c) the composite criterion score.

Zero-Order Validities

It is important to know how closely each predictor score was related to job performance. Only the predictor scores related to the criteria are useful for predicting job performance. In addition, it is often wise to exclude tests from a test battery if their scores are only slightly related to the criteria. A shorter test battery is cheaper to develop, maintain, and administer and is more enjoyable for the examinees.

Therefore, the zero-order correlation was computed between each predictor score and each of the three criteria (CBPM, Ratings, and Composite). Because some tests produced more than one score, the multiple correlation of each criterion with the *set* of scores for each multi-measure test was also computed. This allowed the assessment of the relationship between each test, as a whole, and the criteria. These correlations are shown in Table 5.5.1 below.

Ideally, we would like to know the correlation between the predictors and the criteria among job applicants. In this study, however, we did not have criteria information for the applicants (we did not actually use real applicants but rather pseudo-applicants). That would require a predictive study design. The current study uses

a concurrent design: We computed the predictor-criteria correlations using current controllers. Correlations are affected by the amount of variation in the scores. Scales with little variation among the scores tend to have low correlations with other scales. In this study, the variation in the predictor scores was much greater among the pseudo-applicants than among the controllers. Therefore, we would expect the correlations to be higher within the pseudo-applicant sample. A statistical formula, called *correction for range restriction*, was used to estimate what these correlations would be among the pseudo-applicants. The formula requires three values: (a) the uncorrected correlation, (b) the predictor's standard deviation for the pseudo-applicant sample, and (c) the predictor's standard deviation for the controller sample.

Table 5.5.1 shows both the corrected and uncorrected correlations. The amount of correction varies among the predictors because the ratio of the pseudo-applicant vs. controller standard deviations also varies. The greatest correction occurs for predictors which exhibit the greatest differences in standard deviation between the two samples (e.g., Applied Math). The least correction (or even downward correction) occurs for predictors whose standard deviation differs little between the two samples (e.g., the EQ scales).

Table 5.5.1 shows that most of the tests exhibit moderate to high correlations with the CBPM and low to moderate correlations with the Ratings. Some scales, however had no significant ($p < .05$) correlations with the criteria: the Information Processing Latency scale from the Analogies test and 2 of the 14 scales from the Experiences Questionnaire (*Tolerance for High Intensity* and *Taking Charge*). In addition, these two EQ scales along with the EQ scale, *Working Cooperatively*, correlated negatively with the CBPM and composite criteria. Thus, it is doubtful that these scores would be very useful in predicting job performance. Analyses of their incremental validities, discussed below, confirmed that these scores do not significantly improve the prediction of the criteria.

The EQ (Experiences Questionnaire) is a self-report personality inventory. It is not surprising, then, that its scales do not perform as well as the other tests—which are all cognitive measures—in predicting the CBPM which is largely a cognitive measure. The cognitive tests were generally on a par with the EQ in predicting the Ratings criterion. A notable exception was the Applied Math test, which greatly outperformed all other tests in predicting either the CBPM or the Ratings. Note that the Ratings criterion is a unit-weighted composite of the 10 behavior summary scales completed by supervisors. The EQ correlated quite highly with a number of these behavior summary scales, e.g., the four scales making up the Technical Effort factor, and the single scale in the teamwork factor, but not very highly with the composite Ratings criterion.

Composure and *Concentration* are the only EQ scales that correlate above .08 with the CBPM, whereas eight scales correlate this highly with the Ratings. This is not surprising because both personality measures and performance ratings incorporate non-cognitive performance tors such as motivation. The moderate size of the multiple correlation of the EQ with the CBPM of .16 is misleadingly high because three of the EQ scales correlate negatively with the CBPM. The size of a multiple correlation is usually just as large when some of the correlations are negative as when all are positive. Scales that correlate negatively with the criterion, however, should not be used in a test battery. Otherwise, examinees scoring higher on these scales would get lower scores on the battery. When the three scales that correlate negatively with the CBPM are excluded, the EQ has a multiple correlation of only .10 (corrected for shrinkage) with the CBPM.

Incremental Validities

At this point, all the scores—except for the Information Processing score from the Analogies test and 7 of the 14 scores from the Experiences Questionnaire—have demonstrated that they are related to the criteria. The next step was to determine which scales have a unique contribution in predicting the criteria. That is, some scales might not add anything to the prediction because they are predicting the same aspects of the criteria as some other scales.

If two tests predict the same aspects of the criteria then they are redundant. Only one of the tests is needed. The amount of the unique contribution that a test makes toward predicting a criterion is called *incremental validity*. More precisely, the incremental validity of a test is the increase in the validity of the test battery (i.e., multiple correlation of the criterion with the predictors) when that test is added to a battery.

Table 5.5.2 shows the incremental validities for each test and scale. There are two values for most tests. The first value shows the incremental validity when the test is added to a battery that contains *all* the other tests; the other value shows the incremental validity when the test

is added to *only the tests in the final AT-SAT battery*. In addition, incremental validities for the final version of the EQ test (in which three of the original EQ scales were dropped) are shown.

Three tests have a substantial unique contribution to the prediction of the criteria. Each has an incremental validity greater that .10 (corrected for shrinkage but not for range restriction). They are, in order of decreasing incremental validity, Applied Math, EQ, and Air Traffic Scenarios.

Determination of Scale Weights for the Test Battery

The full AT-SAT battery would require more than a day of testing time. Thus, it was desired to drop some of the tests for this reason alone. Therefore, several tests were excluded from the final test battery taking into consideration the following goals:

1. Maintain high concurrent validity.
2. Limit the test administration time to a reasonable amount.
3. Reduce differences between gender/racial group means.
4. No significant differences in prediction equations (i.e., regression slopes or intercepts) favoring males or whites (i.e., no unfairness).
5. Retain enough tests to allow the possibility of increasing the predictive validity as data becomes available in the future.

There are typically three main types of weighting schemes: regression weighting, unit weighting, and validity weighting. In regression weighting, the scales are weighted to maximize the validity of the predictor composite in the sample of examinees. The main problem with this scheme is that the validity drops when the predictor weights are used in the population. Unit weighting gives equal weight to each scale or test. It tends to sacrifice some sample validity, but its validity does not typically drop in the population because the weights are chosen independent of the sample. Validity weighting assigns each scale's simple validity as its weight. This scheme is a compromise between the two methods. Validity weights do almost as well as regression weights in the sample. More importantly, validity weights are less sensitive to differences in samples than regression weights.

The large numbers of scales and parameters to consider for each scale made it difficult to subjectively decide which tests to drop. For each scale, ten parameters were relevant to this decision. To aid in this decision, a computer program was written (using Visual Basic) which essentially considered all these parameters simultaneously. In choosing the set of optimal scale weights, the program considered the following sets of parameters of the resulting predictor composite: overall validity, differences in group means, differences in the groups' regression slopes, and differences in the groups' intercepts. There were three parameters for each type of group difference: females vs. males, blacks vs. whites, Hispanics vs. whites. One final feature of the program is that it would not allow negative weights. That is, if a scale's computed weight was such that a high score on the scale would lower the score on the overall score then the scale's weight was set to zero.

Several computer runs were made. For each run, the relative importance of the parameters were varied. The goal was to maximize the overall validity while minimizing group differences. In the end, the group difference with the greatest effect on the overall validity was the black vs. white group mean on the composite predictor. Thus, the ultimate goal became to reduce the differences between the black and white means without reducing the maximum overall validity by a statistically significant amount.

There were only nine scales remaining with non-zero weights after this process. This low number of scales was undesirable. It is possible that some of the excluded tests might perform better in a future predictive validity study than in the concurrent study. If these tests are excluded from the battery, then there will be no data on them for the predictive validity study. Another limitation of this technique is that the weights will change, possibly substantially, if applied to another sample.

Therefore, a combination of the validity weighting and optimal weighting schemes was used. For each scale, the weight used was the mean of the optimal and validity weights. A description of the computation of the validity and optimal weights follows.

The computation of the validity weights for a single-scale test was straightforward. It was merely the correlation, corrected for range restriction, of the scale with the composite criterion. The computation for the multi-scale tests was somewhat more complex. First, the multiple correlation, corrected for range restriction, of the test with the composite criterion was computed. This represents contribution of the test to the composite predictor. Then, the correlations of each of the test's scales with the composite criterion, corrected for range restriction, were computed. The validity weights of the scales were computed according to the following formula:

$$w_i = R\frac{r_i}{\sum_{j=1}^{k} r_j} \qquad \text{[Equation 5.5.1]}$$

where w_i = validity weight of scale i, r_i = correlation of the predictor scale with the criterion, R = multiple correlation of the test with the criterion, r_j = the correlation with the criterion of the scale j of the k scales within the test. All correlations were corrected for range restriction.

The validity weights and optimal weights had to be put on a common metric before they could be combined. Each validity weight was multiplied by a constant such that all the weights summed to 1.00. Similarly, each optimal weight was multiplied by a constant such that all the weights summed to 1.00. Each predictor's combined weight was then computed as the mean of its rescaled optimal and validity weights. Finally, the combined weight was rescaled in the same manner as the validity and optimal weights. That is, each combined weight was multiplied by a constant such that all the weights summed to 1.00. This rescaling was done to aid interpretation of the weights. Each weight represents a predictor's relative contribution, expressed as a proportion, to the predictor composite.

Predictor Composite

The predictor composite was computed using the combined predictor weights described above. Before applying the weights, the predictor scores had to be transformed to a common metric. Thus, each predictor was standardized according to the pseudo-applicant sample. That is, a predictor's transformed score was computed as a z-score according to the following formula:

$$z = \frac{x - \hat{\mu}_p}{\hat{\sigma}_p} \qquad \text{[Equation 5.5.2]}$$

where z = the predictor's z-score, x = the raw predictor score, $\hat{\mu}_p$ = the predictor's mean score in the pseudo-applicant sample, and $\hat{\sigma}_p$ = the predictor's standard deviation in the pseudo-applicant sample (i.e., the estimate of the predictor's standard deviation in the population based on the pseudo-applicant sample data).

