UPDATING THE PERSONNEL SELECTION SYSTEM FOR BEHAVIOR DETECTION OFFICERS

VALIDATION STUDY
VOLUME I: TECHNICAL REPORT

20 MAY 2012

PREPARED FOR:
U.S. Naval Research Laboratory
Washington, DC

UNDER CONTRACT:
N00173-11-F-0494

SPONSORED BY:
U.S. Department of Homeland Security
Science and Technology Directorate
Washington, DC

Homeland Security

Science and Technology

AUTHORS

This report was authored by Cheryl Hendrickson (Ph.D.), Tara L. Myers (Ph.D.), Andrew C. Loignon (M.S.), Sarah N. Gilbert (M.S.), James Kurtessis (M.A.), Timothy P. Clayton (M.S.), Dwayne G. Norris (Ph.D.), and Scott A. Davies (Ph.D.) of the American Institutes for Research.

TABLE OF CONTENTS

Page

Sensitive Security Information (SSI)

LIST OF TABLES

LIST OF FIGURES

OVERVIEW AND ORGANIZATION OF THE REPORT

This report documents research conducted by the American Institutes for Research (AIR) to develop and validate an experimental selection battery for use in screening Behavior Detection Officer (BDO) job candidates. Throughout the process, AIR followed professional guidelines for measurement development as presented in the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999) and the Principles for Validation and Use of Personnel Selection Procedures (Society for Industrial and Organizational Psychology [SIOP], 2003). In this vein, the Validation Study Report describes the steps of the test development and validation process. It is composed of two volumes described below: the Technical Report (Volume I, the current document) and Appendices (Volume II).

Volume I contains a description of the work plan used to conduct the work, each phase of the research, and key findings and recommendations. This information is presented in chapters, which are summarized below:

Chapter I: This chapter provides background information about the BDO job, the purpose and goals of the validation study, and a project overview.

Chapter II: This chapter describes the purposes and development of procedures of early phases of the project, including the BDO job analysis (AIR, 2010a) and the test specifications (AIR, 2010b).

Chapter III: Chapter III describes the procedures AIR followed to procure and develop the selection measures assessed during the pilot test and validation study.

Chapter IV: The next chapter describes the purpose, method, and results of the pilot test conducted to collect preliminary data on the feasibility and psychometric properties of the experimental selection battery.

Chapter V: Chapter V describes the purpose, method, and results of the validation study conducted to examine each component of the experimental selection battery.

Chapter VI: The next chapter describes the changes made to each of the assessments based on analysis of both the quantitative and qualitative data collected during the validation study.

Chapter VII: Chapter VII provides a summary of the BDO validation study, notes important products resulting from this work, and revisits some key findings of the research. In addition, potential next steps related to the implementation of the proposed selection battery are provided.

Chapter VIII: This chapter provides an overview of the approach AIR used to set preliminary qualifying scores for the recommended selection battery.

Chapter IX: The technical report concludes with future considerations related to the enhancement of BDO selection, training, and performance management.

Sensitive Security Information (SSI)

Volume II contains supporting appendices, such as the Glossary of Key Terms and Acronyms (Appendix A), results from the job analysis (Appendices B-E), training materials (Appendices K, M, R, and S), the tools and protocols used in data collection (e.g., Appendices I, L, P, T), and detailed data tables from the pilot test and validation study (Appendices N, O, V, W, and X). To ensure test security, selection measure materials (e.g., administration guides, selection measures) are not included in either Volume I or Volume II.

EXECUTIVE SUMMARY

The Transportation Security Administration (TSA) employs a risk-based strategy that leverages intelligence-driven screening procedures to improve both the effectiveness and efficiency of the security process. A major component of TSA's risk-based strategy is its Screening of Passengers by Observation Techniques (SPOT) program, which was developed to identify high-risk passengers and augment airport security screening. Specifically, specially trained TSA airport screeners called Behavior Detection Officers (BDOs) identify passengers whose behaviors indicate they may be a threat to aviation and/or transportation security.

The work performed by BDOs can have significant consequences. Failure to identify someone who intends harm could result in catastrophic loss of life or property. Therefore, individual BDO performance is a significant contributor to the success of the program and, ultimately, the safety of the traveling public.

Since its implementation in 2007,[1] the program has evidenced considerable success, as demonstrated by the number of passengers stopped who were, in fact, of concern to law enforcement (Costigan et al., 2011). As a result, the program has been rapidly expanding. The SPOT program currently has over 2,900 BDOs operating in 176 airports nationwide. Their success is noteworthy given the difficulty of identifying individuals who intend harm but have not yet committed any known or observable offense. Furthermore, the ever-present need to balance the nation's desire to protect itself while simultaneously respecting individual rights makes the SPOT program politically relevant and highly visible.

To build upon the SPOT program's early success and to ensure that the expanding workforce of BDOs will continue to be able to accomplish its mission, it is necessary to maintain a comprehensive human capital system, including recruitment and selection processes that ensure the systematic hiring of individuals who are best qualified for the job as well as initial and recurrent training that positions job incumbents for success on the job.

With these goals in mind, TSA, the Department of Homeland Security (DHS), the U.S. Naval Research Laboratory (NRL), and the American Institutes for Research (AIR) partnered to develop and validate an experimental selection battery for use in screening BDOs. This current report summarizes the research and is organized to reflect the key steps required by professional best practices (Society for Industrial and Organizational Psychology [SIOP], 2003) and established legal guidelines (Equal Employment Opportunity Commission [EEOC], Civil Service Commission, Department of Labor, & Department of Justice, 1978).

[1] The SPOT program was piloted in 2006, and fully implemented in 2007.

Purpose of the Validation Study

The primary objective of the validation study was to provide empirical evidence supporting the reliability and validity of the selection measures contained in the experimental selection battery. To accomplish this goal, AIR sought to answer five primary research questions:

- **Are the scores within the selection measures consistent?** (Do they evidence reliability?)
- **Do the selection measures assess important aspects of the job?** (Do they evidence content validity?)
- **Do the selection measures assess what they were designed to measure?** (Do they evidence construct validity?)
- **How well do the selection measures predict job-relevant outcomes?** (Do they evidence criterion-related validity?)
- **How easy is it to implement and/or score the selection measures?** (What are the practical implications of using them?)

Reliability of a measure can refer to consistency across items (i.e., internal consistency), raters (i.e., interrater reliability), or time (i.e., test-retest reliability) (Nunnally & Bernstein, 1994). At least some form of reliability is required for selection measures before one can make informed hiring decisions because the hiring decision may otherwise be unduly influenced by measurement error. In the current study, AIR evaluated the reliability of the measures by either examining the consistency of test items or the ratings provided by assessors.

In addition, the *Uniform Guidelines on Employee Selection Procedures* (EEOC, 1978) requires all tests used in hiring decisions to demonstrate appropriate levels of validity. Ultimately, the validity of a selection measure is based on evidence that one can appropriately draw inferences about job candidates' potential job performance from their test scores (Guion, 1998; Putka & Sackett, 2010). To ensure that the experimental selection battery for the BDO job met all best practice and legal requirements, AIR sought to establish the validity of the assessments by documenting evidence of content, construct, and criterion-related validity. Content validity (i.e., the extent to which the measure resembles important aspects of the job) was established by basing the experimental selection battery on a recent, comprehensive job analysis (AIR, 2010a). Construct validity (i.e., whether the measure assesses what it was designed to measure) was assessed by examining the degree to which each selection measure converged and/or diverged from other measures in expected ways according to the underlying abilities or existing literature

(McPhail & Stelly, 2010; Landon & Arvey, 2007). Criterion-related validity (i.e., the extent to which scores on the selection measure predict job-relevant outcomes) was examined by calculating the statistical relationship between the assessment scores and ratings of job performance.

To answer the final research question, AIR considered the resources necessary for implementing the updated BDO selection system. This analysis included, for example, consideration of the amount of training required for Administrators to be able to proctor a selection measure, the number of job candidates that could complete a measure during a single administration, and the additional resources that would be required for each measure. These considerations allowed AIR to identify a system that was not only legally defensible, but could be implemented on a nationwide basis.

Design of the Validation Study

The goals of a validation study, as outlined above, can be accomplished in one of two ways: with a predictive design or with a concurrent design (Guion, 1998; McPhail & Stelly, 2010). In both designs, researchers administer the experimental selection battery and collect criteria data (e.g., job performance ratings). Then, the researchers examine the psychometric properties of the selection measures and assess whether the tests scores significantly predict the outcome of interest.

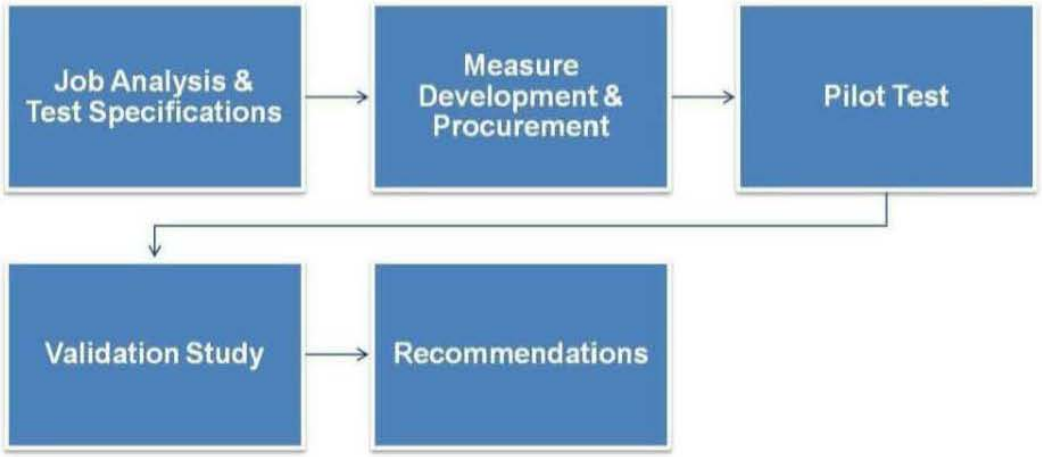
With a predictive design, assessment scores come from job candidates. They complete the selection measures as part of an experimental selection battery, the organization makes hiring decisions on the basis of some other measure (e.g., the existing selection battery), and criteria such as performance ratings are obtained after the new hires have been on the job for a specified period of time (McPhail & Stelly, 2010). With a concurrent design, assessment scores come from job incumbents. They can complete the selection measures at the same time managers are providing job performance ratings for each participant.

Each design option has advantages and disadvantages. For example, the results of a predictive validation study are more likely to directly represent the operational validity of a system (i.e., directly estimate the performance of job candidates) than would the results of a concurrent validation study. On the other hand, a predictive design often requires an extended timeframe for data collection and exposure of the experimental selection battery to participants from the existing applicant pool. After careful consideration, AIR, along with NRL, DHS, and TSA, chose to conduct a concurrent validation study. This approach allowed AIR to collect the

necessary data in a more timely fashion, preserve the applicant pool, and include data from TSA's existing performance appraisal system— the Performance Accountability and Standards System (PASS). Furthermore, because BDOs are familiar with the requisite tasks and duties for their position, they were able to provide direct feedback on the job-relatedness of the measures.

The validation study required a multi-step research plan comprised of several phases (as shown in Figure 1). Each research step is described below as a high-level overview and in the technical report in detail.

Figure 1. Phases of Research



Job Analysis

Job analysis is the process of defining the work activities (i.e., job duties and tasks) and worker characteristics (i.e., knowledge, skills, abilities, and other characteristics [KSAOs]) required for successful performance in a particular job (Brannick & Levine, 2002). Given the expansion of the SPOT program in recent years, the BDO job itself has changed and expanded to include a variety of new tasks such as —walking the line,‖ observing individuals outside of the security checkpoint, working in pairs with other BDOs, and using applied behavior detection methods. Thus, AIR collected job analytic data to capture updates to the work and to identify the KSAOs essential for successful BDO performance. Conducting a comprehensive job analysis prior to updating and/or developing a selection system represents commonly accepted best practices

(SIOP, 2003) and meets legal requirements according to the *Uniform Guidelines on Employee Selection* (EEOC, 1978).

Although the BDO job analytic information was relevant for both selection and training purposes, the current report focuses on selection. In essence, AIR identified the most important pre-hire KSAOs to ensure that the BDO experimental selection battery was driven by a current and accurate understanding of the job. A thorough review of this phase of the research is provided in the job analysis technical report (AIR, 2010a).

Method and Results

To begin, AIR reviewed TSA documents, including the SPOT Standard Operating Procedures, training materials, and position descriptions. This review informed the development of the draft lists of BDO tasks, KSAOs, and tools and equipment.

Next, AIR visited 11 airports across the U.S. to conduct: (1) observations of BDO job performance; (2) one-on-one semi-structured interviews with BDOs, SPOT Transportation Security Managers (STSMs), trainers, Facility Security Directors (FSDs), Assistant Facility Security Directors (AFSDS), and Human Resource (HR) representatives about the BDO job; and (3) focus groups with BDOs to review the initial task, KSAO, and tool and equipment lists. Immediately following the site visits, AIR updated the draft lists and developed a job analysis survey that was distributed to a sample of BDOs and STSMs. These subject matter experts were asked to rate the job tasks in terms of importance, frequency, and difficulty to learn and the KSAOs in terms of importance and time when required. From this information, AIR identified the most critical tasks and most critical pre-hire KSAOs. AIR used the finalized list of critical tasks and KSAOs in a KSAO-Task linkage workshop during which BDOs and STSMs rated the extent to which each KSAO was required for each task. This allowed AIR to document the job relevance of the pre-hire KSAOs.

The job analysis concluded with the development of a KSAO-competency crosswalk.[2] AIR met with representatives of TSA's Office of Human Capital (OHC) to link the critical KSAOs (identified through the job analysis) with relevant competencies contained in TSA's existing competency catalog. In addition, the workshop participants made suggested edits to some of the competencies.[3] The KSAO-competency links provided TSA with additional information about

[2] Competencies are sets of worker attributes that are broader versions of KSAOs. For a more complete discussion of the similarities and differences in these methods, see the Job Analysis Update: Training Decision Paper (AIR, 2009). [3] AIR recommended editing six existing competencies and adding five new competencies. OHC representatives presented these recommendations to their Competency Working Assessment Group (CWAG) for approval. [3] Currently, TSA hires from an internal applicant pool of Transportation Security Officers (TSOs) using a Quick Hire Application, Structured Interview, and medical screening.

their human capital initiatives, and the competency edits offered the potential for closer alignment with the BDO job.

Recommendations

The job analysis provided an updated view (and documentation) of the BDO tasks, KSAOs, and tools and equipment required for the job. Although some core aspects of the job remained unchanged, the expanded nature of the work led AIR to recommend that TSA revise their selection system[3] by refreshing the current content of the Structured Interview and adding new selection measures to the hiring process. AIR began this process for TSA with the test specifications step described below.

Test Specifications

Test specifications provide a framework for the content and structure of selection measures. The ultimate goal is to align the pre-hire KSAOs with the experimental selection measures, demonstrating optimal coverage of the critical KSAOs (Russell & Peterson, 1997). For example, Table 1 illustrates a notional relationship between five pre-hire KSAOs and four potential selection measures. Ability #1 is assessed by the Problem Solving measure. Ability #2 is assessed by both the Work Sample Test and the Structured Interview. Such overlap is typical, ensuring that the KSAOs are adequately measured (Russell & Peterson, 1997).

Table 1. Example of KSAO – Assessment Crosswalk

KSAO	Problem Solving Measure	Work Sample Test	Structured Interview	Role-Play Exercise
Ability #1	X			
Ability #2		X	X	
Skill #3		X	X	
Skill #4		X		X
Other Characteristic #5			X	X

Note. These relationships between KSAOs and selection measures are for demonstrative purposes only.

In keeping with these best practices, AIR used the results of the job analysis to develop test specifications for an experimental selection battery. A thorough review of this phase of the research is provided in the test specifications technical report (AIR, 2010b).

Method

AIR began by prioritizing the list of pre-hire KSAOs for inclusion in the experimental selection battery to identify KSAOs that were (1) required at the time of hire, (2) important for successful job performance, and (3) related to the majority of the job. Through a review of the job analysis data, AIR first identified KSAOs that were rated as pre-hire by a majority of the raters, had average importance ratings of 4 or higher (on a 5-point scale), and linked to at least 50% of the critical job tasks. Next, AIR refined the list through a rational review. For example, AIR determined which KSAOs were suitable for use in selection versus training and re-categorized them as appropriate. Then, AIR reviewed the existing research literature to identify measures that would appropriately assess the KSAOs. Potential measures were reviewed using three categories of criteria: (1) descriptive (e.g., relevance to the intended construct), (2) psychometric (e.g., reliability, validity), and (3) operational constraints (e.g., time to develop) (Hendrickson, Matheson, Amodeo, Norris, & Sparano, 2008; Russell, Norris, & Goodwin, 2000).

Results and Recommendations

The test specifications (AIR, 2012b) contained the list of recommended selection measures for the experimental selection battery and outlined the relationship between the prioritized list of KSAOs and the recommended measures. Although TSA currently uses an internal applicant pool when recruiting and selecting BDO job candidates, AIR ensured that the experimental selection battery would also be appropriate for use with an external applicant pool.

AIR recommended continued use of the medical screening process and Structured Interview. Further examination of the medical evaluation was beyond the scope of this research effort, but AIR recommended refreshing the Structured Interview items for the competencies currently assessed and developing new items for an additional competency: Honesty-Integrity. AIR also recommended adding three new measures that had the potential for enhancing the BDO hiring process, including a:

- critical thinking measure, such as the Watson-Glaser Critical Thinking Appraisal – Short Form;
- Work Sample Test, which would require job candidates to perform tasks similar to those they will complete on the job, designed to assess the specific cognitive abilities and underlying skills necessary for BDO screening duties (e.g., speed of closure, sustained attention, selective attention, memory, mathematical operations); and
- Role-Play Exercise to measure job candidates' skill in gathering information while engaging passengers in conversation.

These measures are described in more detail below.

Measurement Development and Procurement

AIR developed and/or procured eight predictor (i.e., selection) measures for the pilot test and validation study (as summarized in Table 2). The Watson-Glaser Critical Thinking Appraisal – Short Form and Wonderlic Personnel Test (included as an additional measure of construct validity) were procured from their publishers. For the Work Sample Test, Structured Interview, and Role-Play Exercise, AIR examined the literature for best practices, created preliminary measures, conducted internal pilot tests with AIR research staff, and revised the measures. Then, SPOT STSMs and representatives from TSA's Office of Security Operations (OSO) reviewed the measures and provided feedback about their difficulty, the clarity of the instructions, and perceived relevance to the BDO job. This feedback informed final revisions completed in preparation for the formal pilot test conducted with job incumbents. **Table 2. Summary of Predictor Measures**

Measure	Description	Constructs Assessed
Wonderlic Personnel Test	Job candidates complete a series of verbal and quantitative reasoning questions on a timed (12 minutes), 50-item, paperandpencil test.	□ Problem Solving

Watson-Glaser Critical Thinking Appraisal – Short Form	Job candidates form inferences, recognize assumptions, deduce and interpret statements, and evaluate arguments on a timed (40 minutes), 40-item, paperandpencil test.	☐ Critical Thinking ☐ DecisionMaking
Passenger Observation Assessment	Job candidates observe passengers waiting in a queue on a five-minute video recording and look for specific behaviors or appearance factors.	☐ Multitasking ☐ Attention to Detail ☐ Speed of Closure ☐ Sustained Attention ☐ Selective Attention ☐ Decision-Making ☐ Critical Thinking ☐
Mental Math Assessment	Job candidates sum values that appear every second without taking notes or using scratch paper.	☐ Attention to Detail ☐ Math
Writing Knowledge Assessment[3]	Job candidates review a brief passage and circle any errors.	☐ Attention to Detail ☐ Written Communication
Visual Recall Assessment	Job candidates review two images of an airport setting for two minutes each and then answer multiple-choice questions about their content.	☐ Attention to Detail ☐ Speed of Closure ☐ Sustained Attention ☐ Selective Attention ☐ Memory ☐

[3] Note that associated documents may refer to this assessment as —Written Summary.‖ AIR changed the name of this measure following the pilot test to better reflect the abilities assessed by this measure.