The predictor composite was then computed by applying the rescaled combined weights to the predictor z-scores. That is, the predictor composite was computed according to the following formula:

$$\textit{raw composite predictor} = \sum_{i=1}^{k} w_i z_i \qquad \text{[Equation 5.5.3]}$$

where k = the number of predictors, w_i = the rescaled combined weight of the ith predictor, and z_i = the z-score of the ith predictor. In other words, the raw composite predictor score is the weighted sum of the z-scores. This score was rescaled such that a score of 70 represented the cut score and 100 represented the maximum possible score. This is the scaled AT-SAT battery score. The determination of the cut score is described later in this chapter. To simplify the programming of the software that would administer and score the AT-SAT battery, a set of weights was computed that could be applied to the raw predictor scores to obtain the scaled AT-SAT battery score. Thus the scaled AT-SAT battery score was computed according to the following formula:

$$\textit{Scaled AT-SAT Battery Score} = \sum_{i=1}^{k} w_i x_i$$

Equation 5.5.4]

where k = the number of predictors, w_i = the raw-score weight of the ith predictor, and x_i = the raw score of the ith predictor.

The effects of using various weighting schemes are shown in Table 5.5.3. The table shows the validities both before and after correcting for shrinkage and range restriction. Because the regression procedure fits an equation to a specific sample of participants, a drop in the validity is likely when the composite predictor is used in the population. The amount of the drop increases as sample size decreases or the number of predictors increases. The correction for shrinkage attempts to estimate the amount of this drop. The formula used to estimate the validity corrected for shrinkage is referred to by Carter (1979) as *Wherry (B)* (Wherry, 1940). The formula is :

$$\hat{R} = \sqrt{1 - \left(1 - R^2\right)\frac{n-1}{n-k-1}} \qquad \text{[Equation 5.5.5]}$$

where $\hat{R}$ = the validity corrected for shrinkage, R is the uncorrected validity, n = the sample size, and k = the number of predictors. Where validities were corrected for both range restriction and shrinkage, the shrinkage correction was performed first.

As noted above, the final AT-SAT score was computed using the Combined method of weighting the predictors. Only the regression method had a higher validity. In fact, the Combined method probably has a higher validity if we consider that its correction for shrinkage overcorrects to some extent. Finally, the regression-weighted validity is based on all 35 scales whereas the Combined validity is based on just 26 tests. Thus, the Combined weighting method produces the best validity results.

The Combined method produced the second-best results in terms of mean group differences and fairness. Only the Optimal low *d*-score weighting method had better results in these areas, and its validity was much lower than the Combined method's validity. None of the weighting methods produced a statistically significant difference in standardized regression slopes among the groups. Thus, the Combined weighting method was the best overall. It had the highest validity and the second-best results in terms of group differences and fairness. Therefore, the Combined weighting method was used to compute the final AT-SAT battery score.

Final AT-SAT Battery Validity

The best estimate of the validity of the AT-SAT battery is .76. This value is extremely high. Table 5.5.4 shows the validity of the AT-SAT battery for various criteria. The table also shows how various statistical corrections affect the validity estimate. The most relevant validity of .76 is the correlation with the composite criterion which is corrected for range restriction, shrinkage, and criterion unreliability.

The low sample size for the high fidelity criteria precludes accurate estimates of validity. The purpose of the high-fidelity criteria was to obtain independent evidence that the CBPM and Ratings were related to job performance. As shown in a previous chapter, the high correlations of the CBPM and Ratings with the high fidelity criteria are strong evidence that the CBPM and Ratings are accurate indicators of job performance.

Interrater agreement reliability was used to correct the validities for the Ratings and HiFi criteria. Reliability for the CBPM was estimated by computing its internal consistency (coefficient alpha = .59), but this figure is probably an underestimate because the CBPM appears to be multidimensional (according to factor analyses). Ideally, the reliability for the CBPM should be computed as a test-retest correlation. This could not be computed, however, because each examinee took the CBPM only once. Previous research has found that similar measures (i.e., situational judgement tests) have test-retest reliabilities of about .80, with most in the range between .7-.9. Thus, three different reliabilities were used to correct the CBPM's validity for unreliability: .8 (best guess), .9 (upper bound estimate), and .7 (lower bound estimate), respectively. The reliability of the composite measure could not be directly measured. Therefore, an approximation of the composite criterion reliability was computed as the mean of the ratings and CBPM reliabilities.

Determining the Cut Score

One of the specifications for the AT-SAT battery was that a score of 70 would represent the cut score and a score of 100 would represent the highest possible score. The cut score and maximum score were first determined on the AT-SAT battery's original scale. Then these two scores were transformed to scores of 70 and 100 on the scaled AT-SAT battery scale.

The determination of the highest possible score was relatively straightforward. There was, however, one complication. The maximum possible scores for the simulation scales (i.e., Letter Factory scales, Air Traffic Scenarios scales) and some of the other scales (e.g., Analogies information processing scores) were unknown. Thus, the determination of the highest possible score was not simply a matter of adding up the maximum scores possible for each scale. For the scales with an unknown maximum possible score, the maximum scores attained during the study were used to estimate the highest scores likely to be attained on these scales in the future.

The determination of the cut score was more involved. The main goal in setting the cut score was to at least maintain the current level of job performance in the controller workforce. After examining the effects of various possible cut scores on controller performance, a cut score was selected that would slightly improve the job performance of the overall controller workforce. Specifically, the cut score was set such that the mean predicted criterion score, among pseudo-applicants passing the battery, was at the 56th percentile of the current controller distribution of criterion scores.

Table 5.5.5 shows the effects of this cut score on selection rates and predicted job performance. If all the pseudo-applicants were hired, their mean job performance would be at only the 33rd percentile of the current controller distribution. Thus, using the AT-SAT Battery, with the chosen cut score, is considerably better

than using no screening. That is, if all of the pseudo-applicants were hired (or some were randomly selected to be hired), their performance level would be much lower than the current Controllers.

Impact of AT-SAT on Workforce Capabilities

Figure 5.5.1 shows the relationship between scores on the AT-SAT battery and the expected or average performance of examinees at each score level. For comparison purposes, the previous OPM battery, which had a (generously corrected) validity of about .30 has been placed on the same scale as the AT-SAT composite. The primary point is that applicants who score very high (at 90) on the AT-SAT are expected to perform near the top of the distribution of current controllers (at the 86th percentile). Applicants who score very high (at 90) on the OPM test, however, are expected to perform only at the middle of the distribution of current controllers (at the 50th percentile). Only 1 out of 147 applicants would be expected to get an OPM score this high (90 or above). Someone with an OPM score of 100 would be expected to perform at the 58th percentile. Consequently, there is no way that the OPM test, by itself, could be used to select applicants much above the mean of current controllers. In the past, of course, the OPM test was combined with a nine-week screening program resulting in current controller performance levels. The AT-SAT is expected to achieve about this same level of selectivity through the pre-hire screening alone.

Table 5.5.6 shows the percent of high performers expected for different cutpoints on the AT-SAT and OPM batteries. This same information is shown graphically in Figure 5.5.2. Here, high performance is defined as the upper third of the distribution of performance in the current workforce as measured by our composite criterion measure. If all applicants scoring 70 or above on the AT-SAT are selected, slightly over one-third would be expected to be high performers. With slightly greater selectivity, taking only applicants scoring 75.1 or above, the proportion of high performers could be increased to nearly half. With a cutscore of 70, it should be necessary to test about 5 applicants to find each hire. At a cutscore of 75.1, the number of applicants tested per hire goes up to about 10. By comparison, 1,376 applicants would have to be tested for each hire to obtain exactly one-third high performers using the OPM screen.

CHAPTER 5.6

Analyses Oof Group Differences and Fairness

Gordon Waugh, HumRRO

SUMMARY

The group means on the composite predictor for females, blacks, and Hispanics were significantly lower than the means for the relevant reference groups (males, whites). The difference was greatest for blacks. The cognitive tests displayed much greater differences than did the EQ scales. However, the EQ scales had much lower validity as well. Although the predictor composite exhibited lower group means for minorities, no evidence of unfairness was found. In fact, the composite predictor over-predicted the performance of all three minority groups (females, blacks, and Hispanics) at the cut score. The validity coefficients and regression slopes were remarkably similar among the groups. Among the individual test scales, there were no cases (out of a possible 111) in which the slopes of the regression lines differed significantly between a minority and reference group. These results show that the test battery is fair for all groups.

INTRODUCTION

A personnel selection test may result in differences between white and minority groups. In order to continue to use a test that has this result, it is required to demonstrate that the test is job- related or valid. Two types of statistical analyses are commonly used to assess this issue. The analysis of mean group differences determines the degree to which test scores differ for a minority group as a whole (e.g., females, blacks, Hispanics) when compared with its reference group (i.e., usually whites or males). Fairness analysis determines the extent to which the relationship between test scores and job performance differs for a minority group compared to its reference group.

Our sample contained enough blacks and Hispanics to analyze these groups separately but too few members of other minority groups to include in the analyses. It was decided not to run additional analyses with either all minorities combined or with blacks and Hispanics combined because the results differed considerably for blacks vs. Hispanics. Thus, the following pairs of comparison groups were used in the fairness analyses: male vs. female, white vs. black, and white vs. Hispanic. The descriptive statistics for the predictors and criteria are shown in Tables 5.6.1–5.6.3.

Cut Scores

Both the analyses of sub-group differences and fairness required a cut score (i.e., a specified passing score) for each test and for the predictor composite score. Therefore, hypothetical cut scores had to be determined. The cut score on the predictor composite was set at the 32nd percentile on the controller distribution. (This score was at the 78th percentile on the pseudo-applicant distribution.) Thus, the hypothetical cut score for each test was also set at the 32nd percentile on the controller distribution for the purposes of the fairness and group mean difference analyses. The determination of the cut score is discussed elsewhere in this report. Regression analyses predicted that the mean level of job performance for applicants passing the AT-SAT battery would be at the 56th percentile of the job performance of current controllers. That is, it is predicted that applicants passing the battery will perform slightly better than current controllers.

Estimation of Missing Values

There were few blacks in the controller (n = 98) and pseudo-applicant samples (n = 62). In addition, there were even fewer in the analyses because of missing values on some tests. When the composite predictor was computed, missing values on the individual scales were estimated. Otherwise, a participant would have received a missing value on the composite if any of his/her test scores were missing. Each missing score was estimated using a regression equation. The regression used the variable with the missing score as the dependent variable and the scale that best predicted the missing score as the independent variable. The predictor scale had to be from a different test than the missing score. For example, if an examinee's Applied Math score was missing then his/her Angles score was used to estimate it. If both the Applied Math and Angles scores were missing, then the estimated

composite predictor score would also be missing. Each missing EQ score, however, was predicted using another EQ scale. Missing scores were estimated only when building the composite predictor. That is, missing values were not estimated for analyses that used the individual test scores. This was judged to be a conservative estimation procedure because (a) only one independent variable was used in each estimation regression (b) none of the blacks and few of the other examinees were missing more than one test score, and (c) each test score contributed only a small amount to the final composite predictor score. The amount of error caused by the estimation of missing values is very likely to be trivial. To ensure that the covariances were not artificially increased by the estimation of missing values, random error was added to each estimated value.