Structured Interview	Job candidates answer eight questions designed to measure eight competencies.	▯ Attention to Detail ▯ Decisiveness ▯ Flexibility ▯ Honesty-Integrity ▯ Interpersonal Skills ▯ Multitasking ▯ Self-Management ▯ Teamwork ▯ Oral Communication
Role-Play Exercise	Job candidates engage in a conversation with a fictitious passenger to elicit three pieces of information about the person's trip.	▯ Oral Communication ▯ Social Skills ▯ Decision-Making ▯ Memory

In addition to the selection measures, AIR created a feedback questionnaire that captured BDOs' perceptions of the selection system to ensure that the newly developed measures will be well received by future job candidates. This was important because job candidates' reactions to selection measures can influence whether they view the selection process as fair and transparent, complete the selection process, accept job offers, and refer other job candidates to the organization (Gilliland, 1993; Hausknecht, Day & Thomas, 2004; Macan, Avedon, Paese, & Smith, 1994). Such impressions and decisions are especially pertinent for the SPOT program because the TSOs may return to their former positions if they are not selected as BDOs. AIR developed the feedback questionnaire by adapting items from well established job candidate feedback assessments (Arvey, Strickland, Drauden, and Martin, 1990; Bauer et al., 2001; Schmitt, Oswald, Kim, Gillespie, & Ramsay, 2004).

To evaluate the experimental selection battery using the concurrent validity design, AIR collected job performance ratings for each participating BDO. One source of ratings was the existing TSA PASS, which focuses on technical proficiency, BDO competencies, and other jobrelevant criteria (i.e., Training and Development Evaluation, Readiness for Duty Evaluation). In addition, AIR developed a performance measure (i.e., the BDO Job Performance Measure [BDO JPM]) for use in the validation study. This measure contained nine performance dimensions that were identified during the job analysis phase of the research:

1. Observation and Assessment of Passengers

2. Calculation and Assignment of Points
3. Referral of Passengers
4. Interaction with Passengers
5. Cooperation and Communication with Teammates
6. Interaction with Other Security Personnel
7. Documentation
8. Search of Accessible Property and Review of Travel Documentation
9. Preparedness and Dutifulness

These dimensions, and associated behavioral anchors,[4] were created through an iterative process that included the facilitation of external reviews with SPOT subject matter experts (e.g., BDOs).

Pilot Test

Examination of the experimental selection battery began with a pilot test, which took place in five airports across the U.S. over the course of six weeks beginning in February 2010 (see Table 3). AIR administered the assessments to incumbent BDOs, solicited their feedback about the assessments, examined the test scores, and obtained performance ratings for the participating BDOs. Following best practices for test administration (American Educational Research Association [AERA], American Psychological Association, and National Council on Measurement in Education, 1999), AIR standardized the test administration processes at each site visit by using administration guides for each assessment and by following a semi-structured pilot test protocol.

Table 3. Pilot Test Sites

Airport Code – Location	Date of Visit
BWI – Baltimore, MD	March 23 – 24, 2011
BOS – Boston, MA	March 29 – April 1, 2011
MCO – Orlando, FL	March 4 – 6, 2011
PVD – Providence, RI	February 22 – 24, 2011
SLC – Salt Lake City, UT	March 14 – 17, 2011

[4] Behavioral anchors are specific examples of behavior that reflect a particular level of job performance. Unique behavioral anchors were developed for each performance dimension. These behavioral anchors were developed to assist raters when making performance ratings – that is, these behavioral anchors were intended to reduce rating errors and biases and, more generally, to improve the validity and reliability of performance ratings.

Through the pilot test, AIR examined the proposed data collection process and the degree to which the measures were functioning as planned (AERA, 1999). For example, AIR explored the efficiency of the data collection process, the psychometric quality of the measures, the clarity of the instructions, and the perceived utility of the measures. These data were used to inform revisions to the measures prior to the validation study.

Following the pilot test, TSA's Office of Human Capital recommended removing the Structured
Interview from the validation study. Nevertheless, AIR modified the updated Structured Interview on the basis of the pilot test results and feedback from BDOs. [5]

Validation Study

Next, AIR conducted a concurrent, criterion-related validation study. During the study, AIR administered the assessments to incumbent BDOs, solicited their feedback, and collected job performance ratings. In addition to examining the test scores, AIR assessed the relationship of those scores with the job performance ratings.

[5] To ensure test security, selection measure materials (e.g., administration guides, selection measures) are not included in this or other AIR technical reports.

At the conclusion of the validation study, AIR implemented final changes to the selection and job performance measures. Furthermore, the research culminated with the development of recommendations about use of the experimental selection battery, potential steps for setting cut scores, and additional, related research efforts that DHS/TSA may want to consider pursuing to further accomplish their goal of maintaining a successful BDO workforce.

Validation Study Participants

AIR collected data at 11 sites (shown in Table 4) from early September 2011 to late November 2011. The final sample for the study consisted of 214 BDOs.

Table 4. Validation Study Sites

Primary Airport (Code)	Date of Visit
IAD – Washington, DC	September 13, 2012 – September 14, 2012
MDW – Chicago, IL	September 21, 2012 – September 23, 2012
RDU – Raleigh, NC	October 4, 2012 – October 6, 2012
PHX – Phoenix, AZ	September 28, 2012 – September 29, 2012
SJC – San Jose, CA	October 12, 2012 – October 13, 2012
LAX – Los Angeles, CA	October 18, 2012 – October 20, 2012
DEN – Denver, CO	October 25, 2012 – October 27, 2012
PHL – Philadelphia, PA	November 1, 2012 – November 3, 2012
LAS – Las Vegas, NV	November 9, 2012 – November 11, 2012
JFK – New York, NY	November 14, 2012 – November 17, 2012
DCA – Washington, DC	November 29, 2012 – November 30, 2012

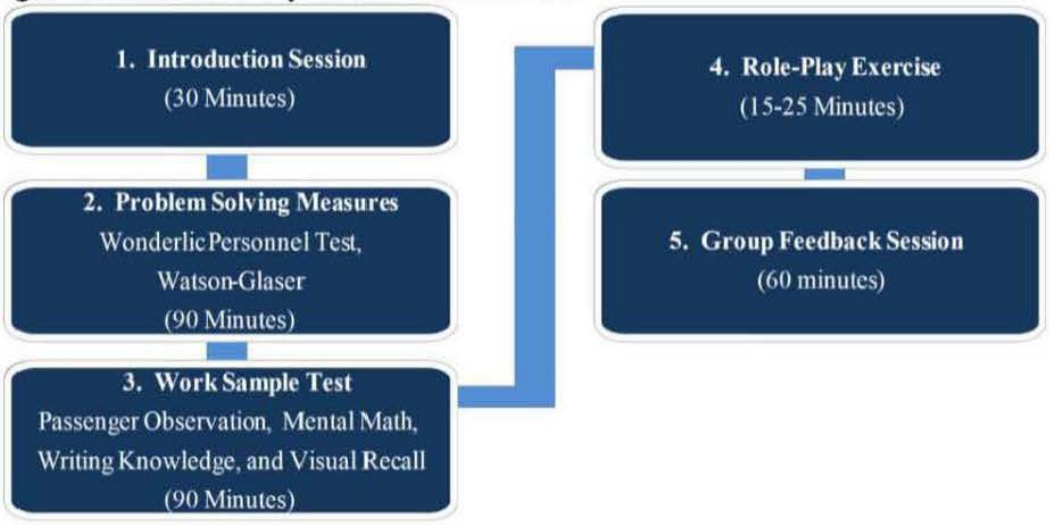
Note. Candidates from Wilmington, NC (ILM) participated during the RDU site visit.

Ideally, the sample of BDOs would be similar to an applicant population to better simulate a typical hiring scenario. Given the voluntary nature of the study and the recent hiring freeze for the SPOT program, however, this was difficult to achieve. For instance, nearly a quarter of the participants were G-Band BDOs (23%) and the average tenure for the participants was 2.94 years (SD = 1.41). These figures suggest that the majority of BDOs who participated in the validation study were experienced job incumbents and were likely to perform differently on the selection measures than job candidates. AIR considered these differences while conducting the validation study analyses and developing recommendations.

Validation Study Method

As with the pilot test, AIR standardized the test administration processes at each site visit by using administration guides for each assessment and by following a semi-structured validation study protocol. These protocols and administration guides were similar to those used during the pilot test. Figure 2 outlines the order in which assessments were delivered as well as the amount of time generally required for each.

Figure 2. Validation Study Administration Process



As opposed to this on-site data collection, the performance data were collected via email. AIR received PASS data from TSA for each BDO who participated in the validation study for the previous performance period (i.e., 2011-2012). In addition, STSMs, trained in using the BDO JPM, completed and sent BDO JPM ratings for each participating BDO.

Key Findings

The results of the validation study indicated that some of the experimental selection measures performed better than others. More specifically, the current study provided evidence that supports the use of the Passenger Observation Assessment, the Mental Math Assessment, and the Role-Play Exercise. The results of the study also suggested that the Watson-Glaser Critical Thinking Appraisal – Short Form, the Writing Knowledge Assessment, and the Visual Recall Assessment were less promising for selection purposes.

Recommended Assessments

Findings from the validation study provide more support for the Passenger Observation Assessment, the Mental Math Assessment, and the Role-Play Exercise, which demonstrated evidence of face, content, construct, and criterion-related validity. In addition, the minimal level of overlap among these assessments suggests that each can contribute unique information when predicting BDO job performance. Though differences existed across forms for each of the selection measures, at least one form of each selection measure was strongly correlated with ratings from either the BDO JPM or PASS.

Assessments Not Included in the Recommended Selection Battery

Results from the validation study suggested that three of the measures in the experimental selection battery were less promising: the Watson-Glaser Critical Thinking Appraisal – Short Form, the Writing Knowledge Assessment, and the Visual Recall Assessment. For example, these measures demonstrated less content and face validity than did the chosen measures discussed above. They are also more likely to be susceptible to adverse impact than the other assessments in the experimental selection battery and thus could increase the likelihood of a legal challenge to the BDO selection system (Hough, Oswald, & Ployhart, 2001). The Visual Recall Assessment evidenced additional limitations (e.g., problem with clarity of the images, ineffective distracters, low item-total correlations, failure to diverge from the other measures).

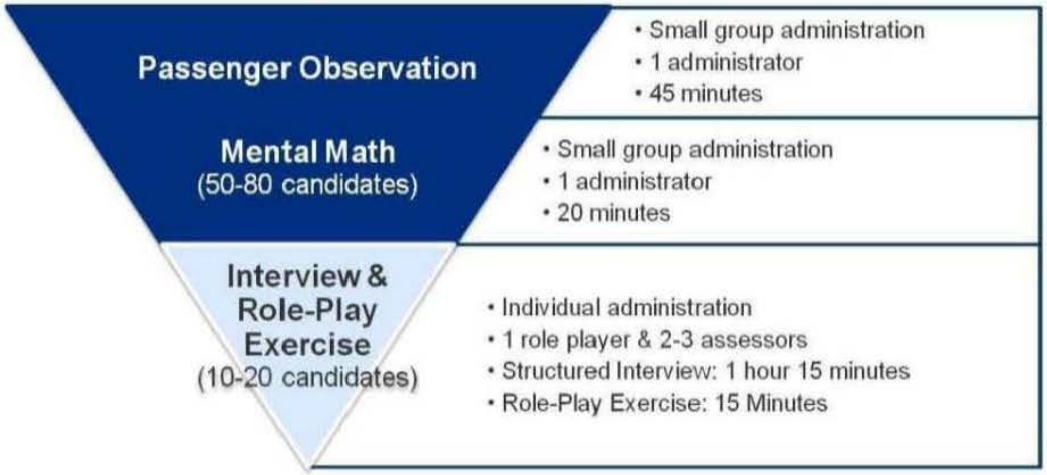
Proposed Selection System

After identifying the most promising selection measures, AIR considered different approaches for combining the measures into a single selection system. One approach consists of using a compensatory model. In this model, the selection measures are weighted and combined into an overall composite. Selection decisions are based on a candidate's overall composite score. Thus, higher scores on one selection measure can compensate for lower scores on another measure. For compensatory measures, weighting schemes should be based on rational criteria, such as choosing to weight selection measures in order to emphasize certain abilities or skills that are critical to job performance (Guion, 1998).

An alternative approach involves a non-compensatory model. In these types of selection systems, job candidates must reach a certain score on each selection measure to be considered for employment. Thus, each selection measure within the battery is a —hurdlel that candidates must pass in order to be selected for the job. These types of models are appropriate in situations where each skill and ability measured within the system is uniquely critical for performance in a way that other strengths cannot compensate for weaknesses in them (Guion, 1998).

AIR recommends using a non-compensatory model to implement the recommended selection battery. In the first hurdle, AIR recommends administering the Passenger Observation and Mental Math Assessments to job candidates in small groups (i.e., 3-4 job candidates). Next, job candidates who score above the qualifying scores for these two measures would then complete the revised Structured Interview and the newly developed Role-Play Exercise via a one-on-one administration. This recommended system is depicted in Figure 3. The figure includes estimates of the number of job candidates that would complete each hurdle in the selection system using an example for illustrative purposes. This example shows how the number of job candidates could be reduced at each step.[6] The typical administration time and resource requirements for both sets of measures are depicted on the right-hand side of the figure.

Figure 3. Increased Efficiency of the Recommended Selection Battery



This system is recommended for two reasons. First, each of the underlying skills and abilities that are assessed by the recommended selection battery are critical for BDO job performance. As such, demonstrating skill in one area cannot compensate for lack of skill in another. For example, if a candidate demonstrates the ability to perform mental arithmetic quickly and accurately, but is unable to observe and distinguish behaviors or appearance factors, he or she would be unlikely to perform the BDO job successfully. Likewise, if a job candidate

[6] This example assumes that 50-80 job candidates are initially considered for a few job openings during one selection cycle. During the conduct of the validation study, some sites reported that they were interviewing 50-80 job candidates for 2-3 job openings.

demonstrates skill in observing behaviors and appearance factors but is unable to elicit information during a conversation with a passenger, he or she would not likely be a good fit for the BDO job.

Second, by reevaluating job candidates after the first hurdle, TSA would be able to reduce the costs associated with administering the Structured Interview. While collecting data during the validation study, some sites commented that they were interviewing up to 80 candidates for 2 – 5 job openings. Because each Structured Interview requires at least two Assessors and can take 45 to 75 minutes to administer, this can be a time-consuming and resource-intensive process. By including a preliminary hurdle that consists of measures that can be administered in small groups, TSA can reduce the number of candidates to be interviewed to fill a BDO job opening. Furthermore, because the Role-Play Exercise was designed to be administered immediately after the Structured Interview, the only additional resource requirement for that phase of the system involves procuring and training a role-player.

The system would also provide increased efficiency by providing job candidates with a realistic job preview (Premack & Wanous, 1985). Specifically, by placing the Passenger Observation and Mental Math Assessments early in the administration process, job candidates will have more information to determine whether they are qualified for or interested in the BDO. Some job candidates who are not suited for the BDO job may withdraw from the selection process at this stage and thereby reduce the number of candidates that need to be interviewed, thus reducing the cost and resources required to fill the job opening.

Summary and Recommended Next Steps

This section reviews the work conducted for the current project, the important outcomes and products, and recommendations for next steps.

Review of Completed Work

The research consisted of multiple steps, each of which provided important information about the BDO job and the experimental selection measures. These steps included the following:

- Conducted and documented a job analysis of the BDO job (AIR, 2010a);

- Developed test specifications to determine the appropriate selection measures for assessing important pre-hire knowledge, skills, abilities, and other characteristics (KSAOs) (AIR, 2010b);
- Developed a Work Sample Test and a Role-Play Exercise to include in the experimental selection battery;
- Identified and procured two Problem Solving measures to include in the experimental selection battery;
- Updated Structured Interview items for inclusion in the experimental selection battery;
- Developed the BDO Job Performance Measure (JPM) and collected existing Performance Accountability and Standards System (PASS) ratings for the criteria in the validation study;
- Pilot tested the selection and criteria measures using AIR staff and at five airports prior to the validation study and analyzed these data to make key decisions regarding the validation study (e.g., procedures to follow, selection measure revisions, forms to administer);
- Conducted validation study at 11 sites and analyzed these data to determine the validity of each selection measure as a predictor of BDO job performance;
- Finalized the selection measures based on results of the validation study; and □

Developed scoring and standard setting recommendations. **Important Products**

The research resulted in the development of several important products. These include:

- A comprehensive job analysis (AIR, 2010a), including lists of current job tasks, KSAOs and competencies, and tools and equipment, as well as information about the work environment. In addition to guiding the current research, this information can be used to inform training, career development, and other human capital initiatives and programs;
- Several experimental selection measures, including the Passenger Observation Assessment, the Mental Math, the Writing Knowledge, and the Visual Recall Assessments, a Role-Play Exercise, and updated Structured Interview questions;
- Critical support documents and other resources for ensuring the proper and standardized administration and use of each of the selection measures, including administration guides, assessor guides, response booklets, and multimedia (e.g., images, videos);
- Training materials for the Role-Play Exercise, including protocols and materials for the Role-Play Exercise Assessors and Resource Person trainings;

- A behaviorally-based performance measure designed specifically for the BDO job and based on the current job analysis (AIR, 2010a);
- Training materials for the BDO Job Performance Measure;
- A scoring technical report (AIR, 2012a), including scoring and standard setting recommendations; and
- This validation study technical report (AIR, 2012b), which documents the procedures followed to validate the recommended selection measures, as well as key findings.

Recommended Next Steps

As described above, the results from the validation study suggests that three of the selection measures are likely to be useful components of the BDO hiring system. Although the validation process is the final phase of the current study, AIR is recommending several additional steps:

- Gathering additional data regarding the recommended selection battery from job candidates using a predictive validation design;
- Considering best practices when implementing the final system (e.g., orient STSMs to the new selection measures, standardize data management and record keeping processes, communicate validation results to all participants);
- Implementing the updated Structured Interview;
- Using the BDO JPM as part of ongoing or future research efforts; and Considering recommendations regarding setting qualifying scores.

Future Considerations

During discussions with the BDOs and STSMs throughout the validation study, these SMEs also provided recommendations for training, performance management, and career progression. These recommendations are documented for TSA's consideration and include:

- Converting the Work Sample Test measures to an electronic and un-proctored administration format;
- Examining the utility of a personality measure (to be used for selection purposes);
- Examining the advantages and disadvantages of using an external applicant pool for the SPOT program;
- Evaluating the validity of the recommended selection battery for selecting candidates into the SPOT Assessor Program;

- Using job analysis data to conduct a training alignment study;
- Developing recurring training/performance measures modeled after the experimental selection battery;
- Examining the BDO career path;
- Examining the human capital processes used for contractor BDOs;
- Investigating the reciprocal relationship between a BDO's behavior and a passenger's behavior during the screening process; and
- Examining the affect of fatigue on a BDO's vigilance.

CHAPTER I: INTRODUCTION

This chapter begins by briefly reviewing the background for the current project. Next are descriptions of the purpose and design of the validation study. Finally, the chapter concludes with an overview of the research process.

Background

The Transportation Security Administration (TSA) employs a risk-based strategy that leverages intelligence-driven screening procedures to improve both the effectiveness and efficiency of the security process. A major component of TSA's risk-based strategy is its Screening of Passengers by Observation Techniques (SPOT) program, which was developed to identify highrisk passengers and augment airport security screening. Specifically, specially trained TSA airport screeners called Behavior Detection Officers (BDOs) identify passengers whose behaviors indicate they may be a threat to aviation and/or transportation security.

The work performed by BDOs can have significant consequences. Failure to identify someone who intends harm could result in catastrophic loss of life or property. Therefore, individual BDO performance is a significant contributor to the success of the program and ultimately the safety of the traveling public.

Since its implementation in 2007,[9] the program has evidenced considerable success, as demonstrated by the number of passengers stopped who were, in fact, of concern to law enforcement (Costigan, Makonnen, Taylor, Sawyer, Myers, & Toplitz, 2011). As a result, the program has been rapidly expanding. The SPOT program currently has over 2,900 BDOs operating in 176 airports nationwide. Their success is noteworthy given the difficulty of identifying individuals who intend harm but have not yet committed any known or observable offense. Furthermore, the ever-present need to balance the nation's desire to protect itself while simultaneously respecting individual rights makes the SPOT program politically relevant and highly visible.

To build upon the SPOT program's early success and to ensure that the expanding workforce of BDOs will continue to be able to accomplish its mission, it is necessary to maintain a comprehensive human capital system. This system includes recruitment and selection processes that ensure the systematic hiring of individuals who are best qualified for the job as well as initial and recurrent training that positions job incumbents for success on the job. Effective selection and training programs, by providing a realistic job preview prior to entry into the BDO

[9] The SPOT program was piloted in 2006, and fully implemented in 2007.

Sensitive Security Information (SSI)

workforce and strengthening the pool of qualified BDOs, could help TSA ensure successful BDO job performance and reduce the likelihood of turnover.[8]

With these goals in mind, TSA, the Department of Homeland Security (DHS), the U.S. Naval Research Laboratory (NRL), and the AIR partnered to develop and validate an experimental selection battery for use in screening BDOs. This current report summarizes the research and, as noted above, is organized to reflect the key steps required by professional best practices (Society for Industrial and Organizational Psychology [SIOP], 2003) and established legal guidelines (Equal Employment Opportunity Commission [EEOC], Civil Service Commission, Department of Labor, & Department of Justice, 1978).

Purpose of Validation Study

The primary objective of the validation study was to provide empirical evidence supporting the reliability and validity of the selection measures in the experimental selection battery. This included evidence demonstrating the consistency of the scores on the selection measures, the jobrelatedness of these measures, and the relationship between performance on the selection measures and performance on the job.

To accomplish this goal, AIR sought to answer five primary research questions, which are described in more detail throughout this section:

- **Are the scores within the selection measures consistent?** (Do they evidence reliability?)
- **Do the selection measures assess important aspects of the job?** (Do they evidence content validity?)
- **Do the selection measures assess what they were designed to measure?** (Do they evidence construct validity?)
- **How well do the selection measures predict job-relevant outcomes?** (Do they evidence criterion-related validity?)
- **How easy is it to implement and/or score the selection measures?** (What are the practical implications of using them?)

Reliability of a measure can refer to consistency across items (i.e., internal consistency), raters (i.e., interrater reliability), or time (i.e., test-retest reliability) (Nunnally & Bernstein, 1994). At least some form of reliability is required for selection measures before one can make informed

[8] Turnover is a particular concern for the SPOT program because of the extensive pre-hire training that BDOs receive and the possible security implications of having individuals who are knowledgeable about SPOT external to the program and/or TSA.

hiring decisions. If scores are not consistent, then hiring decisions may be unduly influenced by measurement error. In the current study, AIR evaluated the reliability of the measures by either examining the consistency of test items or the ratings provided by Assessors.

To further answer the questions above, AIR examined the validity of the measures. The *Uniform Guidelines on Employee Selection Procedures* (EEOC, 1978) requires all tests used in hiring decisions to be job-relevant and to demonstrate appropriate levels of validity. Ultimately, the validity of a selection measure is based on evidence that one can appropriately draw inferences about job candidates' potential job performance from their test scores (Guion, 1998; Putka & Sackett, 2010). To ensure that the experimental selection battery for the BDO job met all best practice and legal requirements, AIR sought to establish the validity of the assessments by documenting evidence of content, construct, and criterion-related validity. First, content validity (i.e., the extent to which the measure resembles important aspects of the job) was established by basing the experimental selection battery on a recent, comprehensive job analysis (AIR, 2010a). Construct validity (i.e., whether the measure assesses what it was designed to measure) was assessed by examining the degree to which each selection measure converged and/or diverged from other measures in expected ways according to the underlying abilities or existing literature (McPhail & Stelly, 2010; Landon & Arvey, 2007). Criterion-related validity (i.e., the extent to which scores on the selection measure predict job-relevant outcomes) was examined by calculating the statistical relationship between the assessment scores and ratings of job performance.

To answer the final research question, AIR considered the resources necessary for implementing the updated BDO selection system. This included, for example, the amount of training required for Administrators to be able to proctor a selection measure, the number of job candidates that could complete a measure during a single administration, and the additional resources that would be required for each selection measure. These considerations allowed AIR to identify a system that was not only legally defensible, but could be implemented on a nationwide basis.

Design of the Validation Study

The goals of a validation study, as outlined above, can be accomplished in one of two ways: with a predictive design or with a concurrent design (Guion, 1998; McPhail & Stelly, 2010). In both designs, researchers administer the experimental selection battery and collect criteria data (e.g., job performance, ratings, training scores). Then, researchers examine the psychometric properties of the selection measures and assess whether the tests scores significantly predict the outcome of interest.

With a predictive design, assessment scores come from job candidates. They complete the selection measures as part of an experimental selection battery and the organization makes hiring decisions on the basis of some other measure (e.g., the existing selection battery). Then, criteria such as performance ratings are obtained after the new hires have been on the job for a specified period of time (McPhail & Stelly, 2010). This approach can be beneficial because it provides a direct estimate of the performance of job candidates. Thus, the results of a predictive study are more likely to directly represent the operational validity of a system than would the results of a concurrent validation study. On the other hand, a predictive design often requires an extended timeframe for data collection, and the time-lapse between hire and collecting criterion data can be prohibitive. This limitation was especially relevant in the current study because TSA instated

a temporary hiring freeze for the SPOT program during the timeframe this research was being conducted. This hiring freeze severely limited the number of available Transportation Security Officer (TSO) candidates who could have participated in a predictive validation study. Furthermore, given that the job candidate pool consists solely of TSOs, DHS/TSA was reluctant to have them participate in the study because they would then have knowledge of the hiring system prior to applying for a BDO job.

With a concurrent design, assessment scores come from job incumbents. They can complete the selection measures at the same time managers are providing job performance ratings for each participant. Despite the more efficient timeline, there are some drawbacks to this design. First, it is possible that job incumbents will perform differently on some selection measures than do job candidates. This was of particular concern for the current study because BDO incumbents have gone through rigorous training and have practiced the skills and abilities that are assessed by the experimental selection battery on a daily basis. Thus, they may perform better on the selection measures than would job candidates.[9] These differences in performance require researchers to interpret concurrent validation data cautiously. Second, concurrent validation studies tend to evidence more range restriction than do predictive designs. Range restriction refers to a phenomenon in which participant scores do not vary as much as would be expected for job candidate scores. In concurrent designs, the job incumbents have already been screened on a job relevant measure, and then, presumably, very poor job performers have been dismissed. Essentially, the incumbent sample has been screened at least twice, leaving participants who will most likely score higher on the selection measures than would a full range of job candidates. Such restriction in the range of assessment scores can result in an attenuated criterion-related validity estimate (Guion, 1998; Van Iddekinge & Ployhart, 2008).

Upon weighing the pros and cons of the two validation designs, AIR, along with NRL, DHS, and TSA, chose to conduct a concurrent validation study. This approach allowed AIR to collect the

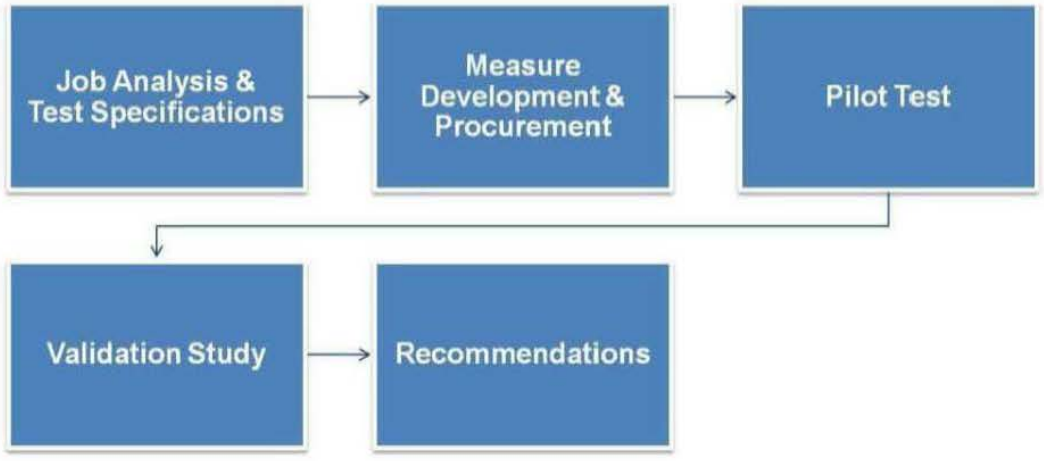
[9] On the other hand, on-the-job practice may limit incumbent BDOs' performance. Their job knowledge/acquired skills could interfere with their performance. The impact of both enhancement and interference due to job knowledge and acquired skills is discussed further in the subsequent chapters of the current report. [12] These corrections are presented in Chapter V of this report.

necessary data in a more timely fashion, administering the selection measures to BDOs while their managers completed performance ratings. AIR was also able to gather additional performance data from TSA's existing performance appraisal system— the Performance Accountability and Standards System (PASS). Furthermore, because BDOs are familiar with the requisite tasks and duties for their position, they were able to provide direct feedback on the jobrelatedness of the measures. Finally, AIR was able to minimize the range restriction limitation by using established statistical techniques to estimate the operational validity of the experimental selection battery.[12]

Overview of the Research Process

To assist DHS and TSA in developing an expanded selection system for BDOs, AIR developed a multi-step research plan comprised of several phases (see Figure 1). This section of the report provides a high-level overview to preview the subsequent chapters, which describe each phase in detail.

Figure 1. Phases of Research



For the job analysis, AIR collected information about the BDO job from job incumbents, managers, and trainers. Of particular importance for the validation study, the job analysis provided up-to-date information regarding which knowledge, skills, abilities, and other attributes (KSAOs) were most important for successful performance as a BDO. Starting with a thorough job analysis represents best practices (SIOP, 2003) and is also a legal requirement according to the Uniform Guidelines on Employee Selection (EEOC, 1978). Accordingly, AIR used the results of the job analysis during the measure development phase. Measure development began with the development of test specifications, in which AIR created plans for the content and structure of the selection system. This —blueprintl was based on the KSAOs appropriate for selection, identified through the job analysis, as well as on knowledge of research and best practice, legal requirements, and operational constraints.

Having weighed these various factors and specified the desired features of the selection system, AIR used the test specifications to guide the development of an experimental battery of tests. To the ensure identification of a set of tests that would have the potential to effectively predict BDO job performance, meet legal standards, and satisfy DHS/TSA requirements, AIR carefully examined the feasibility and psychometric properties of various assessments and ultimately included in the experimental selection battery more tests than would likely be needed. In addition, AIR developed a job performance measure against which tests scores could be compared to inform the usefulness of the experimental selection battery.

Sensitive Security Information (SSI)

Examination of the experimental selection battery began with a pilot test. The objective was to identify problematic test items, ensure the clarity of the test instructions, and preliminarily examine the usefulness of the assessments. To accomplish these goals, AIR administered the assessments to incumbent BDOs, solicited their feedback about the assessments, examined the tests scores, and obtained performance ratings for the participating BDOs. The results and experiences from the pilot test were used to inform revisions of the experimental selection battery and associated materials (e.g., administration guides, rater training materials).

Next, AIR conducted a concurrent, criterion-related validation study with the goal of providing empirical evidence of the reliability and validity of the selection measures. As with the pilot test, AIR administered the assessments to incumbent BDOs, solicited their feedback, and collected job performance ratings. In addition to examining the test scores, AIR assessed the relationship of those scores with the job performance ratings.

At the conclusion of the validation study, AIR implemented final changes to the selection and job performance measures. Furthermore, the research culminated with the development of recommendations about the use of the experimental selection battery, potential steps for setting cut scores, and additional, related research efforts that DHS/TSA may want to consider pursuing to further accomplish their goal of maintaining a successful BDO workforce.

CHAPTER II: PRIOR RESEARCH STEPS

Chapter II reviews the job analysis and test specifications phases of this research. Each step constitutes a prior, completed research phase and has been thoroughly documented in previous technical reports (AIR), 2010a and AIR, 2010b, respectively). Given that the job analysis and test specifications work provided the foundation for measurement development and validation, these steps are briefly described below. This information provides the critical evidence of the content validity of the experimental selection battery and provides the context for the subsequent chapters in this report.

Job Analysis

Job analysis is the process of defining the work activities (i.e., job duties and tasks) and worker characteristics (i.e., knowledge, skills, abilities, and other characteristics [KSAOs]) required for successful performance in a particular job (Brannick & Levine, 2002). Given the expansion of the SPOT program in recent years, the BDO job itself has changed and expanded to include a variety of new tasks such as —walking the line,I observing individuals outside of the security checkpoint, working in pairs with other BDOs, and using applied behavior detection methods. Thus, AIR collected job analytic data to capture updates to the work and to identify the KSAOs essential for successful BDO performance.

Such job analytic data are pertinent to the success of all Human Resource Management (HRM) activities (Siddique, 2004) and are particularly critical for proper development of personnel selection systems (Harvey, 1991; Thompson & Thompson, 1982). In fact, conducting a comprehensive job analysis prior to updating and/or developing a selection system represents commonly accepted best practices (SIOP, 2003) and meets legal requirements according to the *Uniform Guidelines on Employee Selection* (EEOC, 1978).

Although the BDO job analytic information was relevant for both selection and training purposes, the current report focuses on selection. In essence, AIR identified the most important KSAOs that are required at the time of hire to ensure that the BDO experimental selection battery was driven by a current and accurate understanding of the job. This section summarizes the methods, results, and recommendations of that job analysis.

Method and Results

The BDO job analysis consisted of five steps: (1) a review of background materials; (2) the conduct of site visits; (3) the development, implementation, and analysis of a nationwide job analysis survey; (4) the conduct of KSAO–task linkage workshops; and (5) the development of a KSAO-competency crosswalk. Table 1 provides an overview of the steps, along with their key outcomes. As evidenced in the table, subject matter expert (SME) input was critical at each step. SMEs included BDOs, SPOT Transportation Security Managers (STSMs), BDO trainers, Facility Security Directors (FSDs), Assistant Facility Security Directors (AFSDs), Human

 Resource (HR) representatives, and TSA's Office of Human Capital (OHC) representatives. Their roles varied depending on the step of the job analysis process.

Table 1. BDO Job Analysis Steps and Outcomes

Job Analysis Step	Work Completed	Outcome
Background Material Review	AIR reviewed TSA documents, including the SPOT Standard Operating Procedures (SOP), training materials, and position descriptions.	Developed initial task, KSAO, and tool/equipment lists
Site Visits	AIR visited 11 airports across the U.S. Job analysts: • Observed BDO job performance • Interviewed BDOs, STSMs, trainers, FSDs, AFSDs, and HR representatives about the BDO job • Facilitated focus groups with BDOs to review the tasks, KSAOs, tools and equipment, and TSA competencies	Updated task, KSAO, and tool/equipment lists (see Appendix B[10])

Job Analysis Survey	BDOs and STSMS rated: • Tasks in terms of importance, frequency, and difficulty to learn • KSAOs in terms of importance and time when required	Finalized list of critical tasks (see Appendix C) and critical pre-hire KSAOs (see Appendix D)
KSAO-Task Linkage Exercise	BDOs and STSMs rated the extent to which each critical KSAO was required for each critical job task	Documented job relevance of prehire KSAOs

[10] As previously stated, Volume II of this report contains all supporting appendices.

KSAO-Competency Crosswalk[7]	AIR met with TSA OHC to: • Link KSAOs to TSA competencies • Update the competencies to align with the BDO job[8]	• Provided additional information for TSA's human capital initiatives • Aligned competencies more closely to BDO job

[7] Competencies are sets of worker attributes that are broader versions of KSAOs. Linking the KSAOs and competencies combines the benefits of the job analysis and competency modeling methods. For a more complete discussion concerning the similarities and differences in these methods, see the *Job Analysis Update: Training Decision Paper* (AIR, 2009).

[8] AIR recommended edits to the labels and definitions of six existing competencies and language for five new competencies. OHC representatives presented these recommendations to their Competency Working Assessment Group (CWAG) for approval.

Recommendations

The result of the job analysis was an updated view (and documentation) of the BDO tasks, KSAOs, and tools and equipment required for the job. Although some core aspects of the job remained unchanged, the expanded nature of the work (e.g., —walking the line,| working in pairs with other BDOs) led AIR to recommend that TSA revise their selection system[9] by refreshing the current content of the Structured Interview and adding new selection measures to the hiring process. AIR began this process for TSA with the test specifications step described below.

Test Specifications

In the second phase of the study, AIR used the results of the job analysis to develop test specifications for an experimental selection battery. This section of the report summarizes the purpose, methods, results, and recommendations. A thorough review of the test specifications phase of the research is provided in another technical report (AIR, 2010b).

Purpose

Test specifications provide a framework for the content and structure of selection measures. The ultimate goal is to align the pre-hire KSAOs with the experimental selection measures, demonstrating optimal coverage of the critical KSAOs (Russell & Peterson, 1997). For example, Table 2 illustrates a notional relationship between five pre-hire KSAOs and four potential selection measures. In this example, Ability #1 is assessed by the Problem Solving Measure. Ability #2 is assessed by both the Work Sample Test and the Structured Interview. Such overlap is typical in a test specifications document, ensuring that the KSAOs are adequately measured by the selection system (Russell & Peterson, 1997).

Table 2. Example of KSAO – Assessment Crosswalk

KSAO	Problem Solving Measure	Work Sample Test	Structured Interview	Role-Play Exercise
Ability #1	X			
Ability #2		X	X	
Skill #3		X	X	
Skill #4		X		X
Other Characteristic #5			X	X

Note. The relationships between KSAOs and selection measures presented in this table are for demonstrative purposes only.

[9] Currently, TSA hires from an internal applicant pool of Transportation Security Officers (TSOs) using a Quick Hire Application, Structured Interview, and medical screening.