GROUP DIFFERENCES

Analyses

Only the pseudo-applicant sample was used for the group difference analyses. This sample best represented the population of *applicants*. Therefore, air traffic controllers were excluded from these analyses.

The *Uniform Guidelines on Employee Selection Procedures* (Federal Register, 1978, Section 4.D.) state that evidence of adverse impact exists when the passing rate for any group is less than four-fifths of the passing rate for the highest group:

> A selection rate for any race, sex, or ethnic group which is less than four-fifths ($^4/_5$) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.

Therefore, the passing rates for each test were computed for all five groups (males, females, whites, blacks, Hispanics). Then the passing rates among the groups were compared to see if the ratio of the passing rates fell below four-fifths. Separate comparisons were done within the gender groups and within the racial groups. That is, males and females were compared; and blacks and Hispanics were compared to whites.

The Uniform Guidelines (Section D.4.) state that adverse impact might exist even if the passing rate for the minority group is greater than four-fifths the reference group's passing rate:

> Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms . . .

Therefore, the differences in the passing rates were tested for statistical significance using 2 ´ 2 chi-square tests of association. For each predictor score, one chi-square analysis was done for each of the following pairs of groups: male-female, white-black, and white-Hispanic. An example is shown in Table 5.6.4 below. This shows how the chi-square test was computed which compared male and female passing rates.

The groups were also compared by computing the mean test score for each group. The differences in the means between the minority groups and reference groups (i.e., males or whites) were then tested for statistical significance using independent-groups *t*-tests. The differences between the means were then converted to *d*-scores which express these differences in terms of standard deviation units based on the reference group's standard deviation. For example, a *d*-score of –.48 for females indicates that the mean female score is –.48 standard deviations below the mean of the male distribution of scores (i.e., at the 32nd percentile of the male distribution according to a table of the normal distribution).

Results and Conclusions

Table 5.6.5 shows the results for the passing rate analyses. Several tests—including the predictor composite—exhibited evidence of group differences for females, blacks, and Hispanics according to the four-fifths rule. In most of these cases, the difference in passing rates was statistically significant. Females and Hispanics had similar passing rates; blacks had by far the lowest passing rates.

Table 5.6.5 also shows the differences between the group means expressed as *d*-scores. The significant *d*-scores are asterisked in the table. These results were very similar to those for the passing rates. The group ´ predictor combinations that had significantly lower passing scores (compared to the reference group) also tended to have significantly lower *d*-scores. All three minority groups tended to score below their reference groups, but the differences were often not statistically significant. Blacks scored lowest on most tests. On the composite predictor, Hispanics had the highest *d*-score, followed by females and blacks, respectively. The Hispanic *d*-score was not statistically significant.

The group differences for the EQ scales were much lower than for the cognitive tests. (The Memory Test and the Memory Retest, however, had very small group

differences. In fact, females did better than males on these two tests.) For example, for blacks, the median *d*-score was –.48 among the 23 cognitive scores but only –.20 among the 14 EQ scales. However, the EQ scales also had much lower validity than did the other tests. This is probably why the passing rates are much higher for the EQ. In fact, the passing rates on half of the EQ scales were higher for the pseudo-applicants than for the controllers (i.e., half of the passing rates were higher than 68%, which is the passing rate for each test in the controller sample). In all the other tests, the passing rate was much lower for the pseudo-applicants than for the controllers.

There are two possible reasons for the high passing rates for the EQ scales: (a) the pseudo-applicants and current controllers possess nearly the same levels of the personality traits supposedly measured by the EQ or (b) the EQ scales are measuring some unwanted constructs (probably in addition to the traits that the scales were designed to measure). If the first possibility is true, then one must conclude that either these traits are not really needed on the job or that the current controllers would perform even better on the job if they improved in these traits. If the second possibility is true, then some unwanted constructs, such as social desirability, are being measured to some degree by the EQ scales.

In conclusion, the predictor composite for the final AT-SAT battery exhibited lower scores for all three minority groups (i.e., females, blacks, and Hispanics) compared to their reference groups (i.e., males and whites) in terms of both passing rates and *d*-scores. All of these differences, except for the Hispanic *d*-score, were statistically significant. The relative passing rates on the predictor composite for females, blacks, and Hispanics (compared to the passing rates for the reference groups: males and whites) were .54, .11, and .46, respectively. Thus, there was evidence of sub-group differences in test performance for the three minority groups.

It should be noted that subgroup differences in predictor scores do not necessarily imply bias or unfairness. If low test scores are associated with low criterion performance and high test scores are related to high criterion performance, the test is valid and fair. The fairness issue is discussed below.

FAIRNESS

Analyses

The fairness analyses requires analyses of job performance as well as test scores. As a consequence, all fairness analyses were performed on the concurrent validation controller sample. A test is considered fair when the relationship between the predictor test and job performance is the same for all groups. In our analyses, only differences that aid whites or males were considered to be unfair. Fairness is assessed by performing regression analyses using the test score as the independent variable and the criterion measure as the dependent variable. To assess the fairness of a predictor for females, for example, two regressions are performed: one for males and one for females. In theory, the predictor is considered to be fair if the male and female regression lines are identical. In practice, the test is considered to be fair if the difference between the equations of the two regression lines is not statistically significant (given a reasonable amount of power).

The equations of the two regression lines (e.g., male vs. female regression lines) can differ in their slopes or their intercepts. If the slopes differ significantly then the predictor is not fair. If the slopes do not differ significantly, then the intercepts are examined. In this study, to maximize interpretability, the predictor scores were scaled such that all the intercepts occurred at the cut point (i.e., passing score). Specifically, the cut score was subtracted from the predictor score.

Although fairness analysis is based on a separate regression line for each of the two groups being compared, a quicker method uses a single regression analysis. The significance tests in this analysis are equivalent to the tests that would be done using two lines. In this analysis, there is one dependent variable and three independent variables. The dependent variable is the criterion. The independent variables are shown below:

- The predictor.
- The group (a nominal dichotomous variable which indicates whether the person is in the focal or reference group). If this independent variable is significant, it indicates that, if a separate regression were done for each of the two groups, the intercepts of the regression lines would be significantly different. Because the predictors in this study were rescaled for these analyses such that the intercepts occurred at the cut scores, a difference in intercepts means that the two regression lines are at different elevations at the cut score. That is, they have different criterion scores at the predictor's cut score.
- The predictor by group interaction term. This is the product of group (i.e., 0 or 1) and the predictor score. If this independent variable is significant, it indicates that, if a separate regression were done for each of the two groups, the slopes of the regression lines would be significantly different. The standardized slopes equal the validities.

The regression equation is shown below:

$$criterion = b_0 + b_{predictor}\, predictor + b_{group}\, group + b_{interaction}\, interaction + error \qquad \text{[Equation 5.6.1]}$$

The composite criterion and the composite predictor were used for the fairness analyses. The composite criterion was the weighted sum of the composite rating and the CBPM. Based on their relationships with the high fidelity criterion measures, the ratings and CBPM were assigned weights of .4 and .6 respectively. The ratings and CBPM scores were standardized before they were added.

RESULTS AND CONCLUSIONS

Examples of the fairness regression scatterplots are shown in Figures 5.6.1, 5.6.2, and 5.6.3 below. The regression lines for both groups (i.e., reference and minority) are shown in each plot. The slopes of the two regression lines are very similar in each of the three graphs. Thus, the validities differ little between the groups in each graph. The near-parallelism of the regression lines is reflected in the similar values of the two groups' standardized slopes listed in the graphs and in Table 5.6.6. In terms of the intercepts, however, the white and male regression lines are above the female, Hispanic, and especially the black regression lines at the cut score. Thus, the predictor composite *over*-predicts performance for the three minority groups compared with the reference groups, which means that the test actually favors the minority groups. Under these circumstances, a regression equation based on the total sample produces predicted job performance levels that are higher than the actual performance levels observed for minorities. In a selection situation, minorities would be favored in that they would achieve a higher ranking on a selection list than would be indicated by actual performance.

Table 5.6.6 shows the results of the fairness regressions for all of the predictor scores. It displays the standardized slopes for each regression line. These are equivalent to validity coefficients. The table also shows the *Regression Lines' Difference at Cut Score (in Std. Dev. Units)*. This is the difference between the intercepts divided by the reference group's standard error of estimate. Thus it can be considered to be the difference between minority vs. reference groups' predicted criterion scores at the cut score scaled in standard deviation units about the regression line[17]. A negative value indicates that the minority's regression line was below the reference group's line.

The table shows that the slopes of the regression lines are very similar for almost all of the predictors. There are no significant differences in either the slopes or intercepts that favor the whites or males, except for the EQ Self-Awareness scale whose slope favors males. Therefore, the test battery is equally valid for all groups. In addition, the intercepts for males and whites are above the intercepts for females, blacks and Hispanics for every predictor. Thus, there is no evidence of unfairness whatsoever.

The absence of significant differences between intercepts (at the cut score*)* in Table 5.6.6 shows that the minority group's intercept (at the cut score) was never significantly above the reference group's intercept. In fact, the reverse was often true. That is, for many predictors, the performance of the minority group was *over*-predicted by the predictor score. The degree of overprediction was greatest for blacks and least for females.

Another way to examine fairness is to see if the group differences are similar in the composite predictor and composite criterion. Table 5.6.7 shows this analysis. Although females, blacks, and Hispanics had lower scores and passing rates on the composite predictor than males and whites, these differences were virtually identical using the criterion scores. None of the discrepancies were statistically significant.

Both the fairness analyses and the comparison of the group differences on the predictor and criterion strongly support the fairness of the final predictor battery score. The slopes among the groups are very similar and the differences in intercepts always favor the minority group. The group differences in terms of passing rates and differences in means are remarkably similar in the predictor compared to the criterion. The fairness analyses provide strong evidence of fairness for the individual tests as well.