Method

AIR began by prioritizing the list of pre-hire KSAOs for inclusion in the experimental selection battery. The goal of this first step was to identify KSAOs that were (1) required at the time of hire, (2) important for successful performance of the job, and (3) related to the greatest

proportion of the job. First, through an empirical review of the job analysis data, AIR identified KSAOs that were rated as —required before training‖ by a majority of the SMEs, had average importance ratings of 4 or higher (on a 5-point scale),[17] and linked to at least 50% of the critical job tasks.

Next, AIR refined the list through a rational review. AIR reviewed the list of KSAOs to determine which were suitable for use in selection versus training. This was a slightly different question than the SMEs had answered when completing the job analysis survey (i.e., at what time is the KSAO required?). Accordingly, AIR re-categorized seven KSAOs as appropriate for selection and two KSAOs as appropriate for training. AIR also removed KSAOs to decrease redundancy in measurement and added some to include KSAOs revealed as important through the qualitative job analysis results and current BDO selection system.

Upon refinement of the list of KSAOs appropriate for inclusion in the experimental selection battery, AIR reviewed the existing research literature to identify selection measures that would appropriately assess those critical KSAOs. Potential measures were reviewed using three categories of criteria: (1) descriptive (e.g., the relevance to the intended construct), (2) psychometric (e.g., reliability, validity, potential for subgroup differences), and (3) operational constraints (e.g., time to develop, administration requirements) (Hendrickson, Matheson, Amodeo, Norris, & Sparano, 2008; Russell, Norris, & Goodwin, 2000). Using this information, AIR recommended revisions and additions to the existing BDO selection system.

Results and Recommendations

The test specifications represent the results and recommendations of this phase of the research (see Appendix E). These illustrate the plan for the experimental selection battery and outline the relationship between the prioritized list of KSAOs and the recommended selection measures. Although TSA currently uses an internal applicant pool when recruiting and selecting BDO job candidates, AIR ensured that the experimental selection battery would also be appropriate for use with an external applicant pool. As described below, the test specifications included an experimental selection battery that would include both currently used (i.e., the Structured Interview) and additional measures.

Currently Used Measures

TSA currently requires internal BDO job candidates to undergo a medical screening. AIR recommended continued use of the screening and added that, if an external applicant pool is used, a similar medical evaluation should be required for those individuals. [18]

[17] SMEs indicated how important a KSAO was for successful job performance using a 5-point scale (1 – *not important* to 5 – *extremely important*).

[18] AIR did not examine the medical evaluation in subsequent phases of the study because it was beyond the scope of this research effort.

Internal TSO job candidates currently take part in a Structured Interview, and this assessment could also be given to external candidates. The current Structured Interview measures eight TSA competencies: Oral Communication, Teamwork, Social Skills, Multi-Tasking, Decision-Making, Memory, Critical Thinking, and Conscientiousness (see Appendix F for a complete list of the TSA competencies and definitions). Despite the breadth of content that the current Structured Interview covers, AIR recommended adding items to assess job candidates' Honesty-Integrity.

Additional Measures

AIR recommended adding three new measures to the experimental selection battery. These measures would allow TSA to assess additional aspects of competencies that are already measured by the selection system as well as evaluate additional competencies.

First, AIR recommended including a measure of critical thinking ability. The existing BDO selection system assesses this skill with a single question during the Structured Interview. Using a separate critical thinking measure, such as the Watson-Glaser Critical Thinking Appraisal – Short Form (herein referred to as the Watson-Glaser Critical Thinking Appraisal), would allow for direct assessment of job candidates' ability to infer information, recognize assumptions, perform deductive reasoning, interpret information, and evaluate arguments.

Second, AIR recommended developing a Work Sample Test[19] to assess the specific cognitive abilities and underlying skills necessary for BDO screening duties (e.g., speed of closure, sustained attention, selective attention, memory, mathematical operations). Although individual measures of these specific abilities exist, a BDO-specific Work Sample Test would allow for simultaneous assessment of numerous abilities.

Third, AIR suggested adding a Role-Play Exercise to measure job candidates' skill in gathering information while engaging passengers in conversation. This Role-Play Exercise would allow for a more specific assessment of Oral Communication as it relates to the context of the BDO job. To capitalize on the resources that were available, AIR recommended administering the Role-Play Exercise at the end of the Structured Interview. This approach would reduce the number of times that Assessors would need to convene during the hiring process.

These additional measures had the potential for enhancing the BDO hiring process. With this in mind and approval from TSA, AIR began the third research phase: measurement development. Measurement development is described in Chapter III, which provides an overview of the measures, the underlying competencies they assess, and the process used to procure and/or develop them.

[19] A work sample test asks job candidates to perform tasks that are similar to those they will complete on the job.

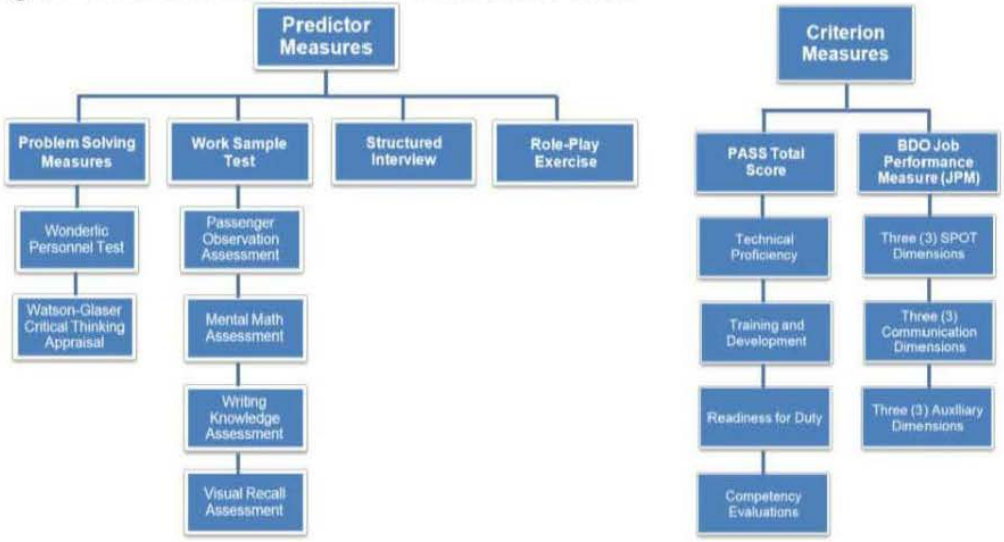
Sensitive Security Information (SSI)

CHAPTER III: MEASUREMENT DEVELOPMENT AND PROCUREMENT

Overview

Based on the list of knowledge, skill, abilities, and other characteristics (KSAOs) identified during the job analysis (AIR, 2010a) and the recommendations from the test specifications, AIR developed and/or procured eight predictor (i.e., selection) measures for the pilot test and validation study. To evaluate these assessments using a criterion-related design, AIR also collected job performance ratings using two criterion measures. An overview of the measures included in the pilot test and validation study is presented in Figure 2.

Figure 2. Overview of Predictor and Criterion Measures



For the predictor measures, AIR procured two Problem Solving measures and developed a Work Sample Test[20] and Role-Play Exercise. AIR also updated TSA's existing Structured Interview. In addition, AIR developed a feedback questionnaire to collect additional information from the BDOs during the pilot test and validation study. For the criteria measures, AIR obtained performance ratings from TSA's Performance Accountability and Standards System (PASS) and developed a measure specifically for this study, the BDO Job Performance Measure (BDO JPM).

A brief summary of the predictor measures is presented in Table 3. Specific relationships between each measure and a prioritized list of pre-hire KSAOs are presented in Appendix E.

[20] A work sample test asks job candidates to perform tasks that are similar to those they will complete on the job.

Sensitive Security Information (SSI)

Table 3. Summary of Predictor Measures

Measure	Description	Constructs Assessed
Wonderlic Personnel Test	Job candidates complete a series of verbal and quantitative reasoning questions on a timed (12 minutes), 50-item, paperandpencil test.	□ Problem Solving
Watson-Glaser Critical Thinking Appraisal	Job candidates form inferences, recognize assumptions, deduce and interpret statements, and evaluate arguments on a timed (40 minutes), 40-item, paperandpencil test.	□ Critical Thinking □ DecisionMaking
Passenger Observation Assessment	Job candidates observe passengers waiting in a queue on a five minute video recording and look for specific behaviors or appearance factors.	□ Multitasking □ Attention to Detail □ Speed of Closure □ Sustained Attention □ Selective Attention □ Decision-Making □ Critical Thinking □
Mental Math Assessment	Job candidates sum values that appear every second without taking notes or using scratch paper.	□ Attention to Detail □ Math

Sensitive Security Information (SSI)

WARNING: THIS RECORD CONTAINS SENSITIVE SECURITY INFORMATION THAT IS CONTROLLED UNDER 49 C.F.R. PARTS 15 AND 1520. NO PART OF THIS RECORD MAY BE DISCLOSED TO PERSONS WITHOUT A —NEED TO KNOW,I AS DEFINED IN 49 C.F.R PARTS 15 AND 1520, EXCEPT
WITH THE WRITTEN PERMISSION OF THE ADMINISTRATOR OF THE TRANSPORTATION SECURITY ADMINISTRATION OR THE SECRETARY OF TRANSPORTATION. UNAUTHORIZED RELEASE MAY RESULT IN CIVIL PENALTIES OR OTHER ACTION. FOR U.S. GOVERNMENT AGENCIES, PUBLIC DISCLOSURE GOVERNED BY 5 U.S.C. 552 AND 49 C.F.R. PARTS 15 AND 1520.

Writing Knowledge Assessment[10]	Job candidates review a brief passage and circle any errors.	☐ Attention to Detail ☐ Written Communication
Visual Recall Assessment	Job candidates review two images of an airport setting for two minutes each and then answer multiple-choice questions about their content.	☐ Attention to Detail ☐ Speed of Closure ☐ Sustained Attention ☐ Selective Attention ☐ Memory ☐

[10] Note that associated documents may refer to this assessment as —Written Summary.l AIR changed the name of this measure following the pilot test to better reflect the abilities assessed by this measure.

Table 3. Summary of Predictor Measures—Continued

Measure	Description	Constructs Assessed
Structured Interview	Job candidates are asked eight questions designed to measure eight competencies.	Attention to Detail Decisiveness Flexibility Honesty-Integrity Interpersonal Skills Multitasking Self-Management Teamwork Oral Communication
Role-Play Exercise	Job candidates engage in a conversation with a fictitious passenger to elicit three pieces of information about the person's trip.	Oral Communication Social Skills Decision-Making Memory

A detailed description of the procurement or development process of all measures is presented in the following sections.

Problem Solving Measures

The results of the job analysis (AIR, 2010a) indicated that critical thinking was an important preemployment KSAO for the BDO job. As a result, AIR recommended including a separate critical thinking measure, the Watson-Glaser Critical Thinking Appraisal (Watson-Glaser Critical Thinking Appraisal, 1998), in the experimental selection battery. This measure allows for a direct assessment of a job candidate's ability to infer information, recognize assumptions, perform deductive reasoning, interpret information, and evaluate arguments. The Watson-Glaser Critical Thinking Appraisal has a large body of psychometric evidence. For instance, this measure yields scores that are reliable (α = .80) and that relate to supervisor ratings of performance (r = .23 - .33) (Watson-Glaser Critical Thinking Appraisal, 1998).

Given this large body of evidence, AIR identified the Watson-Glaser Critical Thinking Appraisal as a potential alternative for the newly developed selection measures (i.e., the Work Sample Test and Role-Play Exercise). That is, if the newly developed measures did not demonstrate adequate levels of reliability and validity, then the Watson-Glaser Critical Thinking Appraisal could be a viable alternative. On the other hand, if the newly developed selection measures were functioning appropriately, then the Watson-Glaser Critical Thinking Appraisal could serve as an additional measure of construct validity.

In addition, AIR recommended including the Wonderlic Personnel Test (Wonderlic Personnel Test, 2001), a measure of problem solving that emphasizes verbal and quantitative reasoning, as a comparison measure during the validation study. This measure has been studied extensively. For example, higher scores on the Wonderlic Personnel Test have been associated with higher job performance ratings for employees in a number of jobs (r = .22 - .67). Also, it has demonstrated adequate levels of test-retest reliability and internal consistency (Wonderlic Personnel Test, 2001). Finally, scores on the test have been found to be associated with other measures of problem solving (e.g., the Wechsler Adult Intelligence Scale). Given its established psychometric properties, AIR included the Wonderlic Personnel Test in the validation study as an additional measure of construct validity. In other words, AIR could determine whether scores on the Wonderlic Personnel Test overlapped with scores from the newly developed selection measures in patterns that would be expected given their proposed underlying constructs and the existing literature.

To procure the Problem Solving measures, AIR contacted the publisher of each test to identify the administration processes and explore different pricing options. AIR planned to administer the assessments on-site during the pilot test and validation study. TSA indicated that many sites may have difficulty providing computer access to all of the BDOs who were participating. Thus, both Problem Solving measures needed to be in paper-and-pencil format. Nonetheless, both measures are available in an electronic format, which may be more appropriate if the Problem Solving measure was administered on a large-scale, nationwide basis.

For the Wonderlic Personnel Test, AIR contacted Wonderlic Inc. Wonderlic Inc, does offer discounted prices to Government agencies, but only for purchases that include more than 500 copies of the test. Wonderlic, Inc. is currently revising their discounted prices, thus the cost estimates in this report represent the non-discounted prices (see Table 4).

Table 4. Pricing Options for the Wonderlic Personnel Test

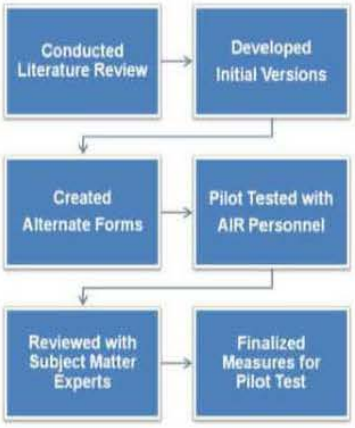
Number of Tests Purchased	Cost Per Test	Total Cost
25	$15.00	$375.00
50	$10.75	$537.50
100	$6.25	$625.00
500	$5.75	$2,875.00
1,000	$5.25	$5,250.00

For the Watson-Glaser Critical Thinking Appraisal, AIR contacted Pearson, Inc. Pearson Inc. sells the Watson-Glaser Critical Thinking Appraisal as part of a larger assessment package. Specifically, users can purchase an assessment package that includes two test manuals, a scoring key, 25 test booklets, and 25 answer sheets for $472.00. Packages of 25 additional answer sheets can be purchased for $125.00. Currently, Pearson Inc. does not provide discounted rates for Government agencies.

Work Sample Test, Structured Interview, and Role-Play Exercise

The development process for the predictor measures consisted of six steps, depicted in Figure 3 below. These steps were taken to ensure that the selection measures were based on relevant research, adequately assessed the appropriate constructs, and were relevant to the BDO job. Steps were also taken to ensure the test security of the selection measures. These included, for example, establishing a separate shared network space for qualified project personnel, encrypting work files with an internal password, and following prescribed procedures for working with Sensitive Security Information (SSI).

Figure 3. Overview of the Development Process



First, AIR reviewed the existing literature for recommendations on building specific types of selection measures. Second, AIR developed preliminary versions of the selection measures. During these initial stages, AIR considered a number of formats and item-types for each measure. Third, AIR used the preliminary versions of each of the selection measures as a prototype for developing alternate forms. Fourth, each version of the selection measures, excluding the Structured Interview, was pilot tested internally with AIR research staff.[11] Next, AIR reviewed the selection measures with SPOT Transportation Security Managers (STSMs) and representatives from TSA's Office of Security Operations (OSO) during four separate workshops. Subject matter experts (SMEs) were asked to provide feedback on the difficulty of the measures, clarity of the instructions, and relevance to the BDO job. Last, AIR finalized the measures in preparation for the formal pilot test to be conducted with job incumbents and managers. After each step in the process, AIR edited and revised the selection measures. The following sections describe each of these steps in more detail, starting with steps one and two (literature review and initial measure development).

[11] The existing Structured Interview had been validated by TSA and was widely used thoroughout the agency. AIR made only minor changes to the administration process, and development of the new items was modeled directly after the TSA items.

Literature Review and Initial Development

During the early phases of the development process, AIR reviewed relevant professional literature and created initial versions of each selection measure. These initial versions consisted of test stimuli (e.g., photographs, videos), response booklets, and administration guides. The administration guides included information on establishing a proper testing environment, scripts for proctoring each assessment, and scoring keys. The response booklets included the instructions that were read aloud by the Administrator. This section describes the early stages of development for each selection measure.

Passenger Observation Assessment

For the Passenger Observation Assessment, AIR reviewed literature that pertained to developing the Work Sample Tests. This review revealed that an important consideration is whether the Work Sample Test will emphasize the outcome of a job candidate's performance or evaluate the process used to reach that outcome (Borman, Bryant, & Dorio, 2010). For instance, a Work Sample Test for mechanics may emphasize whether the job candidate fixes a piece of equipment (i.e., the outcome) or completes the appropriate steps in the correct order (i.e., process). This consideration, although subtle, influences how a work sample test is designed and scored. For the Passenger Observation Assessment, AIR chose to emphasize the process that job candidates use to identify passengers (i.e., what behaviors were selected and what calculations were completed), rather than simply evaluating the outcome (i.e., which passengers were selected for additional screening). This approach would provide a more finite assessment of a job candidate's critical KSAOs.

Another consideration when developing work sample tests is whether performance is evaluated using a continuous rating scale or dichotomous scoring (Borman et al., 2010; Truxillo, Donahue, & Kuang, 2004). Continuous rating scales allow one to make more refined distinctions among job candidates (e.g., identify those who are more skillful, faster, or more efficient). These scales, however, typically require a trained observer to provide subjective ratings. Given the complexity and unobservable nature of the task being evaluated with the Passenger Observation Assessment, it would be very difficult for evaluators to reliably assess job candidates' performance using a continuous rating scale. A dichotomous scoring system (i.e., the job candidate did or did not identify the correct behavior), on the other hand, would not require subjective evaluations and could be completed after a job candidate had finished the assessment. Because a dichotomous approach would be more efficient, while still providing useful information about a job

candidate's ability, AIR adopted it as the method of scoring for the Passenger Observation Assessment.

Developing the Stimuli

During the early phases of development, AIR considered different methods of displaying passengers to job candidates. One potential approach consisted of showing photographs of an airport terminal to job candidates and asking them to indicate whether passengers displayed particular behaviors or appearance factors. This approach would be relatively easy to administer and would resemble the job more closely than using non-visual stimulus (i.e., written scenarios). Presenting photographs of passengers, however, limited the types of behaviors that could be observed because the passengers were not moving or interacting with each other. Thus, to ensure a high-fidelity assessment, AIR decided to use a video recording of passengers in an airport setting. This presentation method, although requiring more resources and administration time, greatly enhanced the realism of the Passenger Observation Assessment. Also, the dynamic nature of the video recording increased the number of potential items by providing more passengers and a wider variety of behaviors and appearance factors for job candidates to observe. The Department of Homeland Security (DHS) provided AIR with a video recording on November 12, 2010. The video recording consisted of 30 minutes of footage of a security screening line. The video recording faced the screening line so that passengers came closer to the camera as they progressed through the queue.

After selecting a video-based presentation method, AIR considered the appropriate length of footage to use. The video needed to be long enough to provide a valid evaluation of a job candidate's sustained attention. If the video was too long, though, it may introduce fatigue effects or become overly difficult to administer. AIR considered using video recordings that ranged in length from two to ten minutes and decided that five-minute video recordings provided an adequate evaluation of sustained attention without being overly lengthy. This AIR decision was based on a consideration of candidate's fatigue, the number of items that could be developed, the amount of time required to adequately assess the competencies, and the amount of overall testing time.

Developing the Items

AIR also considered the number of passengers, behaviors, and appearance factors that a job candidate would be required to observe. These numbers would have a direct impact on the difficulty of the assessment. Thus, AIR sought to include enough passengers to make the assessment challenging, but not overly difficult. Likewise, AIR sought the optimal number of

behaviors and/or appearance factors for each passenger to display. Passengers who displayed multiple behaviors and/or appearance factors were more likely to be more challenging than those with few or no behaviors and/or appearance factors.

The type of behaviors and appearance factors could also impact the difficulty of the assessment. AIR avoided selecting behaviors or appearance factors that were tricky, misleading, or difficult to observe clearly. For instance, —wearing a backpack‖ could not be used as an appearance factor because job candidates may infer the presence of a shoulder strap to mean that a passenger was wearing this item. Similarly, a backpack may be confused with other types of baggage (e.g., purses). These ambiguous behaviors or appearance factors would make the scoring process more challenging. AIR also avoided selecting behaviors or appearance factors that were similar to existing SPOT indicators. When possible, passengers that displayed SPOT-related behaviors or appearance factors were removed from the video.

To identify passengers, behaviors, and appearance factors that were suitable, AIR thoroughly reviewed the 30-minute video recording. Each passenger in the video was observed from when they entered the screening line to when they exited the screen. During these observations a list of potential passengers, behaviors, and appearance factors was developed. This list was then reviewed by the research team to identify behaviors or appearance factors that could be problematic as foci of the assessment. Problematic passengers, behaviors, or appearance factors were removed. The revised list of passengers, behaviors, and appearance factors was used as the foundation for developing the initial version of the Passenger Observation Assessment.

After identifying an initial set of passengers, AIR developed brief written descriptions of each passenger to be included in the job candidate‘s response booklet. These descriptions were intended to help job candidates ensure they had identified the appropriate passenger. While developing the descriptions, AIR did not refer to the passengers by the color of their clothes or belongings. Since color-vision is not currently a requirement of the BDO job, this approach was taken so that individuals who were color-blind would not be at a disadvantage while completing the assessment.

Finally, the initial version of the Passenger Observation Assessment included an evaluation of job candidates‘ ability to compute simple arithmetic in their heads. Specifically, while watching the video recording, job candidates were instructed to indicate what behaviors or appearance

factors they observed and then were asked to compute a passenger's total score using those observed factors and their associated point values in the fictitious screening scenario.

Although this approach closely resembles the duties of a BDO, it presented a number of challenges for scoring a job candidate's performance. For instance, if an individual were to incorrectly select a passenger for additional screening, it could be attributed to (1) overlooking a behavior or inferring a behavior that did not occur, or (2) incorrectly summing the point values. AIR concluded that differentiating between these two sets of abilities (i.e., mental arithmetic and observation) was necessary and would provide important information for hiring officials. Thus, the Passenger Observation Assessment was redesigned so that job candidates were only asked to indicate whether they had observed a particular passenger or appearance factor. An additional assessment that focused solely on mental arithmetic was created.

Mental Math Assessment

For the Mental Math Assessment, AIR reviewed studies that examined other measures of quantitative reasoning (e.g., Ryan & Paolo, 2001). Most existing quantitative reasoning measures assessed mathematical abilities (e.g., multiplication, algebraic principles) that were more complex than the abilities required for the BDO job. Furthermore, many of the existing measures did not require time limits that resembled the pace at which BDOs need to complete their mental calculations. Thus, AIR developed an assessment that closely resembled the type of math performed by BDOs and the pace at which these calculations needed to be made. AIR also reviewed literature that provided suggestions for timing the presentation of stimuli (Ayers, 1953). This research provided recommendations on the pace at which items could be presented to job candidates.

As with the Passenger Observation Assessment, AIR considered different methods for presenting the items on the Mental Math Assessment to job candidates. Initial drafts were in paper-based format. Job candidates would be given a limited amount of time to complete a series of basic addition problems in their response booklet. These problems included values that ranged from one to three and did not exceed totals of thirteen. The timing of the test was established at a point where few, if any, participants would complete all of the items. Although this approach was an efficient assessment of mental arithmetic, it overemphasized processing speed to a degree that did not correspond with the duties of a BDO. Specifically, to obtain sufficient variability in scores, job candidates would need to complete too many problems too quickly. Furthermore, because all of the values for each problem were presented simultaneously, job candidates were able to use multiplication rather than rely solely on addition. For instance, if an item consisted of

multiple —2‖ values, then a job candidate could simply multiply these values to calculate a subtotal. Because BDOs observe behaviors in a serial fashion (i.e., as they appear), the reliance on multiplication minimized the degree of fidelity between the Mental Math Assessment and the BDO job.

In order to address these issues, AIR developed an alternative presentation method that relied on a Microsoft PowerPoint slideshow. Specifically, each item involved the serial presentation of numeric values that ranged from one to three. Each value appeared on the screen for 1.5 seconds. After each item, job candidates were given three seconds to record their score. This presentation method improved upon the original paper-and-pencil design in two ways. First, it required job candidates to complete their mental computations quickly without overemphasizing processing speed. That is, by providing job candidates with three seconds to record their scores, they were able to complete each item and prepare for the subsequent items. Second, because values were presented serially, as opposed to simultaneously, job candidates were unable to use multiplication to compute the total scores. This helped increase the degree of fidelity between the Mental Math Assessment and the BDO job.

Writing Knowledge Assessment

AIR initially considered several different approaches for measuring job candidates‘ written communication skills, along with a review of criteria for evaluating a job candidate‘s writing sample (Mansilla, Duraisingh, Wolfe, & Haynes, 2009; Rawson, Quinlan, Cooper, Fewtrell, & Matlow, 2005). These options are described below, starting with three that were ultimately not used. As noted below, the final approach was chosen for many reasons, including implications for measure effectiveness and implementation.

One option involved asking job candidates to draft a brief summary of a prescribed incident or event. Although this option would resemble the duties of a BDO, it was deemed to be timeconsuming and difficult to score. Furthermore, a job candidate‘s ability to discern the type of information that should be included in a report may be most appropriate for training, rather than pre-employment selection.

As a second approach, AIR developed other methods for assessing Writing Knowledge that featured closed-ended items, such as multiple-choice questions or binary response options (Haladyna, 2004). With one option, job candidates would review a paragraph that featured a series of dichotomous items. Job candidates would then select the option that correctly followed spelling, grammar, and punctuation rules (e.g., using the appropriate form of —their/there‖).

This option was far more efficient than a writing sample, but it also afforded job candidates with a 50% chance of receiving credit for each item by simply guessing.

Next, to increase the difficulty of this assessment using a closed-ended item approach, AIR considered presenting job candidates with a passage and having multiple-choice questions refer to particular underlined sentences or words. Some items would require job candidates to identify the correct spelling of a word, whereas others would involve selecting the section of sentence that featured an error. Having four plausible response options limited the effect of guessing, but it was difficult to develop multiple plausible response options. In many instances, a job candidate could compare the different response options and easily identify the correct response.

Finally, to ensure that the Writing Knowledge Assessment could differentiate among job candidates while maintaining its ease of administration, AIR ultimately developed a passage in which job candidates would have to identify the errors. Specifically, job candidates were asked to review the passage and circle spelling, grammar, and punctuation errors. This approach limited the effects of guessing and allowed for the administration to small groups of job candidates in a few minutes.[12]

During development, AIR reviewed examples of SPOT Referral reports to ensure that the passage would resemble the content of typical BDO reports. Specifically, AIR examined exemplary and poor examples of reports from BDOs in the field. To ensure the confidentiality of the parties involved, TSA removed identifying information pertaining to BDOs or passengers from the reports. These examples informed the type of content that was included in the passage to enhance the realism of this assessment.

Visual Recall Assessment

AIR chose to use a picture-based, multiple-choice, paper-and-pencil format to assess a job candidate's ability to retain and recall visual information. AIR had previously developed similar tests for measuring recall abilities in previous research (Jeanneret & Associates, & AIR, 2000).

[12] AIR also considered requiring job candidates to identify which aspects of the passage would be relevant for a witness statement or internal security report. Although this would help enhance the realism of the assessment and its fidelity to the BDO job, it would focus on the type of knowledge provided during training, rather than the knowledge required for pre-employment selection.

Initially, AIR derived images for the Visual Recall Assessment from the footage of the security screening line that was used in the Passenger Observation Assessment. Specifically, a screenshot was taken of a portion of the video recording that was not used for that assessment. The image was then resized so that job candidates would focus on a particular section of the screening line. Although this approach provided a realistic stimulus, the resolution of the image limited the content of the questions. For instance, questions that referred to items in a passenger's hand (e.g., a bottle or cell phone) were not included in the initial item pool because of the poor resolution with these images.[13]

AIR considered a number of formats for the multiple-choice items on the Visual Recall Assessment such as using three, four, or five response options. AIR chose to use items with four response options because these types of items typically minimize the effect of guessing while avoiding poor functioning distractors that do not increase the difficulty of an item or may be misleading (Haladyna, 2004). Then, AIR reviewed the literature on best practices for item writing (Haladyna, 2004). Based on this review, AIR avoided incomplete item stems, chose not to use fill-in-the-blank questions, and avoided items that referred to ambiguous topics. For instance, items referring to the number of male or female passengers were excluded because those items would require test-takers to infer the gender of a passenger. AIR also avoided referencing the color of clothing or objects within the items. Again, this approach was taken so that individuals who were color-blind would not be at a disadvantage when completing the Visual Recall Assessment.

Structured Interview

Rather than design a new Structured Interview, AIR updated the existing measure by developing new items and modifying the existing rating scales. Before developing items for the Structured Interview, AIR reviewed literature to identify best practices. These sources outlined criteria for effective interview items and recommendations for developing rating scales (Pursell, Campion, & Gaylord, 1980). Researchers recommend developing interview items that:

- Were accurate, complete, and unambiguous;
- Reflected the content of the job;
- Resembled the appropriate level of complexity for the job; and

[13] As discussed in the Alternate Forms section below, AIR later replaced the screen shots with photos taken by TSA, which provided higher resolution images and more flexibility for item development.

- Minimized or eliminated biased language or jargon.

AIR also reviewed TSA's existing Structured Interview items and rating scales as well as the definitions for the eight competencies that are assessed with the existing measure (TSA, 2006). This process ensured that the new items did not duplicate an existing item, but still resembled the original measure.

In addition to updating items for the Structured Interview, AIR recommended providing participants with a handout—distributed at the beginning of the session—that described the competencies assessed during the Structured Interview. This change was recommended on the basis of feedback received during the job analysis. Some BDOs indicated that it was difficult to determine the relationship between the existing Structured Interview items and the BDO job. The job analysis as a whole, though, suggested that the competencies being measured by the Structured Interview were job relevant. Thus, AIR recommended providing job candidates with a list of the competencies assessed in order to improve the overall acceptance of the measure without reducing its reliability or validity (Klehe, Konig, Richter, Kleinmann, & Melchers, 2008).

AIR also considered different approaches for evaluating an additional job-related competency: Honesty-Integrity. First, AIR reviewed off-the shelf Honesty-Integrity measures. These measures are often grouped into two categories: overt measures (i.e., those that ask directly about previous stealing) and covert measures (i.e., those that emphasized underlying personality traits) (Sackett & Wanek, 1996). Although these measures typically show adequate levels of reliability and validity, they can be susceptible to faking and are illegal in Massachusetts and Rhode Island (Cullen & Sackett, 2004). Furthermore, an Honesty-Integrity measure that focused solely on a job candidate's likelihood of stealing was judged to be too narrow in scope for the BDO selection system.

Given the challenges associated with off-the-shelf integrity measures, AIR identified literature on using Structured Interview questions to assess honesty (Hollowitz, 1999; Jones & Terris, 1991). Although this body of research was limited, it did provide examples of previously developed items and suggestions for creating rating scales. Using these examples as templates, AIR generated new Structured Interview items to measure job candidates' Honesty-Integrity. For the initial version of the Structured Interview, AIR created behavior-based items.

Behaviorbased items ask job candidates to describe past experiences in order to assess a competency.

Items were based on TSA's definition of the Honesty-Integrity competency and other definitions from the literature. AIR also used task statements from the job analysis that were strongly linked to Honesty-Integrity as a foundation for writing the new items. In addition, using the existing Structured Interview as a model, a new rating scale was also developed for these items.

An initial challenge with these behavior-based Honesty-Integrity items was creating questions that provided adequate levels of variability without being invasive. It was important that the correct response to an item was not so obvious that the job candidate could easily provide the appropriate response. However, some of the initial items were discarded because, although they were challenging, they addressed sensitive subjects (e.g., instances where someone felt guilty, condoning dishonest behavior).

Role-Play Exercise

As with the initial development phases of the Work Sample Test and the Structured Interview, AIR identified literature relevant to the development of Role-Play Exercises. In particular, AIR considered recommendations for developing and implementing one-on-one simulations and oralfact finding exercises (Lievens, 1999; Thornton & Mueller-Hanson, 2004). One-on-one simulations consist of a job candidate interacting with a trained Resource Person (i.e., roleplayer). Job candidates are typically evaluated on their oral communication skills. Oral-fact finding exercises, on the other hand, require job candidates to elicit specific pieces of information from a panel or role-player. Job candidates are typically evaluated on their questioning techniques and the amount of information they gather. These sources provided recommendations for training role-players, scoring performance, and standardizing the administration process.

AIR also reviewed materials about the content of Role-Play Exercises, such as articles pertaining to active listening and other questioning techniques (Knippen & Green, 1994). AIR also leveraged its knowledge of the SPOT program by reviewing previous deliverables for the BDO selection research and other related research efforts (e.g., Mullaney, Makonnen, & Costigan, 2009). AIR used this information as a foundation for the Resource Person's[14] script and ratings scales for the Role-Play Exercise.

[14] Resource Person refers to the individual playing the role of the fictitious passenger (i.e., role-player).

After reviewing the literature and other sources, AIR began initially conceptualizing the RolePlay Exercise. First, AIR developed two rating scales to evaluate a job candidate's performance. These rating scales were designed to measure two important aspects of the casual conversation process. The first rating scale (Active Listening) assessed how well a job candidate could conduct the conversation in a non-threatening, friendly manner. The second rating scale (Elicitation of Information) evaluated a job candidate's questioning techniques. The structure of both rating scales resembled that of the Structured Interview rating scales.

Initial development also informed the appropriate length of the conversation between the job candidate and the fictitious passenger. It was important to find a time limit that was not too long and would be unrealistic in a BDO's natural job setting. On the other hand, the time limit needed to afford job candidates the opportunity to gather answers to their assigned questions. AIR set the initial time limit for the Role-Play Exercise at five minutes.

The next decision to make was who should play the role of the fictitious passenger (i.e., Resource Person)? Ideally, this individual would be someone who could be trained to standardize his or her performance across job candidates. A standardized performance ensures that job candidates receive the same testing process during the Role-Play Exercise. Furthermore, this person would need to remain impartial during the Role-Play Exercise. Some options included SPOT managers, human resource (HR) personnel, and members of the National Deployment Operations (NDO) team. After consulting with TSA's Office of Security Operations, AIR recommended using a member of the NDO.

Another decision point was determining the appropriate level of difficulty for the Role-Play Exercise. Specifically, AIR considered how reticent or uncommunicative the Resource Person should be while interacting with the job candidate. For instance, the Resource Person could provide the job candidate with inconsistent information or be evasive during the conversation. Although this could enhance the difficulty of the assessment, and may reflect some of the situations that BDOs face, AIR decided not to include these types of challenges. Instead, the Resource Person's character was designed to be as forthcoming and friendly as possible. This was done to avoid any instance of inconsistency where one Resource Person may provide a more challenging performance than another. Also, AIR anticipated that this assessment would be inherently difficult for most job candidates and would differentiate among high and low performers without these additional challenges. Furthermore, the Role-Play Exercise was designed to assess a candidate's basic communication abilities, and not advanced questioning

techniques or skills. These basic abilities could then be enhanced with job knowledge and further training.

Finally, during the initial phase of development, AIR decided that job candidates should stand during the Role-Play Exercise. Asking job candidates to stand may increase their performance anxiety. On the other hand, it would increase the realism of the Role-Play Exercise. AIR concluded that the benefits of the additional realism outweighed the potential increase in performance anxiety.

Alternate Forms

After creating initial versions of each selection measure, AIR developed alternate forms of each predictor. These alternate forms were developed for two reasons. First, alternate forms help enhance the test security of an assessment because if a single form is compromised, the alternate form may be implemented in its place. Second, developing alternate forms of an assessment allowed AIR to evaluate separate tests and determine which one was most likely to provide useful information during the hiring process. These evaluations were made following the pilot test and validation study.

Developing alternate forms essentially requires creating multiple forms that were as similar as possible. This ensures that job candidates receiving different versions of an assessment undergo the same testing experience. Alternate forms are generated using the same development processes and by relying on the same test specifications (Kolen & Brennan, 2004).

Alternate forms are not the same as equated forms, however. Formal equating processes require that measures include the same items or are administered to the same participants during the development process (or randomly equivalent groups). In the current study, a formal equating process was not possible for two reasons. First, some measures relied on distinct stimuli. For example, one form of the Passenger Observation Assessment consisted of a separate segment of video recording than its counterparts. Likewise, each form of the Visual Recall Assessment featured different images of an airport scene. Thus, formally equating the forms of these assessments using equivalent items was not possible. Second, administering different forms of each assessment to the same BDOs or to randomly equivalent groups of BDOs would have required a substantial amount of resources. Doubling the number of participants who would need to complete each assessment would have significantly impacted the security operations at each airport. Also, exposing multiple forms of a single assessment to the same participants would have limited the test security of these measures. Given these

challenges, AIR decided to rely on alternate, or parallel, forms of these assessments and forego a formal equating process. This section describes the development process for these alternate forms.

Passenger Observation Assessment

AIR developed four versions of the Passenger Observation Assessment. AIR used three criteria to ensure that each version was parallel. First, AIR considered the total number of passengers that a job candidate would observe. As the total number of passengers increased, the difficulty of the assessment increased because the job candidate was responsible for observing and evaluating more information. Second, AIR examined the passenger sequencing across the four forms. In other words, AIR examined the amount of time that elapsed between each passenger being observed. Decreasing the amount of time between identified passengers would increase the difficulty because job candidates would need to observe more information in a shorter period of time. Third, AIR required job candidates to observe a similar number of behaviors and appearance factors for each version. This number affects the total maximum score that a job candidate could receive for each form. Although AIR sought to ensure the parallelism of the four forms, there were instances when the three criteria conflicted with one another. Nevertheless, after reviewing each of the versions, AIR determined that the forms were similar enough to merit further investigation during internal and external pilot testing.

Mental Math Assessment

AIR created two forms of the Mental Math Assessment. Two criteria were used to ensure the parallelism of these measures. First, AIR designed both forms to include an equal proportion of different types of mental math items. Mental math items differed based on the number of values that a candidate was required to add (e.g., 2 values, 3 values, 4 values). Each form required job candidates to complete items that included two to seven values with a majority of items asking job candidates to sum four to five values. Second, AIR developed each assessment to gradually increase in difficulty. Specifically, earlier items featured two to three values for a job candidate to add, whereas items near the end of the assessment required job candidates to sum six to seven values. This approach was taken to allow job candidates to build their confidence and familiarity with the assessment before reaching the more difficult items.