---

[17] Linear regression assumes that the standard deviation of the criterion scores is the same at every predictor score. This is called homoscedasdicity. In practice, this assumption is violated to varying degrees. Thus, in theory, the standard error of estimate should equal the standard deviation of the criterion scores at the predictor's cut score—and at every other predictor score as well. In practice, this is only an approximation.

The sample size of each of the groups is an important issue in fairness regressions. If the samples are too small, the analyses will be unable to detect statistically significant evidence of unfairness. Figure 5.6.4 below shows the 95% confidence intervals for the slope. The graph clearly shows the wide confidence band for Hispanics; the moderate bands for females and blacks; and the narrow bands for males, whites, and the entire sample. The slopes at the bottom of all confidence bands are well above zero which shows that the validity is statistically significant for each group.

The power analyses were done to consider the possibility that the analyses were not sensitive enough (i.e., the sample size was too small) to have discovered evidence of unfairness (see Table 5.6.8). From the fairness regressions, the reference groups were compared with the minority groups in terms of their slopes and intercepts. For each pair of slopes and intercepts, the analyses determined how small the difference (i.e., a difference favoring the reference groups) between the groups would have to be in the population to achieve a power level of 80%. A power level of 80% means that, if we ran the analysis for 100 different samples, we would find a statistically significant difference between the two groups (i.e., minority vs. reference group) in 80 of those samples.

The power analyses showed that even relatively small differences between groups would have been detected in our fairness analyses. Due to its smaller sample size, the Hispanic group has the largest detectable differences. Table 5.6.8 shows the sizes of the smallest detectable differences at 80% power and $p < .05$.

DISCUSSION

Although many of the tests, including the final AT-SAT battery score, exhibited differences between groups, there is no reliable evidence that the battery is unfair. The fairness analyses show that the regression slopes are very similar among the groups (white, male, female, black, Hispanic). There are differences among the intercepts (at the cut score), but these differences favor the minority groups. Thus, there is strong evidence that the battery is fair for females, blacks, and Hispanics. These results show that the test battery is equally valid for all comparison groups. In addition, differences in mean test scores are associated with corresponding differences in job performance measures. For all groups, high test scores are associated with high levels of job performance and low scores are associated with lower levels of job performance.

TARGETED RECRUITMENT

As indicated above, the AT-SAT Battery is equally valid and fair for white, African American and Hispanics as well as male and female groups. It was also shown in Chapter 5.5 that there is a strong positive relationship between AT-SAT test scores and job performance as an air traffic controller. At the same time, the FAA has the responsibility to try to have the workforce demographics reflect the population of the nation in spite of mean test score differences between groups. We believe that the solution to the apparent contradictory goals of hiring applicants with the highest potential for high job performance and maintaining an employee demographic profile that reflects the nation's population is to staff the ATCS positions with the use of targeted recruiting efforts. Simply stated, targeting recruiting is the process of searching for applicants who have a higher than average probability of doing well on the AT-SAT test battery and, therefore, have the skills and abilities required for performance as an ATCS. For example, one recruiting effort might focus on schools that attract students with high math ability.

Figure 5.6.5 shows the distribution of AT-SAT scores from the pseudo-applicant sample, including scores for all sample members, females, Hispanics, and African Americans. Two important observations can be made from an examination of Figure 5.6.5. First, there are obvious differences in mean test scores between the various groups. Secondly, there is a high degree of overlap in the test score distributions of the various groups. This high degree of overlap means that there are many individuals from each of the different groups who score above the test cut score. These are the individuals one would seek in a targeted recruiting effort. It should be noted that the targeted recruiting effort needs to be a proactive process of searching for qualified candidates. If no proactive recruitment effort is made, the distribution of applicants is likely to be similar to that observed in Figure 5.6.5.

On the other hand, the potential impact of targeted recruiting on mean test scores is shown in Table 5.6.9. In the total applicant sample, 18.8% of the applicants would likely pass at the 70 cut off. If applicants from the top 10% of the black population were recruited so that they were 6 times more likely to apply, about 15.5% would be expected to pass at the 70 cut off. The change from 3.9% (no targeted recruiting) to 15.5% (with targeted recruiting) represents an increase of about 300% in the black pass rate.

CHAPTER 6

The Relationship of FAA Archival Data to AT-SAT Predictor and Criterion Measures

Carol A. Manning and Michael C. Heil
Federal Aviation Administration, Civil Aeromedical Institute

The FAA Civil Aeromedical Institute (CAMI) has conducted research in the area of air traffic controller selection and training for nearly 3 decades. As a result of this research, CAMI established several Air Traffic Control Specialist (ATCS) data bases that contain selection and training scores, ratings, and measures as well as demographic information and other indices of career progression. The archival data described below were matched with AT-SAT predictor test and criterion performance scores for controllers participating in the concurrent validation study who agreed to have their historical data retrieved and linked with the experimental selection and performance data.

PREVIOUS ATC SELECTION TESTS

The United States ATCS selection process between 1981 and 1992 consisted of two testing phases: (a) a 4 hour written aptitude examination administered by the United States Office of Personnel Management (OPM); and (b) a multi-week screening program administered by the FAA Academy. A description of these tests is presented below.

OPM Test Battery

The OPM test battery included the Multiplex Controller Aptitude Test, the Abstract Reasoning Test, and the Occupational Knowledge Test. The Multiplex Controller Aptitude Test (MCAT) required the applicant to combine visually presented information about the positions and direction of flight of several aircraft with tabular data about their altitude and speed. The applicant's task was to decide whether pairs of aircraft would conflict by examining the information to answer the questions. Other items required computing time-distance functions, interpreting information, and spatial orientation. Performance on the MCAT was reported as a single score. The Abstract Reasoning Test (ABSR) was a civil service examination (OPM-157) that included questions about logical relationships between either symbols or letters. This was the only test retained from the previous Civil Service Commission (CSC) battery in use before 1981. (The other CSC tests were Computations, Spatial Patterns, Following Oral Directions, and a test that slightly resembled the MCAT). The Occupational Knowledge Test was a job knowledge test that contained items related to air traffic control phraseology and procedures. The purpose of using the Occupational Knowledge Test was to provide candidates with extra credit for demonstrated job knowledge.

The MCAT comprised 80% of the initial qualifying score for the OPM battery, while the Abstract Reasoning Test comprised 20%. After these weights were applied to the raw scores for each test, the resulting score was transmuted to a distribution with a mean of 70 and a maximum score of 100. If the resulting Transmuted Composite score (TMC) was less than 70, the applicant was eliminated from further consideration. If, however, the applicant earned a TMC of 70 or above, he or she could receive up to 15 extra credit points (up to a maximum score of 100) based upon the score earned on the Occupational Knowledge Test (OKT). Up to 10 extra credit points (up to a maximum score of 110) could also be added based on Veteran's Preference. The sum of the TMC and all earned extra credit points was the final OPM Rating.

This version of the OPM ATCS battery was implemented in September 1981, just after the Air Traffic Controller strike. For some time after the strike, applicants were selected using either a score on the earlier CSC battery or on the later OPM battery. Because of concerns about artificial increases in test scores as a function of training, changes were made in October 1985 to 1) replace the versions of the MCAT that were used, 2) change the procedures used to administer the MCAT, and 3) change eligibility requirements for re-testing.

Academy Nonradar Screening programs

Because tens of thousands of people applied for the job of Air Traffic Control Specialist (ATCS), it was necessary to use a paper-and-pencil format to administer the CSC/OPM batteries. With paper-and-pencil testing, it was difficult to measure aptitudes that would be utilized in a dynamic environment. Consequently, there

continued to be a high attrition rate in ATCS field training even for candidates who successfully completed the initial selection process (earning a qualifying score on the CSC/OPM selection battery, and passing both a medical examination and a background investigation.) In 1975, the Committee on Government Operations authorized the FAA Academy to develop and administer a second-stage selection procedure to "provide early and continued screening to insure prompt elimination of unsuccessful trainees and relieve the regional facilities of much of this burden."

In January of 1976, two programs were introduced at the FAA Academy to evaluate students' ability to apply a set of procedures in an appropriate manner for the non-radar control of air traffic. From 1976 until 1985, candidates entered either the 12-week En Route Initial Qualification Training program (designed for new hires assigned to en route facilities) or the 16-week Terminal Initial Qualification Training program (designed for new hires assigned to terminal facilities). While both programs were based on non-radar air traffic control, they used different procedures and were applied in different types of airspace. Academy entrants were assigned to one program or the other on a more-or-less random basis (i.e., no information about their aptitude, as measured by the CSC/OPM rating, was used to assign them to an "option" or facility). Those who successfully completed one of the programs went on to a facility in the corresponding option. Those who did not successfully complete one of the programs were separated from the GS-2152 job series.

Both the En Route and Terminal Screen programs contained academic tests, laboratory problems, and a Controller Skills Test. The laboratory problems, each one-half hour in length, required the student to apply the principles of non-radar air traffic control learned during the academic portions of the course to situations in which simulated aircraft moved through a synthetic airspace. Student performance was evaluated by certified air traffic control instructors. Two scores, a Technical Assessment (based on observable errors made) and an Instructor Assessment (based on the instructor's rating of the student's potential) were assigned by the grading instructor for each problem. These assessment scores were then averaged to yield an overall laboratory score for a single problem.

The Controller Skills Test (CST) measured the application of air traffic control principles to resolve air traffic situations in a speeded paper-and-pencil testing situation. The composite score in the program was based on a weighted sum of the Block Average (BA; the average of scores from the academic block tests), the Comprehensive Phase Test (CPT; a comprehensive test covering all academic material), the Lab Average (the average score on the best 5 of 6 graded laboratory problems), and the Controller Skills Test (CST). A composite grade of 70 was required to pass. From 1976 until 1985, the same weights were applied to the program components of both the En Route and Terminal Screen programs to yield the overall composite score: 2% for the Block Average, 8% for the Comprehensive Phase Test, 65% for the Lab Average, and 25% for the CST.

For those candidates entering the Academy after the Air Traffic Controller strike of 1981, the pass rate in the En Route Screen program was 52.3% and the pass rate in the Terminal Screen program was 67.8%. The pass rate in both programs combined was 58.0%. In October of 1985, the two programs were combined to create the Nonradar Screen program. The purpose of using a single program was to allow facility assignments to be based, when possible, upon the final grade earned in the program. The Nonradar Screen program was based upon the En Route screen program (containing the same lessons and comparable tests and laboratory problems). It was necessary to change the weights applied to the individual component scores of the Nonradar Screen program to maintain the average pass rate obtained for both the En Route and Terminal screen programs. The weights used in the Nonradar Screen program to yield the overall composite score were: 8% for the Block Average, 12% for the Comprehensive Phase Test, 60% for the Lab Average, and 20% for the CST. The pass rate for the Nonradar Screen program was 56.6%.