Writing Knowledge Assessment

AIR developed two forms of the Writing Knowledge Assessment. While writing the passages for the two forms, AIR sought to align the content but change superficial details. For instance, one passage describes the behaviors of a male passenger, whereas the other describes a female

passenger. Likewise, one version describes the events of a Tuesday morning, whereas the other refers to a Thursday evening. Changing this type of superficial information minimizes the extent to which job candidates can disclose information about the test without requiring the creation of divergent measures (Lievens & Anseel, 2007). AIR also reviewed the types of errors that job candidates were asked to identify in each form. This rational review was conducted to ensure that the forms included the same number and types of errors (i.e., spelling, punctuation, and grammar).

Visual Recall Assessment

For the initial version of the Visual Recall Assessment, AIR used one image with a large pool of items. These items were then separated into two different forms. This approach enhanced the parallelism of the items by using the same image for both forms. AIR ensured that items were not duplicated across forms. Also, items that could serve as a cue for the correct response in subsequent items were separated across the two forms.

For subsequent versions, TSA provided AIR with multiple images of an airport setting. TSA provided these images on January 7, 2011. AIR reviewed these images and selected those that (1) clearly depicted a variety of stimuli in the background and (2) featured multiple passengers in the screening line. AIR developed three forms of the Visual Recall Assessment. AIR included two images for each form in order to develop a large enough item pool (Haladyna, 2004). AIR drafted items for each image and these were reviewed to ensure the clarity of the item stem and response options. AIR also reviewed the item pool to ensure items were not duplicated across images within form, and previous response options or items stems would not signal the correct response for subsequent items.

Structured Interview

AIR developed four separate sets of interview questions. Because the current Structured Interview allows Administrators to choose between two interview items, AIR developed the new items in pairs. In other words, two of the items were designed to be administered together, and the other two were designed as a pair. While writing each interview question, AIR sought to ensure that the pairs of items were parallel in content. This was done by targeting the pair of questions to specific components of the TSA competency. For instance, while developing questions that evaluated the Flexibility competency, AIR developed two items that measured task flexibility (i.e., the ability to adapt to changing work conditions) and two items that measured interpersonal flexibility (i.e., the ability to adapt one's beliefs about another person).

Role-Play Exercise

Initially, AIR created two characters for Role-Play Exercise. To enhance the test security of the Role-Play Exercise, each character provided different responses to the questions a job candidate was tasked with answering. For instance, one character was traveling to Miami, Florida while another was going to Kansas City, Missouri. In addition to modifying the answers to the questions, other superficial information was also changed. For example, the passenger's profession and hometown differed. Other information that could impact the Resource Person's performance (e.g., the passenger's attitudes about security screening) was held constant (Lievens & Anseel, 2007). This approach was taken to increase the standardization of the exercise and was similar to the one used for designing the two separate versions of the Writing Knowledge Assessment.

Internal Pilot Testing

This section describes the internal pilot testing conducted with AIR research staff. First, the purpose and process of these pilot tests is reviewed. This section concludes by presenting the changes that were identified and implemented for each selection measure.

Purpose and Process

As part of the development process, AIR conducted internal pilot tests with AIR personnel. Specifically, between six to ten staff members completed at least one form of each measure (see Table 5).

Table 5. Number of AIR Personnel Who Completed Internal Pilot Tests

Selection Measure	N
Passenger Observation Assessment	10
Mental Math Assessment	6
Writing Knowledge Assessment	9
Visual Recall Assessment	7
Role-Play Exercise	6

These pilot tests were conducted in an iterative fashion. In other words, research staff completed revised versions of each assessment after changes had been made based on the results of the previous pilot test. Each pilot test session began with a brief overview of the project. AIR personnel either completed the assessments individually or in pairs. Then, participants provided

feedback about the difficulty of the assessments, the clarity of the instructions, and ways to improve the tests. Administrators also provided feedback. All feedback was then reviewed by the entire research team and changes were implemented. This section reviews the primary changes that were identified and implemented during this phase of the development process.

Passenger Observation Assessment

The initial internal pilot tests of the Passenger Observation Assessment indicated that this measure was far too difficult. Test-takers indicated that they had difficulty locating the passengers they were tasked with observing. Also, they stated that the number of observation criteria and passengers was too challenging. On the basis of this feedback, AIR reduced the number of behaviors and appearance factors that job candidates were required to observe from ten to eight. Also, the response booklets were reformatted so that responses were recorded on only two pages. This helped minimize the amount time job candidates would spent locating the appropriate response sheet as they completed the assessment.

Test-takers from subsequent pilot tests commented that the assessment was moderately difficult, but not overly challenging. Nevertheless, they identified criteria (e.g., holding a cell phone) and particular passengers that could be considered tricky or misleading. These types of items were removed from both forms of the Passenger Observation Assessment to ensure parallelism. These test-takers also indicated that they had difficulty identifying which passengers they were required to observe. As a result, AIR revised the descriptions of the identified passengers to make them easier to identify.

A majority of the research team who participated in the internal pilot test, many of whom had participated in the job analysis, indicated that the assessment provided a realistic representation of the BDO job. This finding suggested that the Passenger Observation Assessment would provide job candidates with a realistic job preview and allow them to make informed decisions about their degree of fit with the BDO job.

Mental Math Assessment

Members of the research team that participated in the internal pilot tests indicated that the Mental Math Assessment was moderately difficult. Some individuals were able to answer all of the items correctly, whereas others answered 80% of the items correctly. Although these percentages were relatively high, AIR anticipated that these values were likely to decrease with a population of job candidates who may have different educational or employment backgrounds.

Also, many test-takers indicated that although the assessment became increasingly challenging as it progressed, they did not feel the items were unfair or overly difficult.

Test-takers also commented that presenting the items using a Microsoft PowerPoint® format prevented them from using multiplication to develop subtotals for each item. This feedback supported the use of a serial presentation of point values in order to provide a higher fidelity measure of BDO job-related mathematical ability.

Writing Knowledge Assessment

During early rounds of internal pilot testing, test-takers described the Writing Knowledge Assessment as too easy and commented that many of the errors were easy to identify. AIR revised the items so that the spelling and grammatical errors were not as salient. These revisions appeared to be effective as test-takers described subsequent versions of the Writing Knowledge Assessment as moderately difficult.

Visual Recall Assessment

Members of the research team who completed earlier versions of the Visual Recall Assessment indicated that it was difficult to respond to specific items due to the poor resolution of the images. Specifically, many test-takers indicated that it was difficult to distinguish between the specific details on some of the more challenging items (e.g., what items the passengers were holding in their hands). AIR used Microsoft Office Picture Manager® software to increase the brightness of the image to improve the resolution of certain areas of the image. Also, AIR resized the image to remove passengers or objects that were misleading or unclear.

Subsequent administrations identified issues with the wording of the items. For example, some test-takers commented that it was unclear which passengers were being referenced in certain items. Also, some response options in certain items were identified as being cues for subsequent items in the assessment. These items were either revised or removed from subsequent versions of the Visual Recall Assessment.

Structured Interview

AIR chose not to internally pilot test the Structured Interview items with project personnel for two reasons. First, the major components of the Structured Interview had been previously validated by TSA (TSA, 2006). Specifically, TSA had conducted studies to collect evidence of content validity for the rating scales and Structured Interview items. Second, internal pilot testing would have required a considerable investment of resources. Specifically, each

administration of the Structured Interview required approximately 60 minutes for two to three research staff. Thus, conducting multiple administrations of each form of this measure was prohibitive. Because the newly developed interview items were modeled after the existing interview items and designed to measure competencies that were previously linked to the BDO job, AIR determined that it would be more beneficial to internally pilot test the other newly developed selection measures.

Role-Play Exercise

Internal pilot tests of the Role-Play Exercise revealed a few of areas for improvement. First, early pilot tests suggested that many of the conversations were unlikely to last longer than two minutes. Thus, to improve the efficiency of the measure, AIR reduced the time limit from five minutes to two minutes.

Second, AIR incorporated instructions into the Resource Person's materials that allowed the passenger to respond to an opening icebreaker. Prior to this change, the Resource Person had been instructed to answer any question that referred to information not contained in his or her script with —I don't know.‖ This response would indicate to a job candidate that he or she needed to redirect the conversation. However, during internal pilot testing, there were instances where a job candidate would ask an opening question to build rapport (i.e., —How's the weather outside?‖) and the Resource Person would respond, —I don't know.‖ This response made the conversation stilted and penalized the job candidate for a natural opening question. To account for this type of question, the instructions for the exercise were changed to allow the Resource Person to respond to an opening icebreaker with an ad-libbed response if relevant information was not contained within the script. For the remainder of the conversation, however, the Resource Person would respond to off-script questions with —I don't know.‖

Third, some staff who served as the Resource Person stated that there was too much information to memorize during the introduction. As a result, AIR created a bulleted summary sheet of the information that the Resource Person could use as a handout while performing. This helped ensure that the information provided by different role-players was consistent across administrations.

Fourth, to provide additional context about the situation for the job candidates, one participant suggested including a photograph of a BDO interacting with another passenger. This information would provide external job candidates, who are less likely to be familiar with the

BDO job, a mental image of the uniform and checkpoint environment. AIR identified a photograph that could be distributed to job candidates during the introductory part of the administration.

Subject Matter Expert Review of Predictors

This section describes the four workshops AIR conducted with SMEs. First, the purpose and methodology for the workshops is reviewed. The section concludes by presenting the results from each of the workshops.

Purpose and Process

In addition to conducting internal pilot tests with members of the research team, AIR reviewed the newly developed selection measures with SMEs. These SMEs were familiar with the BDO job and could provide additional feedback about the difficulty of the items, the clarity of the instructions, and the degree of job-relatedness of the assessments. Four separate workshops were conducted with SMEs. Each workshop focused on one or two different measures. A summary of the participants, process, and major recommendations for each workshop is presented below.

Review of the Structured Interview and the Role-Play Exercise

The first workshop was held remotely on December 2, 2010 with STSMs from different airports from across the country (n = 10). On average, these SMEs had been managers in the SPOT program for 3.2 years (SD = 1.0 year) and had been working for TSA for 7.9 years (SD = 0.6 years). Most SMEs (n = 6) had received additional training in behavioral indicators of deception. On average, the total amount of training equaled 35.3 hours (SD = 18.4 hours). Seven out of 10 of the SMEs were male and most identified themselves as White (n = 6).

Many of these managers had participated in previous workshops during earlier phases of the research (e.g., the job analysis workshop, working meetings to develop the BDO Job Performance Measure). As a result, most SMEs were already familiar with the purpose of the current study.

After a brief overview of the study, SMEs were presented with materials for the Structured Interview and Role-Play Exercise. AIR personnel began the workshop by reviewing the competency handout that would be provided to job candidates. Next, SMEs were asked to review the three pairs of Honesty-Integrity items. Finally, SMEs reviewed materials for the Role-Play Exercise, including the job candidate's instructions, the Resource Person's script, and the rating scales. SMEs also considered the effect of asking job candidates to stand during

the Role-Play Exercise. These materials were presented using Microsoft Live Meeting® webinar software. This allowed SMEs to view the documents without being able to modify them. This approach, as opposed to sending the materials to workshop participants, was taken to ensure the security of the test materials.

Structured Interview

Some SMEs expressed concern about providing the definition of the competencies to job candidates prior to the Structured Interview. These individuals stated that this information may allow job candidates to unduly modify their responses to improve their ratings. Other SMEs, however, indicated that the handout would improve the Structured Interview because job candidates would be more likely to provide answers that corresponded with the questions and rating scales. After further deliberation, the SMEs agreed that the handout would be beneficial, but that it should only be used during the initial introduction session. In other words, job candidates would be presented with the competency definitions, be afforded some time to review these definitions, and then return the handout to the Administrator.

A majority of the SMEs suggested that the addition of the Honest-Integrity competency would add value to the interview process. However, the SMEs were concerned that some of the items may prompt job candidates and help them provide the correct response. For instance, one item asked job candidates to describe a time when a cashier had overcharged them for something. The SMEs noted that using the term —overchargedl would steer job candidates towards a particular response. SMEs were also concerned about the scoring of items that asked individuals about situations in which they had condoned or perpetrated a dishonest act. In particular, the SMEs commented that it would be challenging to evaluate job candidates who indicated that they had never performed a dishonest act. Although this response would receive full credit using the rating scale, some SMEs indicated that they would doubt the sincerity of such a response.

In addition, one SME stated that there was a risk of losing good job candidates who may actually possess adequate levels of Honesty-Integrity because they do not describe behaviors that align with the rating scale. In other words, if a job candidate does not describe his or her past experience using the key terms that were included in the rating scale, then he or she may not receive a passing rating.

Finally, one SME indicated that it may be best to approach the Honesty-Integrity items using situational interview items rather than behavioral-based items. Situational interview items ask job candidates to respond to hypothetical scenarios. Because this type of interview question

would rely on hypothetical situations, they would be less likely to be invasive for job candidates. Furthermore, because this type of question does not ask about past experiences, Assessors would not need to question a job candidate's sincerity if a candidate indicated that he or she had never performed a dishonest act.

Role-Play Exercise

In terms of the Role-Play Exercise, SMEs indicated that this measure would be an important addition to the BDO hiring system. One SME stated that it would be —outstanding‖ to include a formal assessment of the skills needed to engage passengers and elicit information. The SMEs commented that one of the biggest challenges with the Role-Play Exercise would be identifying the appropriate personnel to serve as a Resource Person. Many SMEs stated that the Resource Person should be familiar with the SPOT program and existing protocols. In addition, some SMEs indicated that it would be important to have the same person serve as a Resource Person during a single round of hiring. This would help improve the standardization of the Role-Play Exercise. Some SMEs, however, mentioned that this requirement would be difficult for their airports to meet.

Along with these considerations for the Resource Person, SMEs offered suggestions for improving the clarity of the instructions for job candidates. For example, some of the SMEs wanted to include instructions to inform the job candidates that they cannot directly ask about the requisite information (b)(3):49 U.S.C. § 114(r) These SMEs commented that it was important to evaluate a job candidate's ability to integrate questions into a conversation and that this instruction should be more salient for the job candidate. Also, the SMEs commented that the instructions should mention that the Resource Person has been instructed to respond with —I don't know‖ when he or she is asked information that is not included in the character's script.

With respect to the standing requirement for the Role-Play Exercise, many of the SMEs commented that this it would make the assessment more realistic. Although some SMEs acknowledged that this could be more anxiety provoking for some test-takers, a majority of the SMEs indicated that they would be comfortable performing this exercise while standing.

Finally, SMEs also suggested adding a requirement at the end of the Role-Play Exercise where job candidates would need to summarize the information they had elicited from the job candidate. The SMEs indicated that this part of the assessment would resemble the types of interactions that BDOs have with other security personnel when they need to share the results

of their conversations with passengers. SMEs recommended evaluating these summary statements on their relevance and accuracy to the questions that a job candidate was tasked with answering. One SME questioned whether this summary component would introduce an evaluation of a job candidate's recall ability. One SME responded that it could enhance the Role-Play Exercise because BDOs have to recall their conversations when sharing their findings with other security personnel.

Demonstration of the Work Sample Test

The second workshop was held on-site at AIR in December, 2010. STSMs (n = 2) and representatives from the SPOT program office (n = 2) attended this workshop. On average, these SMEs had worked for TSA for 7.9 years (SD = 1.2 years) and had been in their current position for 3.5 years (SD = 1.4 years). Half of the SMEs had received additional training in behavioral indicators of deception. Three of the four SMEs were male and most identified themselves as White (n = 3).

During this workshop, SMEs were asked to complete the Passenger Observation Assessment, the Writing Knowledge Assessment, and the Visual Recall Assessment and provide feedback. As a facilitator reviewed the correct responses, SMEs scored their own assessments. This procedure allowed SMEs to identify any items that were confusing or misleading. After scoring their responses, SMEs were asked to comment on the difficulty of the measure, the clarity of its instructions, and barriers to administering it in the field. These conversations were guided with a semi-structured protocol and facilitated by two members of the research team.

Passenger Observation Assessment

After reviewing the Passenger Observation Assessment, the SMEs provided suggestions for improving the clarity of the instructions. These included:

- Providing a more detailed description of the scoring process,
- Including a screenshot from the video recording in the example item so job candidates can understand the different components of the measure (e.g., passenger descriptions, behaviors and appearance factors),
- Providing additional time for the job candidates to review the behaviors and appearance factors after the example item,
- Informing the job candidates that they only need to mark a behavior or appearance factor once per passenger, and

- Instructing job candidates to continue to observe passengers until they exit the screen.

SMEs also provided feedback regarding the difficulty of the Passenger Observation Assessment. Some SMEs mentioned that it was difficult to identify passengers using the written descriptions. Some recommendations for making the passengers more salient included highlighting passengers on the screen with symbols (e.g., boxes or arrows) and indicating when a job candidate is responsible for observing a passenger with a tone or sound. Other SMEs indicated that it was difficult to see specific details in the video recording. These individuals suggested requiring that job candidates are seated within a specified distance from the projection screen or television monitor. (b)(3):49 U.S.C. § 114(r)

(b)(3):49 U.S.C. § 114(r)

Writing Knowledge Assessment

Overall, SMEs recommended including the Writing Knowledge Assessment as part of the experimental selection battery. However, they expressed concern about screening out job candidates who performed well on the other assessments, but did not score highly on the Writing Knowledge Assessment. Some SMEs suggested weighting this measure less than the other measures in the system or setting a lower cut score to ensure that job candidates who possess other important skills and abilities (e.g., observation) would not be screened out from the list of qualified candidates.

In addition, some SMEs suggested that the test would be more job relevant if it required testtakers to identify information that would be pertinent for inclusion in a SPOT Referral report. Despite this suggestion, the SMEs acknowledged that this particular skill may require prior job knowledge and would best be addressed during training.

Visual Recall Assessment

SMEs also provided feedback on the Visual Recall Assessment. First, SMEs suggested including directions in the response booklet that indicated the points at which a job candidate would need to wait for additional instructions before proceeding. This would help ensure that job candidates are not exposed to the test items before returning to the image.

SMEs also provided suggestions about how to improve the test items. Specifically, the SMEs identified items that were ambiguous or unclear. Also, the SMEs suggested drafting additional items that focused on actions or behaviors that were displayed in the image. In other words, instead of asking job candidates to identify amounts (e.g., How many security officers were in the image?), items would focus on what passengers were doing (e.g., Which passenger was

opening his or her suitcase?). SMEs indicated that this would help improve the job-relatedness of the assessment.

Finally, all of the SMEs indicated that the Visual Recall Assessment was difficult. This was primarily attributed to the clarity of the images. For example, the SMEs commented that it was difficult to answer questions about passengers in the second row of the security screening line because the details were unclear. The SMEs indicated that they would provide new images of airport settings with higher resolution.[15] Also, to facilitate the item development process, the SMEs indicated that these images would include fewer passengers and more background details (e.g., signs, luggage).

Review of the Mental Math Assessment and the Structured Interview

The third and fourth workshops were held remotely on January 26, 2011 and February 1, 2011 with representatives from the SPOT program office. Three SMEs participated in the first workshop and two SMEs took part in the second workshop. On average, these SMEs had worked for TSA for 2.5 years (*SD* = 3.2 years). All of the SMEs were Program Analysts and at least one had worked in the SPOT program or served as a Training Manager. All of the SMEs had received additional training in behavioral indicators of deception. All of the SMEs were female and identified themselves White.