The Pre-Training Screen

In 1992, the Nonradar Screen program was replaced with the Pre-Training Screen (PTS) as the second-stage selection procedure for air traffic controllers. The goals of using the PTS were to 1) reduce the costs of ATCS selection (by reducing the time required for screening controllers from approximately 9 weeks to 5 days), 2) maintain the validity of the ATCS selection system, and 3) support agency cultural diversity goals. The PTS consisted of the following tests: Static Vector/Continuous Memory, Time Wall/Pattern Recognition, and Air Traffic Scenarios Test. Broach & Brecht-Clark (1994) conducted a predictive validity study using the final score in the ATCS screen as the criterion measure. They found that the PTS added 20% to the percentage of variance explained in the Nonradar Screen Program final

score, over and above the contribution made by the OPM test. Broach & Brecht-Clark (1994) also described a concurrent validation study conducted using 297 developmental and Full Performance Level (FPL) controllers. The criterion used for this study was a composite of supervisor ratings and times to complete field training, along with performance in the Radar Training program. The corrected multiple correlation between PTS final score and the training composite score was .25 as compared with .19, which was the multiple correlation between the ATCS screen score and the training composite.

Radar Training (Phase XA)

A second screening program, the En Route Basic Radar Training Course (otherwise known as RTF), was administered to en route developmentals who had completed their Radar Associate/Nonradar on-the-job training. The RTF course was a pass/fail course, and developmentals who did not pass were unable to proceed in further radar training at their facilities unless they recycled and later passed the course. However, the pass rate in this phase of training exceeded 98%. The RTF course paralleled the Nonradar Screen program, including an average grade on block tests (2% of the final grade), a comprehensive phase test (8% of the final grade), an average grade for laboratory evaluations (65% of the final grade), and a Controller Skills Test (25% of the final grade.)

OTHER ARCHIVAL DATA OBTAINED FOR ATC CANDIDATES

Biographical Questionnaire

Additional information about controller demographics and experience was obtained from data provided by Academy entrants during the first week they attended one of the Academy screening programs and obtained from the Consolidated Personnel Management Information System (CPMIS). New entrants completed a Biographical Questionnaire (BQ). Different BQ items were used for those entering the Nonradar Screen Program at various times. The BQ questions concerned the amount and type of classes taken, grades earned in high school, amount and type of prior air traffic and/or aviation experience, reason for applying for the job, expectations about the job, and relaxation techniques used.

VanDeventer (1983) found that the biographical question related to grades in high school mathematics courses loaded .31 on a factor defined by pass/fail status in the Academy screening program. Taylor, VanDeventer, Collins, & Boone (1983) found that, for a group of 1980 candidates, younger people with higher grades in high school math and biology, pre-FAA ATC experience, and fewer repetitions of the CSC test, and a self-assessment of performance in the top 10% of all controllers were related to an increased probability of passing the Nonradar Screen program. Collins, Manning, & Taylor (1984) found that, for a group of trainees entering the Academy between 1981 and 1983, the following were related to pass/fail status in the Nonradar Screen program: higher grades in high school math, physical science, and biology classes, a higher overall high school grade point average, younger age, not being a member of the armed forces, taking the OPM test only one time, expectations of staying in ATC work more than 3 years, and a self-assessment that the trainee's performance would be in the top 10% of all ATCSs were positively related to pass/fail status. Collins, Nye, & Manning (1990) found, for a group of Academy entrants between October 1985 and September 1987, that higher mathematics grades in high school, higher overall high school grade point average, self assessment that less time will be required to be effective as an ATCS, self-assessment that the trainee's performance level will be in the top 10% of all ATCSs, and having taken the OPM test fewer times were related to pass/fail status in the Academy screening program.

16PF and Experimental Tests

Also available were scores from the Sixteen Personality Factor (16PF), which is administered during the medical examination and scored with a revised key (Cattell & Eber, 1962; Convey, 1984; Schroeder & Dollar, 1997). Other tests and assessments were administered during the first week of the Academy screening programs; however, they were often administered to a limited number of classes. Consequently, these tests would have been taken by only a few of the controllers who passed the Academy, became certified in an en route facility, and eventually participated in the concurrent validation study. Only the Mathematics Aptitude Test was taken by a sufficient number of participants to include in these analyses.

ARCHIVAL CRITERION MEASURES

Field Training Performance Measures as Criteria

Description of En Route ATCS Field Training

In the en route option, the unit of air traffic control operation is the sector, a piece of airspace for which a team of 2-3 controllers is responsible (during times of slow traffic, only one controller may be responsible for a sector). A group of between 5-8 sectors is combined into what is called an area of specialization. An en route controller is assigned to only one area of specialization, but is responsible for controlling traffic for all sectors within that area. The team of en route controllers working at a sector handles duties related to: Radar separation of aircraft (radar duties; including formulating clearances to ensure separation and delivering them by radio to pilots, handing off responsibility for an aircraft to another controller); assisting the radar controller (radar associate duties; including maintaining records about clearances that have been issued or other changes in the flight plan of an aircraft, identifying potential problems, communicating information not directly related to aircraft separation of aircraft to pilots or other controllers); or supporting other activities (assistant controller duties; including entering data into the computer, ensuring that all records of flight progress are available for the controller in charge).

En route controllers are usually trained as assistant controllers first, then given training on increasingly difficult responsibilities (radar associate duties, then radar). Training on concepts is conducted in the classroom, before being applied in a laboratory setting, and then reinforced during on-the-job training (OJT), which is conducted in a supervised setting. At some facilities, all radar associate training is completed before radar training begins. At other facilities, training is conducted by position: Both radar associate and radar training are provided for a specific position before training begins on the next position. At one point in time, en route controllers could have taken up to 9 phases of field training, depending on the way training was provided at the facility.

Measures of Performance in Field Training

Several measures of training performance were obtained for each phase of air traffic control field training. For each phase of training, the start and completion dates, the number of hours used to complete on-the-job training (OJT), the grade (Pass, Fail, or Withdraw), and a rating of controller potential, measured on a 6-point scale, (provided by an instructor or supervisor who most frequently observed the student during that phase) were collected. This information was compiled to derive measures of training performance, such as the amount of time (in years) required to reach full performance level (FPL) status, mean instructor ratings of potential computed for OJT phases (called the Indication of Performance), the amount of time (in calendar days) required to complete OJT in certain training phases, and the total number of OJT hours required to complete those phases. Data were used from only phases IX and XII because those phases included the first two sectors on which nonradar/radar associate (Phase IX) and radar (Phase XII) training were provided.

These measures of training performance were collected because they were readily available for most trainees, but a number of outside factors besides aptitude and technical proficiency could have affected their value. Time required to reach FPL status could be affected by delays in training caused by a number of factors, including the need for management to use a trainee to control traffic on sectors on which he/she had already certified instead of allowing him/her to participate in OJT, the number of other students undergoing OJT in the same airspace at the same time (limiting an individual's access to OJT), or the number of trainees, (affecting the availability of the training simulation laboratory). The number of OJT hours required to certify on a specific sector could be affected by the type of traffic the student controlled during training or the difficulty of the sector. The subjective rating of trainee potential could be affected by a number of rating biases familiar to psychologists, such as halo, leniency, etc. In spite of the measurement problems associated with these training performance measures, they were the best measures available for many years to describe performance in ATCS technical training programs.

HISTORICAL STUDIES OF VALIDITY OF ARCHIVAL MEASURES

Brokaw (1984) reviewed several studies examining the relationship between aptitude tests and performance in both air traffic control training and on the job. He described an early study (Taylor, 1952) that identified a set of 9 tests having zero-order correlations of .2 or above with supervisor job performance ratings or composite criteria. A selection battery that included the following tests was recommended but not implemented: Memory

for Flight Information, Air Traffic Problems I & II, Flight Location, Coding Flight Data I, Memory for Aircraft Position, Circling Aircraft, Aircraft Position, and Flight Paths.

A more extensive study was performed during a joint Air Force Personnel Laboratory and Civil Aeronautics Administration collaboration (Brokaw, 1957). Thirty-seven tests were administered to 130 trainees in an ATC school. Criteria were based on performance in the ATC course, including grades for the lecture, instructor ratings, and a composite of ratings from multiple instructors. Tests related to one or more of the training criteria involved Computational and Abstract Reasoning (including Dial and Table Reading and Arithmetic Reasoning tests), Perceptual and Abstract Reasoning, Verbal Tests, Perceptual Speed and Accuracy, and Temperament. The multiple correlation of four tests (Air Traffic Problems, Arithmetic Reasoning, Symbol Reasoning and Perceptual Speed, and Code Translation) with the instructor rating was .51.

A follow-up study (Brokaw, 1959) was conducted to examine the relationship between the experimental selection battery and supervisor ratings of on-the-job performance. The multiple correlation of the same four tests with the supervisor rating was .34. Trites (1961) conducted a second follow-up study using Brokaw's 1957 sample, obtaining supervisor ratings after hire. Symbolic Reasoning and Perceptual Speed, Abstract Reasoning (DAT), Space Relations (DAT), and Spatial Orientation (AFOQT), were all significantly related to supervisor ratings provided in 1961 (correlations were .21, .18, .18, and .23, respectively.) The correlations were reduced somewhat when partial correlations were computed holding age constant. Furthermore, the Family Relations Scale from the California Test Bureau (CTB) California Test of Personality had a .21 correlation with the 1961 supervisor ratings. The correlation was not reduced by partialing out the effect of age.

Trites & Cobb (1963), using another sample, found that experience in ATC predicted performance both in ATC training and on the job. However, aptitude tests were better predictors of performance in training than was experience. Five aptitude tests (DAT Space Relations, DAT Numerical Ability, DAT Abstract Reasoning, CTMM Analogies, and Air Traffic Problems) had correlations of .34, .36, .45, .28, and .37 with academic and laboratory grades, while the correlations with supervisor ratings were lower (.04, .09, .12, .13, and .15, respectively) for en route controllers.