During the first workshop, AIR reviewed the Mental Math Assessment and Structured Interview with SMEs. The second workshop focused solely on the Structured Interview. Specifically, AIR provided a rationale for including a separate Mental Math measure and reviewed the current version of this assessment. AIR then asked the SMEs to provide feedback on the presentation method, difficulty, and job-relatedness of this selection measure. SMEs indicated that this assessment would be beneficial to include as part of the pilot test. They indicated that it could contribute unique information and was an important ability to assess.

After reviewing the Mental Math Assessment, the SMEs reviewed the revised Structured Interview. Specifically, for each set of four items, AIR first presented the competency definition and rating scale. This provided the SMEs with the opportunity to familiarize themselves with the underlying competency and performance criteria for the items. Next, AIR presented the four items as a group. SMEs were asked to consider how well the items aligned with the intended competency and whether the item would be problematic during administration. AIR also reviewed the revised interview probes with the SMEs. As with the first workshop, test materials

[15] AIR received these additional images from TSA on January 7, 2011.

were presented using Microsoft Live Meeting® webinar software. This allowed SMEs to view the documents without being able to modify them. This also ensured the test security of these materials during the development process.

For the Structured Interview, SMEs suggested writing additional items that focus on the specific behaviors within TSA's competencies that were most job relevant. For example, for Attention to Detail, SMEs suggested developing items that assessed how well a job candidate can process and recall visual and auditory information about individuals and the environment. SMEs suggested that the new Flexibility items focus on a job candidate's —mental flexibility‖ (e.g., adjusting one's thinking or beliefs in response to new information) rather than —task flexibility‖ (e.g., adjusting one's schedule). Also, SMEs recommended that the new Multitasking items emphasize job candidates' ability to integrate and use multiple pieces of information simultaneously.

In addition to recommendations for how to align the Structured Interview items to the BDO job, SMEs also identified questions that they believed would be problematic. For example, for the Teamwork competency, SMEs indicated that the item —*Describe a time when you worked with teammates to accomplish a goal,*‖ which was meant to lead a job candidate to describe how he or she interacted with his or her teammates, would instead lead job candidates to describe the general processes or actions of a team. The SMEs recommended developing items that emphasized the interaction between a job candidate and his or her team members.

SMEs also provided feedback regarding the Honesty-Integrity items that AIR had developed. Overall, the SMEs indicated that this competency was important for the BDO job and that it needed to be assessed during the hiring process. After reviewing the behavior-based items, though, the SMEs expressed concerns about whether job candidates would provide valid responses. One SME suggested that job candidates would very likely unduly modify their responses to achieve higher ratings. Also, one SME indicated that some job candidates may feel uncomfortable discussing past experiences in which they were dishonest.

On the other hand, SMEs suggested that the situational Structured Interview items may provide useful information. Specifically, they indicated that this format would increase the likelihood of eliciting an appropriate and valid response. Given their stated preference for the situational items, AIR asked the SMEs to describe scenarios in which a BDO's Honesty-Integrity was critical. These scenarios were used to develop the situational Honesty-Integrity items.

Predictor Measures Finalized for Pilot Test

Using the information collected during each step of the development process, AIR finalized the predictor measures in preparation for the pilot test. This section provides a description of each of these measures at this stage of their development.

Passenger Observation Assessment

The pilot test version of the Passenger Observation Assessment assessed test-takers' ability to observe and identify behaviors and appearance factors. This assessment consisted of an example item, practice item, and test items. The example item provided test-takers with an image that was similar to the type of footage included in the assessment and instructions for completing the response sheets. After the example item, test-takers completed a practice item. The practice item provided test-takers with a preview of the type of footage and additional instructions. Specifically, the practice item consisted of a 1-minute video recording from a different segment of footage than the actual test. During the 1-minute video recording, test-takers observed two passengers. After the video recording, the Administrator reviewed the correct responses with test-takers and provided specific instructions. For example, test-takers were informed that they should not infer that a behavior or appearance factor was displayed. Instead, test-takers must actually observe the behavior or appearance factor. During the test, test-takers watched a 5minute video and were responsible for observing eight passengers and eight behaviors and appearance factors.

Mental Math Assessment

The pilot test version of the Mental Math Assessment required test-takers to complete a series of addition problems in their heads. This assessment consisted of one example item, four practice items, and 20 test items. All of the items were presented using a Microsoft PowerPoint® Slideshow that featured automatic timing mechanisms. For each item, test-takers were presented with a series of values that appeared on the screen for 1.5 seconds. After 1.5 seconds, another value appeared. This process continued until three to seven values had appeared. After the last value was on the screen for 1.5 seconds, a —Record Your Score‖ slide replaced the test item slide. During this time, test-takers had three seconds to record the value they had calculated for the previous slide and prepare for the next item. Because the Mental Math Assessment measures test-takers' ability to compute calculations in their head, test-takers were not allowed to take notes or scratch marks on their papers. If they made these marks, they received a score of zero.

Writing Knowledge Assessment

During the pilot test version of the Writing Knowledge Assessment, test-takers were asked to review a brief summary of an airport scenario and identify writing errors. Test-takers first

completed a practice test that consisted of three sentences. There were three errors embedded in the sentences. After the test-takers completed the practice test, the Administrators reviewed the correct responses. For the actual Writing Knowledge Assessment, test-takers were given time to review the paragraph and identify grammatical, spelling, punctuation, and sentence structure errors. There were multiple errors embedded in the paragraph.

Visual Recall Assessment

The pilot test version of the Visual Recall Assessment required test-takers to review images of an airport setting and then answer questions about the images. Specifically, test-takers were presented with an image in an envelope, asked to remove the image, and given time to review the image. After their review, test-takers returned the image to the envelope and answered a series of multiple-choice questions about the content of the image. The process was then repeated with a second image.

Structured Interview

The revised Structured Interview consisted of seven behavioral-based interview items. For these items, test-takers were asked to think of a past experience that pertained to a specific competency. For example, test-takers were asked to describe an experience that demonstrated Teamwork. In addition to these behavioral-based items, the revised Structured Interview also included a situational interview item that assessed a test-takers' Honesty-Integrity. For this item, test-takers were presented with a hypothetical situation and asked how they would respond. Finally, test-takers were evaluated on their Oral Communication abilities based on their performance on both types of items.

Role-Play Exercise

The pilot test version of the Role-Play Exercise required test-takers to engage a fictitious passenger to elicit three pieces of information in a non-threatening manner. They were first provided with a description of the context of the Role-Play Exercise (b)(3):49 U.S.C. § 114(r) (b)(3):49 Next, test-takers were told the information that they needed to elicit from the passenger. Before beginning their conversation, test-takers had one minute to plan their approach. They then had two minutes to engage the passenger and elicit the three pieces of information. At the conclusion of the conversation, test-takers were asked to remain standing and state the three pieces of information that they had gathered. Test-takers were evaluated using two ratings that assessed their ability to maintain a goal-oriented conversation (i.e., Elicitation of Information) and maintain a friendly, non-threatening demeanor (i.e., Active

Listening). Also, the accuracy of the test-takers' responses to the questions was evaluated. Specifically, test-takers received credit for each question that they correctly answered at the conclusion of the Role-Play Exercise.

Feedback Questionnaire

Job candidate's reactions to selection measures can influence whether they view the selection process as fair and transparent (Gilliland, 1993). Furthermore, these perceptions can influence their decision to complete the selection process, accept job offers, and refer other job candidates to the organization (Hausknecht, Day & Thomas, 2004; Macan, Avedon, Paese, & Smith, 1994). The SPOT program currently identifies potential BDOs using an internal job candidate pool (i.e., TSOs). Job candidates who are rejected during the selection process are likely to remain with the organization. Given the impact of a job candidate's reactions, it was critical to evaluate BDOs' perceptions of the selection system in order to ensure that the newly developed measures will be well received by future job candidates.

Feedback from job candidates would be difficult, if not impossible, to gather by just using test scores. Thus, AIR developed a feedback questionnaire for the pilot test and validation study. Similar to the predictor measures, AIR reviewed relevant literature on organizational justice and job candidate reactions to inform the development process (e.g., Gilliland, 1993; Truxillo, Steiner, & Gilliland, 2004). Based on this review, AIR identified three measures that were most relevant to the current study. First, Bauer et al.'s (2001) Selection Procedural Justice Scale was included because it is a widely-used measure of job candidate reactions. This measure consists of two higher-order factors. Each factor is comprised of separate subscales. The first factor, Structure, consists of items that assess specific reactions to the selection process (e.g., jobrelatedness of the measure, opportunity to perform). The second factor, Social, consists of items that assess the communication with and treatment of job candidates.

Second, Arvey, Strickland, Drauden, and Martin's (1990) perceived difficulty measure was included as part of the questionnaire in order to determine test-takers' perspectives on the difficulty of the selection measures. During the internal pilot test, AIR observed that participants stated that some measures were especially challenging (e.g., Visual Recall). As part of the development process, AIR wanted to ensure that job candidates did not perceive the measures as overly difficult or unfair. Thus, this scale allowed AIR to analyze these perceptions across the pilot test and validation study samples.

Third, AIR included a measure of performance beliefs as part of the feedback questionnaire (Schmitt, Oswald, Kim, Gillespie, & Ramsay, 2004). Performance belief refers to how well a job candidate thought he or she performed on a selection measure. This information, like the difficulty subscale, was helpful in determining whether job candidates perceived the selection measures to be overly difficult or unfair.

AIR adapted items from each measure so that they pertained to the current study. For example, AIR reduced the number of items for some measures and modified the wording on some items so that these pertained to the BDO job. These revised items were then reviewed by the research team to ensure that they were clear and assessed their intended construct. Table 6 lists the subscales within the feedback questionnaire. This table also provides a brief definition of each subscale and an example item for each.

Table 6. Feedback Questionnaire Subscales and Example Items

Subscale	Definition	Example Item
Content Job-Relatedness[16]	The extent to which a test appears to measure content relevant to the job situation.	The content of this component was clearly related to the BDO job.
Predictive Job-Relatedness[27]	The extent to which a test appears to be valid.	A person who scored well on this component will be a good BDO.
Opportunity to Perform[27]	Having adequate opportunity to demonstrate one's knowledge, skills, and abilities within the testing situation.	I could really show my skills and abilities through this component.
Two-Way Communication[27]	The opportunity for job candidates to offer input or to have their views considered during the test.	There was enough communication during the testing process.
Propriety of Questions[27]	The extent to which questions avoid personal bias, invasion of privacy, and illegality.	This component did not seem too personal or private.
Perceived Difficulty[17]	The extent to which a test was difficult or challenging.	This component was too easy for me.

[16] Adapted from Bauer et al. (2001).
[17] Adapted from Arvey et al. (1990).

Absolute Performance Beliefs[18]	The extent to which examinees believed they did well on a test.	I am confident that I performed well on this component.
Relative Performance Beliefs[29]	The extent to which examinees believed they performed better than others on a test.	I am confident that I will be evaluated more highly than others on this component.

In addition to these items, AIR included a comment box at the end of the questionnaire. This open-response item allowed BDO participants to provide additional feedback concerning the difficulty of the measures, the clarity of the instructions, and the relevance to the job. BDOs were also encouraged to use this space to provide specific recommendations for improving the selection measures. The feedback questionnaire used during the pilot test is included in Appendix G.

Criterion Measures

Essential to any validation study are well conceptualized, job relevant criteria. Criteria, broadly defined, are standards or evaluations by which employees' success or failure can be measured (Bass & Barrett, 1981; Guion, 1965; Landy, 1989). Regardless of how criteria are specified, the selection of valid and reliable criteria is essential. The validity of the criterion is important as it will affect the degree to which the inferences drawn about the effectiveness of the selection measures are accurate (Ployhart, Schnieder, & Schmitt, 2006). Further, the reliability of a criterion measure is important as it establishes the upper bound for validity and therefore has important implications for evaluating the selection battery (Nunnally & Bernstein, 1994). As part of the validation study, AIR collected job performance data using multiple criterion measures. This section begins with a brief overview of the alternative approaches for collecting criterion measures that AIR considered. Next, the existing performance data that AIR received from TSA is described. Finally, the development process of a criterion measure that was designed specifically for this study is described.

Examined Different Alternatives

This section reviews the two general types of criteria that AIR considered including for both the pilot test and validation study: objective and subjective criteria. Each type of criteria provides different types of information about employee performance and as such are often only weakly correlated and should not be used interchangeably (Bommer, Johnson, Rich, Podsakoff,

[18] Adapted from Schmitt et al. (2004).

& MacKenzie, 1995). These broad classes of criteria are described below, followed by the advantages and disadvantages associated with each.

Objective Criteria

Objective criteria are measures of job performance that focus on countable behaviors or outcomes. Examples of objective criteria include the production of output (quantity), quality of output, avoidable absenteeism, lateness, turnover, promotion rate, salary increase rate, safety rates, accident rates, and error rates (Ployhart, Schneider, & Schmitt, 2006). In the context of the BDO job, objective criteria may include the number of referrals, the number of arrests made from referrals, or the number of prohibited items seized. Below the advantages and disadvantages of using these types of criteria are discussed.

Advantages of Objective Criteria

Among the several types of objective criteria, there are various advantages. First, objective criteria often represent important outcomes for organizations (e.g., quantity of output, frequency of absenteeism, error rates). In turn, these types of criteria are often tracked for administrative purposes and are thus readily available for use. Second, these criteria typically represent data that directly results from (e.g., production quantity) or is reflective of (e.g., number of days absent) job performance. That is, these data represent objectives of the work performed by an individual. Finally, these measures are often accepted by employees, because objective data are thought to be unchallengeable (Gatewood & Feild, 2001). For example, the number of arrests resulting from an individual BDO is likely to be a data point that is free from interpretation or question by an employee. Similarly, these measures do not introduce the rater errors or biases that must be considered when using subjective measures.

Disadvantages of Objective Criteria

Although potential objective criteria for BDOs are identifiable and available (e.g., referrals, arrest rate, the number of prohibited items seized), objective criteria can be problematic. First, objective criteria often reflect the measurement of outcomes of job performance rather than the measurement of job performance itself. For example, using the number of referrals made by a BDO as a criterion is not the same as measuring the behaviors that are required to perform the tasks involved in conducting a referral. Second, objective criteria are often deficient in that they may not measure all of the components that comprise organizationally relevant behavior of interest (Borman, et al., 2010). That is, because performance is a complex, multidimensional construct, measuring a single variable, such as absenteeism, is likely to provide incomplete or misleading information about the nature of employee performance. Third, objective measures do

not account for situational factors that are beyond the employee's control (Ployhart et al., 2006). For example, if arrest rates were used as a criterion for measuring BDO job performance, an employee's performance ratings would be (a) based on a task (e.g., arresting a passenger) that may happen infrequently, and (b) affected by the behavior and characteristics of the passengers at a given time. Also, the final decision to arrest a passenger ultimately lies with the attending law enforcement officer. Therefore, objective criteria, if not carefully selected, may afford BDOs little control over their performance scores.

Subjective Criteria

Subjective criteria involve the appraisal of employee performance by informed raters (e.g., immediate supervisors). Performance ratings can be collected using a variety of methods as well as information from a variety of sources (Cardy & Dobbins, 1994; Dipboye, Smith, & Howell, 1994). Similar to objective measures, there are specific advantages and disadvantages associated with each type of format and each rating source used. For example, different rating *methods* will vary in the amount of time and effort it takes to create the rating tools used to appraise performance. Conversely, different rating *sources* will vary in the degree to which they are able to observe various aspects of an employee's performance. Both rating methods and rating sources vary in the degree to which they are susceptible to a number of rater errors or biases.

Advantages of Subjective Criteria

Unlike objective criteria, subjective criteria (i.e., performance ratings) focus on the process (i.e., behavior) that leads to an outcome, and can account for the full scope of an employee's job performance behaviors. Performance ratings can be completed using a number of methods (e.g., Behavioral Anchored Rating Scales, Behavioral Observation Scales, Behavioral Checklists),[19] which allows organizations to choose the most appropriate approach for their system (Guion, 1998). Evaluations of job performance can be obtained from a variety of rater types that may be supervisory or non-supervisory in nature. Supervisor ratings are most frequently used, as supervisors often have familiarity with the employee being rated and knowledge about what is required to perform the job. Immediate supervisors typically are most familiar with an individual's performance and have had the broadest opportunity to observe the individual (Bernardin & Beatty, 1984). In fact, research has shown that supervisory feedback is more highly related to job performance than feedback from any other source (Becker & Klimonski, 1989). Some performance dimensions, however, may lend themselves to ratings by peers or

[19] Each type of rating scale is described in more detail in the —Developed Performance Measurel section in Chapter III.

subordinates, who have more frequent interaction or opportunity to observe the employee. Last, self ratings are often useful for identifying training and development needs (Guion, 1998).

Disadvantages of Subjective Criteria

One of the biggest disadvantages of subjective criteria is their susceptibility to rater errors. Examples of commonly occurring rater errors include leniency bias (i.e., providing high ratings regardless of whether these reflect the actual performance observed), central tendency bias (i.e., assigning ratings that are predominately neutral, or in the middle of the scale), and halo bias (i.e., providing consistently high or low ratings based on one's general impression of an individual, rather than his or her performance along specific job performance dimensions). These rater errors can have deleterious effects on the psychometric characteristics of a criterion measure (Pulakos, 1984). Research, however, has shown that the quality of ratings can be improved by providing rater training. Training on rater error provides a description and examples of errors such as halo, leniency, severity, and similar-to-me effects, as well as guidance for avoiding these errors (Borman et al., 2010). Frame-of-reference training is utilized to improve the accuracy of ratings. This type of training involves providing information and behavioral examples for each performance dimension, and to the extent possible, practice and feedback on making ratings (Borman et al., 2010; Cascio, 1998).

Obtained Existing Performance Data

TSA currently uses a variety of objective and subjective measures to assess BDO job performance. The section below provides a description of each of these existing measures and their utility for the BDO validation study.

PASS

TSA recently implemented their PASS, a pay-for-performance employee appraisal system (TSA, 2010), which focuses on: technical proficiency, BDO competencies, and other job-relevant criteria (i.e., Training and Development Evaluation, and Readiness for Duty Evaluation). Specific measures for each component are described in Table 7 below.

Table 7. Description of PASS Measures

	PASS Measure	Description
	Job Knowledge Test	On-line job knowledge test that features both true-false and multiple-choice items that assess a BDO's knowledge of SPOT protocols and other screening techniques.

Technical Proficiency	BDO SPOT Proficiency Evaluations	Bi-annual observation of a BDO while he or she performs 22 SPOT-related behaviors in a live environment.
	Practical Skills Evaluation	Annual on-the-job assessment of a BDO's ability to conduct physical bag searches and checkpoint/explosive trace detection tasks.
	Attention to Detail	Bi-annual evaluation of BDOs using three-point rating scale (i.e., Group I – Group III). BDOs are assigned a particular rating if they demonstrate *most but not all* of the behaviors associated with that rating
	Decisiveness	
	Interpersonal Skills	
Competencies	Oral Communication	
	Teamwork	
Other	Training and Development Evaluation	Evaluation of the number of training courses a BDO completed within his or her annual development plan.
	Readiness for Duty Evaluation	Evaluation of the number of leave requests that were submitted using the proper protocol and number of dress code violations during the performance period.

In addition to the separate measures within PASS, each BDO receives a total score. This total score is based on a weighting system in which the technical proficiency and competency evaluations account for (b) of the final rating.