Other studies have examined relationships between experimental tests and performance in the FAA Academy Screening Program. Cobb & Mathews (1972) developed the Directional Headings Test (DHT) to measure speeded perceptual-discrimination and coding skills. They found that the DHT correlated .41 with a measure of training performance for a group of air traffic control trainees who had already been selected using the CSC selection battery. However, the test was highly speeded, and was consequently difficult to administer.

Boone (1979), in a study using 1828 ATC trainees, found that the Dial Reading Test (DRT; developed at Lackland AFB for selecting pilot trainees) and the DHT had correlations of .27 and .23, respectively, with the standardized laboratory score in the Academy screen program. An experimental version of the MCAT correlated .28 with the lab score. In the same study, CSC 24 (Computations) and CSC 157 (Abstract Reasoning) correlated .10 and .07, respectively, with the laboratory score.

Schroeder, Dollar & Nye (1990) administered the DHT and DRT to a group of 1126 ATC trainees after the air traffic control strike of 1981. They found that the DHT correlated .26 (.47 after adjustment for restriction in range) with the final score in the Academy screening program, while the DRT correlated .29 (.52 after adjustment for restriction in range) with the final score in the Academy screening program. MCAT correlated .17 and Abstract Reasoning correlated .16 with the final score, though those two tests had been used to select the trainees.

Manning, Della Rocco, and Bryant, (1989) found statistically significant (though somewhat small) correlations between the OPM component scores and measures of training status, instructor ratings of trainee potential, and time to reach FPL (a negative correlation) for 1981-1985 graduates of the en route Academy screening program. Correlations (not corrected for restriction in range) of the MCAT with training status, OJT hours in Phase IX, mean Indication of Performance for Phases VIII-X, OJT hours in Phase XII, Indication of Performance in Phases XI-XIII, and time to FPL were -.12, .05, .11, .08, .11, and -.11, respectively. Correlations (not corrected for restriction in range) of the Abstract Reasoning Test with the same measures of field training performance were .03, .04, .03, .09, .01, and -.02, respectively.

Manning et al. also examined correlations between component scores in the en route Academy screening program and the same measures of field training performance. Correlations (not corrected for restriction in

range) of the Lab Average with training status, OJT hours in Phase IX, Indication of Performance in Phases VIII-X, OJT hours in Phase XII, Indication of Performance in Phase XII, and Time to FPL were -.24, -.06, .23, -.12, .24, and -.16, respectively. Correlations (not corrected for restriction in range) of the Nonradar Controller Skills Test with the same training performance measures were -.08, -.02, .11, 0, .07, and -.09. Correlations (not corrected for restriction in range) of the Final Score in the Screen with the same training performance measures were -.24, -.06, .24, -.10, .24, and -.18, respectively.

Manning (1991) examined the same relationships for FY-96 graduates of the ATC screen program, assigned to the en route option. Correlations (not corrected for restriction in range) of the MCAT, Abstract Reasoning Test, and OPM rating with status in field training were .09, .03, and .09, respectively. When adjusted for restriction in range, these correlations were .24, .04, and .35, respectively. Correlations (not corrected for restriction in range) of the Lab Average, Controller Skills Test, and Final Score in the Screen with status in field training were .21, .16, and .24, respectively. When adjusted for restriction in range, these correlations were .36, .26, and .44, respectively.

RELATIONSHIPS BETWEEN ARCHIVAL DATA AND AT-SAT MEASURES

Relationship of Archival and AT-SAT Criterion Measures

It is expected that the measures of field training performance used during the 1980s as criterion measures to assess the validity of the OPM test and Academy screening programs will also be significantly correlated with the AT-SAT criterion measures. The magnitude of these correlations might be lower than those computed among the original archival measures because several years have elapsed between the time when field training occurred and the administration of the AT-SAT criterion measures.

Table 6.1 shows correlations between the archival criterion measures and the AT-SAT criterion measures. These correlations have not been adjusted for restriction in the range of the training performance measures. Correlations between days and hours in the same phase of training were high, and correlations between days and hours in different phases of training were moderate. Correlations between the Indication of Performance and time in the same or different phases of training were non-significant, but the correlation between the Indication of Performance in Phase IX and the Indication of Performance in Phase XII was moderately high.

Correlations between time in training phases and the composite criterion rating were statistically significant at the .01 level, but were not very high. The CBPM was significantly correlated with only the days and hours in Phase XII, which described the outcome of training on the first two radar sectors. It makes sense that the CBPM would relate particularly to performance in radar training because the CBPM contains items based on radar concepts. Correlations of both the ratings and the CBPM with the Indication of Performance variables were either non-significant or not in the expected direction (i.e., correlations of AT-SAT criteria with the indication of performance variables should be positive while correlations with training times should be negative.)

Relationship of Archival Predictors with Archival and AT-SAT Criterion Measures

Because the archival and AT-SAT criterion measures are related, and because the ATCS job has changed little in the last 15 years, the selection procedures previously used by the FAA and the AT-SAT criterion measures should be correlated. The following two tables show relationships of the OPM rating and performance in the Academy screen program with both the archival and AT-SAT criterion measures. It should be remembered that the controllers who participated in the concurrent validation study were doubly screened—first on the basis of their OPM rating, then, second on the basis of their score in the Academy Screen program. Current FPLs were also reduced in number because some failed to complete training successfully. Thus, there is considerable restriction in the range of the selection test scores.

Table 6.2 shows correlations of the archival selection test scores (OPM Rating, final score in the Nonradar Screen program, and final score in the Radar Training program) with both the archival criterion measures and the AT-SAT criterion measures. Correlations adjusted for restriction in the range of the predictors are in parentheses after the restricted correlations. The OPM rating correlated .18 with the final score in the Nonradar Screen program and .11 with the final score in the Radar Training program. The OPM rating had very low correlations with archival criterion measures (although it was significantly correlated with the Indication of Performance in initial radar training.) The OPM rating was not significantly correlated with the rating composite, but was significantly correlated with the CBPM score

(r = .22.) The final score in the Nonradar Screen program was significantly correlated with training times in both phases of field training and with time to reach FPL status, but not with either Indication of Performance measure. The final score in the Nonradar Screen program was also significantly correlated with both AT-SAT criterion measures, although the correlation with the CBPM (.34) was much higher than the correlation with the rating composite (.12). The final score in the Radar Training program was also significantly correlated with training times, and was significantly correlated with the Indication of Performance for initial radar training. It was also significantly correlated with both the AT-SAT rating composite (.17) and the CBPM score (.21).

Table 6.3 shows correlations of the performance-based components of the archival selection procedures (Nonradar Screen program and Radar Training program) with both the archival and AT-SAT criterion measures. The correlations at the top of the table are intercorrelations between archival selection procedure components. Of the OPM component scores, only the Abstract Reasoning Test and the MCAT were significantly correlated.

Correlations of components of the OPM battery with component scores from the Nonradar Screen program and the Radar Training program were fairly low, although some statistically significant correlations with scores from the laboratory phases were observed. The MCAT was significantly correlated with Instructor Assessment and Technical Assessment from both the Nonradar Screen and Radar Training programs, and was significantly correlated with the Nonradar CST. Abstract Reasoning was significantly correlated with only the nonradar Average Technical Assessment and the nonradar CST. The OKT had a small but statistically significant correlation with the Nonradar Average Instructor Assessment.

The correlation between the Average Instructor Assessment and Average Technical Assessment from each course was very high (.79 and .83, for the Nonradar Screen program and Radar Training program, respectively.) Across programs the Average Instructor Assessment and Average Technical Assessment had significant correlations that ranged between about .02 and .35. The Controller Skills Tests for both courses had significant correlations with the Nonradar Average Technical and Average Instructor Assessment. While the Nonradar CST was significantly correlated with the Radar Average Instructor and Technical Assessments, the Radar CST was not. Correlation between CSTs was only .25, which was similar to correlations with other components of the Nonradar Screen and Radar Training programs.

Correlations of OPM component scores with the rating criterion measure were all low and non-significant. However, the MCAT and Occupational Knowledge Tests were both significantly correlated with the CBPM score.

Of the components of the Nonradar Screen and Radar Training programs, the Average Technical Assessment had significant negative correlations with training times (though not with the Indication of Performance measures). The Radar Technical Assessment was correlated both with time spent in Radar Associate and Radar field training phases, while the Nonradar Technical Assessment was only correlated with time spent in Radar field training phases. Both were significantly correlated with the Time required to reach FPL status. The Radar Average Instructor Assessment was significantly correlated with time spent in Radar Associate field training. Interestingly, the Nonradar Average Instructor Assessment was not related to time in phases of field training, although its correlation with the Nonradar Average Technical Assessment was about .8. Both the Nonradar and Radar Average Instructor Assessment were significantly correlated with time to reach FPL status.

The Nonradar and Radar Average Technical Assessments and Average Instructor Assessments were all significantly related to the CBPM score, though only the Nonradar Average Instructor Assessment was significantly related to the rating composite. Both the Nonradar and Radar Controller Skills Tests were significantly correlated with the CBPM. This relationship is not surprising because the CSTs and CBPM have similar formats: They all present a sample air traffic situation and ask the respondent to answer a multiple choice question (under time pressure) involving the application of ATC procedures. The CSTs were presented in a paper-and-pencil format while the CBPM was presented using a dynamic computer display.

Relationship of Archival Criteria and High-Fidelity Simulation Criteria

Table 6.4 shows correlations of the criterion measures obtained from the high-fidelity simulation (comprising 107 participants) with archival performance-based predictor and archival criterion measures. The high-fidelity criterion measures used in this analysis included the individual scales used in the Over-the-Shoulder rating

form. Also used was the number of operational errors made during the 7th graded scenario, the most complex scenario included in the simulation test. The high-fidelity rating scales were correlated very highly with each other (.80 and above). The number of operational errors made in the 7th graded scenario was correlated -.20 to -.32 with the high fidelity rating scales, which were based on performance in all 7 graded scenarios. The high-fidelity rating scales (based on assessments of maximum performance) had correlations of about .35 to about .40 with the AT-SAT rating composite (based on assessments of typical performance), and had correlations of about .60 to .65 with the CBPM. The number of operational errors made in the 7th graded scenario was not significantly correlated with either the AT-SAT rating composite or the CBPM.

The high-fidelity rating scales were not correlated with either Indication of Performance measure obtained from field training records. OJT hours in Phase IX (Radar Associate/Nonradar training) had significant negative correlations with several individual high-fidelity rating scales, including the overall rating. OJT hours in Phase XII (field Radar training) had significant negative correlations with all high-fidelity ratings scales except Coordination. Time to reach FPL status had significant negative correlations with only Maintaining efficient air traffic flow and with Attention & Situation Awareness.

The high-fidelity rating scales had higher, significant, correlations with some of the performance-based components of the archival selection procedures. The high-fidelity rating scales were correlated between about .35 and .40 with the Average Instructor Assessment from the Nonradar Screen program, and were correlated between about .5 and .55 with the Average Technical Assessment from the Nonradar Screen program. There were only two significant correlations between the Controller Skills Test from the Nonradar Screen program and the high-fidelity rating scales (Coordination and Managing Sector Workload). The high-fidelity rating scales had almost no correlation with the Average Instructor Assessment from the Radar screen program but were correlated between about .55 and .60 with the Average Technical Assessment from the Radar screen program. Performance on the Controller Skills Test from the Radar screen program was correlated between about .60 and .71 with the high-fidelity rating scales. Though many of these correlations are statistically significant, they were typically based on fewer than 60 participants who allowed their archival data to be matched with their performance in the AT-SAT testing and the high fidelity simulation testing. At the same time, it is interesting to observe correlations of the magnitude seen here between measures of performance from simulations that occurred recently and from performance-based selection procedures that occurred between 5 and 15 years previously.

Relationship of Archival Measures and AT-SAT Predictors

It was also expected that archival measures, including archival selection tests and scores on experimental tests administered at the FAA Academy during the first week of the Academy screen program might have high correlations with AT-SAT predictor tests. High correlations between AT-SAT predictors and other aptitude tests should provide evidence supporting interpretations of the construct validity of the AT-SAT tests. The magnitude of these correlations might be reduced, however, because the experimental tests were administered between 5 and 15 years prior to the concurrent validity study and the OPM test was probably administered between 6 and 16 years previously.

An analysis was conducted to compute correlations between scores on the OPM selection tests: the Multiplex Controller Aptitude Test (MCAT), the Abstract Reasoning Test, and the Occupational Knowledge Test (OKT), and the AT-SAT predictor tests. The MCAT, the highest weighted component of the OPM rating, required integrating air traffic information to make decisions about relationships between aircraft. Thus, aptitudes required to perform well on the MCAT might be related to aptitudes required to perform well on the Air Traffic Scenarios Test (ATST). Furthermore, the skills required to integrate information when taking the MCAT might be related to performance on the Letter Factories Test, Time Wall, Scan, and Planes tests. Positive correlations of the AT-SAT predictors with the MCAT, a test previously used to select controllers, would provide further evidence of the validity of the tests included in the AT-SAT battery.

Table 6.5 shows correlations of the MCAT, Abstract Reasoning Test, and OKT with the AT-SAT predictor tests. The computed correlations are followed in parentheses by correlations adjusted for restriction in the range of each archival selection test. (Correlations for the OKT were not adjusted for restriction in range because the standard deviation of the OKT after candidates were selected was larger than was its standard deviation before applicants were selected.)

MCAT had significant, but small, correlations with many of the AT-SAT predictor tests: all measures derived from the Letter Factories test, Applied Math, Time Wall Time Estimation Accuracy and Perceptual Accuracy scores (but not Perceptual Speed), Air Traffic Scenarios Efficiency and Safety scores (but not Procedural Accuracy), Analogies Reasoning score (but not Latency or Information Processing), Dials, Scan, both Memory tests, Digit Span, Planes Timesharing score (but not Projection or Dynamic Visual/Spatial), and Angles.

Abstract Reasoning was also significantly correlated with several of the AT-SAT predictor tests. The relationship of the most interest is with the component scores of the Analogies test. Abstract Reasoning might be expected to have a high correlation with Analogies because many items in both tests are similar. Thus, it is not surprising to observe a correlation of .33 between Abstract Reasoning and the Analogies: Reasoning score. However, the correlation of Abstract Reasoning with the Latency and Information Processing components was non-significant. Abstract Reasoning also correlated with other AT-SAT predictor tests: all Letter Factories subscores, Angles, Applied Math, Time Wall: Time Estimation Accuracy and Perceptual Accuracy (but not Perceptual Speed), both Memory tests, Dials, Scan, and AT Scenarios: Efficiency and Safety (but not Procedural Accuracy).

The Occupational Knowledge Test measured the knowledge about aviation and air traffic control that applicants brought to the job. The OKT had several significant correlations with AT-SAT predictor tests, although all but one was negative, implying that controllers who entered the occupation with less knowledge of ATC performed better on the AT-SAT aptitude tests. OKT was negatively correlated with Letter Factories Situational Awareness and Planning & Thinking ahead scores (but was not significantly correlated with number of letters correctly placed), both memory tests, Time Wall Perceptual Accuracy score, and Planes Dynamic Visual/Spatial score. OKT had a significant positive correlation with AT Scenarios Procedural Accuracy score.

Although many of these correlations are statistically significant, they are nevertheless small, which might appear to suggest that they do not provide evidence of the construct validity of the AT-SAT predictor tests. Moreover, most of the correlations continued to be rather small after they were adjusted for restriction in the range of the archival selection tests. However, it must be remembered that the participants in the concurrent validity study were doubly (and even triply) selected, because they first qualified on the basis of their performance on the OPM test, then by passing the Nonradar Screen program (which had approximately a 40% loss rate), then again by passing field training (which had approximately an additional 10% loss rate). Thus, even making one adjustment for restriction in range does not compensate for all the range restriction that occurred. Furthermore, performance on the AT-SAT predictor tests may have been influenced by age-related effects.

Archival Experimental Tests and AT-SAT Predictors. The next analysis examined the relationship of the Dial Reading Test (DRT), the Directional Headings Test (DHT), and two other archival measures of mathematical aptitude with AT-SAT predictor tests. The Dial Reading Test is a paper-and-pencil version of the computerized AT-SAT Dials test, and so it would be expected that scores would be highly correlated. The DHT was an experimental test administered to ATC trainees during the 1970s. the DHT required comparing three pieces of information: A letter (N, S, E, or W), a symbol (^, v, <, or >), and a number (0 to 360 degrees), all indicating direction, in order to determine whether they indicated a consistent or inconsistent directional heading. A second part of the test required determining the opposite of the indicated direction. Thus, performance on the DHT might be expected to correlate positively with both Angles and Applied Math.

The Math Aptitude Test was taken from the Educational Testing Service (ETS) Factor Reference Battery (Ekstrom, French, Harman, & Derman, 1976). An item dealing with reported grades in high school math courses was also included in the analysis because this biographical information was previously found to be related to success in the Nonradar Screen program.

Although these tests were administered between 5 and 15 years before the concurrent validation study, it is expected that the DHT and DRT would be at least moderately related to performance on some of the AT-SAT predictor tests, especially those related to mathematical skills. It may be remembered that in past research, the DHT and DRT had moderate correlations with criterion measures of performance in ATC training. Thus, positive correlations of the AT-SAT predictors with the DHT and DRT would provide further evidence of the validity of the AT-SAT tests.

Table 6.6 shows the relationship of three AT-SAT predictor tests with DHT, DRT, the Math Aptitude Test, and a biographical item dealing with high school math grades. Numbers of respondents are shown in parentheses after the correlation coefficient. As expected, Applied Math had a high, positive correlation with the

Math Aptitude Test total score (.63). Applied Math had also statistically significant and reasonably high positive correlations with Dial Reading Number Correct (.52) and Directional Headings Number Correct Part 2 (.40). Applied Math also had moderate, significant negative correlations with Dial Reading Number items wrong (-.36) and the biographical item dealing with high school math grades (-.34).

Angles was significantly correlated with Dial Reading Number Correct (.37) and Dial Reading Number Wrong (-.28). Angles was also significantly correlated with the Math Aptitude Test (.41) and the biographical item dealing with high school math grades (-.21). Unexpectedly, Angles had a small positive (but significant) correlation with Directional Headings number wrong Part 2 (.18).

The results of the comparison of the Dials test and the archival experimental tests was somewhat surprising. Dials had a significant positive correlation with Dial Reading number correct (.22) and a significant negative correlation with Dial Reading number wrong (-.39). However the correlation with Dial Reading number correct was low, considering that Applied Math and Angles had higher correlations than did Dials. However, Dials did not contain all the same items as Dial Reading. After the Alpha testing, certain items present in Dial Reading were removed from Dials, and other items were inserted. Moreover, Dial Reading was presented in a paper-and-pencil format while Dials was presented in a computerized format. One might speculate that the different formats were responsible for the reduced correlation. However, it must be remembered that Dial Reading Test was administered between 5 and 15 years prior to the administration of Dials, and considerable training and aging occurred during the interim. While air traffic controllers in the en route environment may not read dials, they are trained extensively on other tasks involving perceptual speed and accuracy, which is an aptitude that the Dials test is likely to measure. Thus, it is more likely that the low correlation between Dial Reading and Dials resulted from changes in the items, and the effects of time and aging on the individuals taking the test, rather than a change in the format of the test.

Pre-Training Screen and AT-SAT Predictors. In 1991, a group of 297 developmental and FPL controllers participated in a study assessing the validity of the Pre-Training Screen (Broach & Brecht-Clark, 1994). Sixty-one controllers who participated in the concurrent validation of the PTS also participated in the AT-SAT concurrent validation in 1997/1998.

Scoring algorithms used for the PTS version of the ATST differed from those used for the AT-SAT version of the ATST. In the PTS version, the Safety score was a count of safety-related errors and Delay Time measured the amount of time aircraft were delayed. For both the Safety score and Total Delay Time, higher scores indicated worse performance. In the AT-SAT version, the Safety and Efficiency scores were based on counts of errors and measurement of delays, but both variables were transformed so that higher scores indicated better performance. Procedural Accuracy is a new variable based on the occurrence of errors not related to safety. It is expected that the PTS Safety Score would be more highly correlated with the AT-SAT Safety score than with the AT-SAT Efficiency Score and that PTS Total Delay Time would be more highly correlated with the AT-SAT Efficiency Score than with the AT-SAT Safety Score. It is also expected that the two PTS scores would have significant negative correlations with the three AT-SAT scores.

Table 6.7 shows the relationship of the scores from the version of the Air Traffic Scenarios Test included in the Pre-Training Screen with the version of the Air Traffic Scenarios Test included in AT-SAT. As expected, the PTS Safety Score is more highly correlated with the AT-SAT Safety Score than with the AT-SAT Efficiency Score (and those correlations are negative). Also, the correlation between the PTS Average Total Delay Time and AT-SAT Efficiency Score was both significant and negative. The Procedural Accuracy score from the AT-SAT version had little relationship with either PTS ATST score.

Archival Data and the Experience Questionnaire. The merging of the archival data with the AT-SAT concurrent validation data provided an opportunity to investigate the construct validity of the personality test contained in the AT-SAT battery. Construct validity of the Experience Questionnaire (EQ) was investigated using the following methods: principal component analysis to determine structure of the scale; and Pearson product-moment correlations to determine the degree of convergence and divergence with archival 16PF data. The 167 items contained in the EQ were used to calculate 14 personality scales, which were used in the analyses.

In terms of the principal components analysis, a final solution revealing at least two independent factors would provide evidence that the EQ scales measure unique constructs. Relationships between some of the EQ scales

would be anticipated, therefore, certain scales should load on the same factor. However, some scales should be unrelated, meaning that they should load on different factors. For example, "taking charge" and "decisiveness" are likely to be related and therefore load together on a factor. The variable "concentration", on the other hand, would not be expected to have a high degree of relationship with these other two variables and should load on a different factor. An oblique principal components analysis was conducted using data collected during the AT-SAT concurrent validation study. As shown in Table 6.8, the principal components analysis resulted in the extraction of only two factors. The first factor accounts for 56% of the variance, whereas the second factor accounts for only 9.49%. Additionally, these two factors are correlated with each other (r=.54). These results suggest that the variance in EQ scores is best explained by one primary factor, although a small percentage is explained by a related factor. For the most part, the EQ scales are related to each other even when they should theoretically be distinct. The results of this principal components analysis fail to provide support for the independence of the EQ scale scores.

Further support for the construct validity of the EQ was sought by comparing scale scores to archival 16PF scores. Although the 16PF is not necessarily the standard by which all personality tests are measured, it is, in fact, an established measure of personality traits that is widely used in clinical and experimental settings. The merging of these two data bases resulted in 451 usable cases. A description of the 16PF scales included in the analyses is provided in Table 6.9. Certain relationships would be expected to exist between scores from the two tests. Specifically, there would be support for the construct validity of the EQ scales if they correlate with 16PF scales that measure a similar construct. Conversely, the EQ scales would be expected to be unrelated to 16PF scales that measure other constructs. Since the 16PF was administered several years before the EQ, these expected relationships are based on the assumption that measurement of personality characteristics remains relatively stable over time. This assumption is supported by Hogan (1996) and Costa & McCrae (1988). A summary of the expected relationships between EQ and 16PF scale scores is provided below.

The EQ Composure scales should be positively correlated with 16PF Factor C (emotionally stable), which would indicate that people high in composure are also more emotionally stable and calm. EQ Task Closure and EQ Consistency of work behavior should be positively correlated with 16PF Factor G (conscientiousness). EQ Working Cooperatively should be positively correlated with 16PF Factors A (outgoing and participating) and Q_3 (socially precise) as well as negatively correlated with Factor L and Factor N (which would indicate that these people are trusting and genuine). Furthermore, it would be expected that a high score on EQ Decisiveness and EQ Execution would be negatively correlated with 16PF Factor 0, meaning that decisive people would also be expected to be self-assured and secure. EQ Flexibility should have a positive correlation with 16PF Factor A and a negative correlation with Factor Q_4 (relaxed).

The EQ Tolerance for High Intensity scale would be expected to be positively correlated with 16PF Factor H (Adventurous) and negatively correlated with Factor O (Apprehensive). EQ Self-Awareness and EQ Self-Confidence should both be negatively correlated with 16PF Factor O (Apprehensive). A positive correlation between EQ Self-Confidence and 16PF Factor Q2 (Self-sufficient) might also be expected. EQ Sustained Attention and EQ Concentration should be related to 16PF Factor G (conscientiousness) whereas EQ Taking Charge should be related to 16PF Factor H (Adventurous) and Factor E (Assertive). Finally, EQ Interpersonal Tolerance should be positively correlated with 16PF Factor I (Tender-minded), Factor Q3 (socially precise), and Factor C (Emotionally Stable).

Scores on the EQ and 16PF scales were compared using Pearson product-moment correlations, the results of which are presented in Table 6.10. The results of correlational analyses between the EQ scales shows that they are all inter-related. However, this is not surprising considering the results of the principal components analysis described above. Although relationships between some of the scales contained in a personality measure are not unusual, moderate to high correlations between all of the scales is another matter.

As stated earlier, the EQ scores were compared to 16PF Factor scores in an effort to support construct validity by determining whether or not these scales measure what they are purported to measure. Although statistically significant, the correlations between EQ and 16PF scales represent small effect sizes and are not of the magnitude desired when attempting to support the validity of a test. The statistical significance of these relationships is most likely an artifact of sample size. For the most part, the pattern of relationships with 16PF scales was the same for all EQ scales. This would not be expected if the EQ scales did in fact measure different constructs. This pattern is not unexpected given the EQ

inter-scale correlations and the results of the principal components analysis. The results of these analyses fail to provide evidence that the EQ scales measure unique constructs, let alone the specific constructs they are professed to measure. However, there are indications that the EQ contributes to the prediction of AT-SAT criterion measures (Houston & Schneider, 1997). Consequently, CAMI will continue to investigate the construct validity of the EQ by comparing it to other personality measures such as the NEO PI-R.

Regression of Archival Selection Procedures and AT-SAT Tests on AT-SAT Criteria. A multiple linear regression analysis was conducted to assess the contribution of the AT-SAT tests in predicting the AT-SAT criterion, over and above the contribution of the OPM rating and final score from the Nonradar Screen program. The regression analysis used OPM rating, final score in the Nonradar Screen program, and AT-SAT test scores as predictors, and the weighted composite of AT-SAT criterion measures as the criterion variable. To compute the weighted composite criterion measure, the CBPM received a .6 weighting while the AT-SAT rating composite received a .4 weighting. A stepwise regression was used.

Table 6.11 shows the results of the analysis. A model was identified that contained the following variables: Analogies Reasoning score, final score from the Nonradar Screen program, Applied Math Number Correct, Scan Total score, EQ Unlikely virtues scale, and Air Traffic Scenarios Procedural Accuracy score produced a multiple regression coefficient of .465, accounting for about 22% of the variance in the AT-SAT composite criterion variable. It is interesting that the final score in the Nonradar Screen program contributed so much to the prediction of the criterion measure, because there was considerable restriction in the range of that variable. At least 40% of the candidates failed the Nonradar Screen program and were removed from employment, and another 10% failed field training and were also removed from employment or reassigned to another type of air traffic facility.

It may appear surprising that more of the AT-SAT predictor tests were not included in this model, but they probably accounted for similar parts of the variance in the AT-SAT composite criterion measure that were also accounted for by the final score in the Nonradar Screen program. For example, the Safety and Efficiency scores from Air Traffic Scenarios Test, Applied Math, Angles, the Letter Factories: Number of letters correctly placed, Planning & Thinking Ahead, and Situation Awareness scores, EQ: Composure & Self-Confidence scales all had significant correlations with the final score in the Nonradar Screen program. On the other hand, the Unlikely Virtues scale from the EQ probably tapped a part of the variance in the AT-SAT composite criterion measure that was not already tapped by another AT-SAT predictor test or by the final score in the Nonradar Screen program. The Unlikely Virtues scale will not be included as part of the selection battery, but will be retained to provide information about whether the applicant is faking responses on the rest of the EQ scales.

Discussion

Several analyses were conducted to examine interrelationships between archival selection tests, archival criterion measures, and experimental tests administered to candidates entering the Academy for the Nonradar Screen program. The purpose of these analyses was to assess the construct validity of the AT-SAT criterion measures and predictors. The results of the analyses supported the interpretation of the AT-SAT measures discussed in other chapters of this report.

For example, the amount of time required to complete various phases of field training, which were used as archival criterion measures, were related to the AT-SAT rating composite. Also, the OPM rating, the final score in the Nonradar Screen program, and the final score in the Radar Training program, were all significantly correlated with the CBPM. The final score in the Nonradar Screen program and the final score in the Radar Training program were both significantly correlated with the AT-SAT rating composite. Also, the component tests of the OPM Battery, the Nonradar Screen program, and the Radar Training program all had significant correlations with the CBPM. Furthermore, all scales from the Over-the-shoulder rating form used in the high-fidelity simulation (which were significantly correlated with both the CBPM and the AT-SAT rating composite) were also significantly correlated with both the Instructor Assessment and Technical Assessment ratings made during both the Nonradar Screen program and the Radar Training program. These results suggest that the CBPM and the composite ratings are related to measures used in the past as criterion measures of performance in air traffic control.

Additional analyses suggest that the AT-SAT predictors are also related to tests previously used to select air traffic controllers. The MCAT was correlated with many

of the AT-SAT predictor tests, especially those involving dynamic activities. The Abstract Reasoning test had a particularly high correlation with the Analogies Reasoning score, but was also correlated with other AT-SAT predictors.

Other tests, administered experimentally to air traffic control candidates between the years of 1981 and 1995, provided additional support for the construct validity of AT-SAT predictor tests. For example, the Math Aptitude Test from the ETS Factor Reference Battery (Ekstrom et al., 1976), the Dial Reading Test, and a biographical item reporting high school math grades (which was previously shown to be correlated with success in the Nonradar Screen program) had high correlations with the Applied Math Test. The Angles and Dials tests were also correlated with Dial Reading, Math Aptitude, and the biographical item reporting high school math grades. These results are not surprising, considering the consistent relationship, observed over years of research, between aptitude for mathematics and various measures of performance in air traffic control.

Finally, a multiple linear regression analysis was conducted which showed that several of the AT-SAT tests contributed to the prediction of the variance in the AT-SAT composite criterion measure over and above the OPM rating and the final score in the Nonradar Screen program. The OPM battery and Nonradar Screen program provided an effective, though expensive, two-stage process for selecting air traffic controllers that was used successfully for many years. It appears that the AT-SAT battery has equivalent, or better, predictive validity than did the former selection procedure, and costs much less to administer. Thus, it should be an improvement over the previous selection process.

To maintain the advantage gained by using this new selection procedure, it will be necessary to monitor its effectiveness and validity over time. This will require developing parallel forms of the AT-SAT tests, conducting predictive validity studies, developing and validating new tests against criterion measures of ATC performance, and replacing old tests with new ones if the former become compromised or prove invalid for any reason.

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.