PASS, being a multi-component evaluation system, allowed AIR to consider different existing performance measures for use in the validation study. Although each measure provides unique information, many of the measures reviewed above, while useful for the purposes of performance evaluation, were not appropriate for the validation of the selection battery.[20] The usefulness of these measures was limited due to statistical considerations. Specifically, many of the subcomponents within the PASS scores showed restricted ranges and low levels of variability. In some instances, range restriction was observed because scoring below a certain threshold on specific components of PASS (e.g., Job Knowledge Test, Practical Skills Evaluations) qualifies one for removal from the SPOT program. Thus, many job incumbents,

[20] It is important to emphasize that measures which are used for performance evaluation in practice may not be appropriate for validation study research. Thus, the limitations of measures from PASS described herein should be interpreted using this context; they should not be interpreted as limitations or criticisms of their use in practice.

and participants in the current validation study, were unlikely to score low on these measures. Other measures demonstrated low levels of variability because of dichotomous scoring systems (i.e., the Readiness for Duty and Training and Development Evaluation).

Because these components of PASS limit the total variance of BDOs' scores which is likely (or possible) to be observed, this limited variance would artificially attenuate validity coefficients. In other words, because of a limited range of possible scores these measures may show an artificially weak relationship with the predictor measures (Nunnally & Bernstein, 1994). Therefore, these component measures were not included as criteria in the concurrent validation study.

It is important to note that although the subcomponents of PASS demonstrated range restriction and limited variability, this was not the case for the overall composite or total score. The total score demonstrated adequate levels of variability and was a useful measure for criterion-related validity analyses. This increased variability with the composite may be attributed to TSA's prescribed weighting scheme and scoring system (TSA, 2010).

Training Evaluations

In addition to evaluating job performance, it is often useful to collect ratings of training performance. These two criteria, although related, often show different relationships with selection measures (e.g., McDaniel, Whetzel, Schmidt, & Maurer, 1994; Schmidt & Hunter, 1998). In addition to providing a unique measure of performance, BDO training evaluation ratings would have been useful because these ratings are used to remove low performing job candidates during the hiring process. Specifically, job candidates who are unable to pass the initial SPOT training course are denied entry into the program and typically return to their position as a TSO. However, despite the potential benefits of these scores, these criterion data were not available for the current study.

Developed Performance Measure

AIR developed a new criterion measure (job performance) for several reasons. First, existing performance measures (i.e., PASS) were used for administrative decisions (e.g., determining pay, promotions) and, like other operational measures, are likely to show limited variability (Landy & Farr, 1980). Second, because PASS was created for the purpose of performance appraisal, it may not fully capture the job performance information needed to provide validity evidence for a selection system. Third, alternate measures of job performance (e.g., training evaluations) were not available for use in this study.

Because of these limitations, AIR undertook a criterion development process. This process was based on a comprehensive job analysis of the BDO job (AIR, 2010a) and a review of the research literature. The process was designed specifically to capture the information needed to collect validity evidence for the current study.

The following sections describe the criterion development process. First, this section describes the results of a literature review that was conducted to determine the optimal type of performance measure that should be created. Second, this section provides a description of the process that was used to create performance dimensions for the BDO job. Third, the procedures that were used to create behavioral anchors for the BDO JPM are described. The section concludes with a description of how these changes were finalized and made into a final product for distribution to users.

Literature Review

A review of the job performance literature was first undertaken to identify the most appropriate way to conceptualize job performance. Current conceptualizations of job performance emphasize that performance is a multidimensional construct (e.g., Borman & Motowidlo, 1997). An implication of this conceptualization is that there are multiple dimensions of job performance which must be measured in order to accurately reflect performance – single, holistic ratings do not capture the full and complete nature of performance. As a result, any measure of job performance must be multidimensional to best reflect the true nature of performance.

Second, AIR undertook a review of the job performance measures literature to identify the most appropriate way to measure job performance. Based on this literature review, as well as the literature review described above, AIR identified three measures of job performance that were considered for this study – Graphic Rating Scales (GRS), Behaviorally Anchored Rating Scales (BARS), and Behavioral Summary Scales (BSS).

Graphic rating scales are the most commonly used type of subjective rating scales. A primary benefit of these scales is that they are easy to use and familiar to raters, but are also prone to rater errors due the ambiguity of their scale anchors (Gatewood, Feild, & Barrick, 2008). This limitation is addressed by BARS because these scale anchors describe actual job behavior and, thus, are less ambiguous (Gatewood et al., 2008). Additionally, BARS are usually based on actual instances of behavior (often collected through the critical incidents technique) and are

developed with significant input and contributions from SMEs (Schwab, Heneman, & DeCotiis, 1975). However, BARS describe only specific behaviors (Schwab et al., 1975) and require an extensive retranslation process (Smith & Kendall, 1963).[21] According to (Guion, 1997), BSS have several advantages over both graphic rating scales and BARS. First, they reduce errors because scale points and anchors are more clearly defined than with other measures such as BARS. Second, they provide summaries of behavior rather than specific examples of behavior like a BARS. And last, because there is not a need for a retranslation of scale anchors, the development process for BSS is more feasible.

AIR ultimately decided to utilize a BSS for this study. This decision was based on the advantages and disadvantages of each scale, the nature of the BDO job, and available resources. The sections that follow describe the scale development process.

Development of Performance Dimensions

An iterative process was taken to develop the performance dimensions which comprise the BDO JPM. First, AIR analysts independently created an initial set of performance dimensions. Second, these performance dimensions were revised through consensus building meetings among research team members. Third, the performance dimensions were reviewed and edited by a panel of BDOs. The following sections describe this process.

Initial development of the performance translations was undertaken using a procedure that has elements in common with retranslation procedures (e.g., Schwab, Heneman, & DeCotiis, 1975; Smith & Kendall, 1963) and Q-sort methodology. This procedure involved having research team members sort task statements into higher order groups based on the underlying behavior needed to complete each task. Task statements were used because they refer to specific instances of behavior (e.g., Review an individual's ID, travel documents, and other paperwork, as needed, to check for fraudulent documents). For this procedure, all 110 task statements and 9 skills from the BDO job analysis were used. Because this task was based on information from the job analysis, it is reflective of a systematic data collection that was undertaken to identify all of the relevant tasks and KSAOs for the BDO job.

[21] During the retranslation process, SMEs typically generate examples of job behaviors and assign these examples to categories of job performance. Examples of behaviors that are not consistently assigned to the same categories by separate SMEs are discarded.

A total of seven AIR experts participated in this task. Each of these experts had extensive knowledge about the BDO job, knowledge and expertise in the field of industrial and organizational psychology, and most of these experts had knowledge about the BDO job through experience conducting a job analysis of the BDO position. To complete this task, each expert was given a standardized set of instructions and asked to independently sort the 110 tasks and 9 skills into groups based on the underlying behavior reflected by each task statement or skill. In other words, the job analysts created groups based on a consideration of the set of behaviors necessary for completing each task statement or utilizing each skill. Behavior was used as the basis for this sorting task because job performance reflects job relevant behavior rather than the outcomes of behavior (e.g., the completion of a task).

Participants in this procedure were not given a set number of groups to sorts their tasks into, and were asked to revise their groupings until they thought that the same (or very similar) behavior was represented in each group. The groupings that were created by each job analyst was recorded and compiled into a single document. At this stage, some task statements or skills were immediately removed from further consideration because they reflected primarily supervisory tasks, were not specific, or were multifaceted. Additionally, all skills were removed from further consideration because: (1) SMEs were unsure how to group them, (2) there was a lack of agreement regarding how they should be categorized, (3) they did not fit well with the groupings that were made (and which consisted primarily of task statements), and (4) it was determined that they did not reflect important components of the job beyond the tasks statements and, therefore, their removal would not result in a loss of important information.

To further refine the performance dimensions, consensus building groups comprised of AIR job analysts were convened. During these meetings the job analysts met to review each performance dimension, and specifically to review the tasks and associated behaviors that were sorted into each performance dimension. An iterative process was used for these revisions, and performance dimensions were refined based on discussion and consensus about each change.

Once the performance dimensions of the BDO JPM had been finalized through consensus building meetings, definitions of each performance dimension were drafted. These definitions were based only on the task statements that were sorted into each performance dimension. More specifically, the dimension definitions reflect a summary of the behaviors needed to complete the group of task statements which comprise each performance dimension.

To finalize the performance dimensions and their associated definition, a group of nine BDOs served as SMEs in a workshop that was held on December 1, 2010. This group of BDOs reviewed the performance dimensions and their associated definitions. This review was conducted remotely, using telephone and Microsoft Live® web conferencing software, with BDOs from multiple airports around the nation. All participants were STSMs (n = 8) and had worked for TSA for an average of 7.7 years (SD = .71 years). Six of the nine SMEs had received additional training in behavioral indicators of deception, and seven of the nine SMEs were male. About half of the SMEs identified themselves as White.

The BDOs who participated in this workshop were first familiarized with the overall research effort, the performance scale, its intended use, the process that was undertaken to develop the measure, and their role in the development process (i.e., to review the dimensions to ensure these categories adequately represented the BDO job). SMEs were then presented with the performance dimensions and their associated definitions, as well as a list of the task statements that were reflected by each performance dimension. Although a list of task statements was presented to SMEs during this meeting, it was emphasized that the performance dimensions and their associated definitions were intended to represent behavior. Changes to the scale were made during the workshop – in real time – so that SMEs could immediately review each change and suggest additional changes if needed.

This workshop had two major outcomes. First, SMEs provided suggestions for better aligning the BDO performance dimensions and their associated definitions with current BDO phrases and terminology. Second, no major edits were necessary for either the performance dimensions or their definitions, which suggests that the initial efforts by AIR were consistent with how job incumbents view their work. Based on these results, AIR then began creating scale anchors, which is described below.

Development of Behaviorally Based Scale Anchors

Behavioral anchors were then developed for the performance scale. Behavioral anchors are specific examples of behavior which reflect a particular level of job performance. Unique behavioral anchors were developed for each performance dimension. These behavioral anchors were developed to assist raters when making performance ratings – that is, these behavioral anchors were intended to reduce rating errors and biases and, more generally, to improve the validity and reliability of performance ratings.

Multiple anchors (between two and five) were written for each of the seven scale points and nine performance dimensions. These behavioral anchors were written based on the performance dimensions and their associated definitions – that is, scale anchors operationalized different levels of performance that were specific to each performance dimension. Behavioral anchors were also developed to maintain consistency across rows and within columns. In other words, multiple levels of performance for a specific behavior were reflected across columns, while multiple behaviors were reflected for a specific performance level within rows. These behavioral anchors were reviewed extensively within AIR.

A group of ten BDOs, serving as SMEs, reviewed these performance anchors in light of the performance dimensions which had previously been created and reviewed. This review was held on December 16, 2010, and was conducted remotely using telephone and web conferencing software. All participants were STSMs (n = 10) and had worked for TSA for an average of 8.0 years (SD = .67 years). Six of the ten SMEs had received additional training in behavioral indicators of deception, and seven of the ten SMEs were male. About half of the SMEs identified themselves as White. All of the BDOs participating in this workshop had also participated in another workshop convened to work on the performance dimensions and definitions. Because of this prior experience, all SMEs were familiar with the performance scale, its intended use, the process that was undertaken to develop to scale, and their role in the development process (i.e., to review scale and make suggestions to improve the accuracy and usability of the measure).

At the beginning of the workshop, SMEs were provided a brief explanation concerning how the performance anchors would be used for making performance ratings. SMEs were asked to offer suggestions on how to align the wording and phrasing of these performance anchors with current and correct usage of relevant BDO terminology. In addition to suggested changes to the wording and terminology of the anchors, the SMEs also reviewed the placement of the performance anchors under each rating point. That is, they reviewed the performance anchors to ensure that each anchor reflected the appropriate level of performance. Changes to the scale were made during the workshop – in real time – so that SMEs could immediately review each change and suggest additional changes if needed.

During this workshop SMEs suggested several more substantive changes to the scale. First, SMEs suggested that the scale should be modified from a seven point scale to a five point scale. As a rationale for this change, SMEs noted that it was difficult to make distinctions

between different levels of performance. For example, SMEs found it difficult to differentiate between performance that would be rated as a —4‖ on the scale and performance that would be rated as a
—5.‖ As a result, the performance anchors were revised by removing two scale points from the scale to create a five point scale.

Second, SMEs noted that the provision of behavioral anchors at every scale point could be overwhelming for some users and recommended that some behavioral anchors be removed. To avoid losing important information, AIR incorporated this feedback by combining some behavioral anchors rather than removing them entirely from the scale.

Last, based on comments from SMEs and internal discussions at AIR, a not applicable (—N/A‖) rating option was added to the scale. This rating option was added primarily because some of the BDOs who would be invited to participate in the study may have been recently hired. As such, at the time supervisors were asked to evaluate BDO performance, some recent hires may not yet have had the opportunity to perform some of the behaviors reflected by the BDO JPM (e.g., referral of passengers). Adding this option was intended to prevent raters from providing inaccurate ratings on performance dimensions for which they had insufficient or no information.

Final BDO Job Performance Measure

The finalized BDO JPM was reviewed in a final workshop with five BDOs who had previously participated in the reviews of the performance dimensions and behavioral anchors. This workshop was also held remotely, using telephone and web conferencing software, on January 20, 2011. Each of the workshop participants were STSMs (n = 5) and had worked for TSA for an average of 8 years (SD = .00 years). Three of the five SMEs had received additional training in behavioral indicators of deception, and four of the five SMEs were male. Almost all of the SMEs identified themselves as White.

SMEs were presented with the finalized BDO JPM, and were asked to suggest any additional edits that should be made. As with prior workshops, changes were made to the scale in real time so that SMEs could immediately see and comment on the changes. The SMEs who participated in this review suggested only minor wording changes and were supportive of the revised scale format (i.e., a five point scale which utilized three sets of behavioral anchors).

Following the workshop, a glossary of terms was then added to the finalized measure to define any terms which might be unfamiliar to users. Additionally, a set of instructions was created to explain how to use the BDO JPM and record performance ratings. These instructions emphasized the fact that all ratings would remain confidential, that ratings were being collected only for research (rather than administrative) purposes, and that raters should consider ratees' performance over the prior three months when making performance ratings. Additionally, raters were instructed to use the —N/A‖ response option if they were unable to provide a rating for some dimensions (i.e., because of an insufficient opportunity to observe a specific aspect of performance). Last, a rating sheet was created in both Microsoft Word® and Excel® format which provided a way for raters to record their performance ratings and return them directly to AIR. Instructions were provided – both in the scale instructions and on the rating sheet – which described how completed rating sheets could be returned (via email or fax). Together, these documents (instructions, scale, glossary, and rating sheet) formed the BDO JPM found in Appendix H.

Opportunity to Observe Measure

In addition to the performance measure, AIR developed three items to measure STSM's opportunity to observe each ratee perform specific duties. These duties included performing and conducting SPOT, writing reports, and demonstrating knowledge of the SPOT standard operating procedure. Ratings could be provided to each item using a five-point scale (1=*almost never had the opportunity to observe* to 5 = *very frequently had the opportunity to observe*).

CHAPTER IV: PILOT TEST

After procuring and/or developing the selection and job performance measures, the AIR pilot tested these instruments over the course of six weeks beginning in February 2010. This chapter describes the purpose, participants, method, and results of the pilot test. Subsequent changes to the measures based on the pilot test results are described at the end of this chapter.

Pilot Test Purpose

The pilot test allowed AIR to examine the proposed data collection process and the degree to which the measures were functioning as planned (American Educational Research Association [AERA], American Psychological Association, and National Council on Measurement in Education, 1999). For example, AIR was able to examine the efficiency of the data collection process, estimate the administration times for each selection measure, and ensure the instructions were clear. In addition, the pilot test provided preliminary data for examining the statistical and psychometric quality of the measures, as well as qualitative feedback on their job-relevance and perceived utility. These data were used to inform revisions to the measures prior to the validation study.

Pilot Test Participants

Prior to analyzing participant data, AIR removed the scores from the Structured Interview and Work Sample Test data for one participant because TSA representatives had been present during the testing sessions.[33] Further data cleaning efforts included looking for individuals with extreme (i.e., outlying) scores on one or more of the measures. Due to the small sample size and because the measures would likely be modified after the pilot test, AIR chose not to remove those individuals from the dataset.

The final pilot test sample included 60 BDOs from five airports (Table 8). Participants were relatively experienced BDOs, with an average tenure of 2.74 years in their current position (SD = 0.86) and 82% of the sample representing Master BDOs (F-Band). Participants were primarily male (73%), non-Hispanic or Latino (85%), and White (73%). Also, participants in the pilot test were generally well educated, with 83% of the sample having at least some college experience. Finally, on average, BDOs who participated in the pilot test were 38.49 years old (SD = 10.53).

[33] A description of the purpose and process for the TSA observation is described in the Method section below.

Sensitive Security Information (SSI)

Table 8. Pilot Test Participant Characteristics

	N	Percentage
Airport Code – Location		
BWI – Baltimore, MD	9	15%
BOS – Boston, MA	16	27%
MCO – Orlando, FL	11	18%
PVD – Providence, RI	13	22%
SLC – Salt Lake City, UT	11	18%
Job Title		
Expert BDO	11	18%
Master BDO	49	82%
Gender		
Female	16	27%
Male	44	73%
Education Level		
High School/GED	10	17%
Some College	19	32%
Associate's Degree	13	22%
Bachelor's Degree	9	15%
Master's Degree	2	3%
Ethnicity		
Hispanic or Latino	10	10%
Not Hispanic or Latino	51	85%
Other	1	2%

Sensitive Security Information (SSI)

Race		
Black of African American	9	15%
Native Hawaiian or Pacific Islander	3	5%
Two or More Races	2	3%
White	44	73%
Other	2	3%
	Mean	**Std. Dev.**
Tenure (Years)		
In Current Position	2.74	0.86
With TSA	6.82	1.93
Age	38.49	10.53

Note. —Other‖ includes participants who indicated they were American/Chilean or Samoan.

Pilot Test Method

This section describes the procedures used for collecting data during the pilot test, including presite visit preparation, the on-site test administration process, as well as the process used to collect the criterion (job performance) data.

Pre-Site Visit Preparation

This section describes the activities AIR conducted prior to each site visit. These included identifying pilot test sites, conducting trainings, and assigning test forms to participants. Each of these activities is described in more detail below.

Site Selection Process

AIR worked closely with TSA's Office of Security Operations (OSO) to identify and communicate with potential pilot test sites, which were selected to represent different geographic locations and airport sizes (i.e., medium and large airports). As a result of the coordination efforts between AIR, TSA's OSO, and the participating airports, site visits were scheduled and conducted over six weeks beginning in late February and concluding the first week of April (see Table 9).

Table 9. Pilot Test Sites

Airport Code – Location	Airport Size	Date of Visit
BWI – Baltimore, MD	Large	March 23 – 24, 2011
BOS – Boston, MA	Large	March 29 – April 1, 2011

MCO – Orlando, FL	Large	March 4 – 6, 2011
PVD – Providence, RI	Medium	February 22 – 24, 2011
SLC – Salt Lake City, UT	Large	March 14 – 17, 2011

Prior to each site visit, AIR conducted teleconferences with representatives from the participating airports using a semi-structured protocol (see Appendix I). During the teleconferences, AIR provided background information about the study, described the purpose of the pilot test, outlined scheduling and room/equipment requirements, and provided criteria for recruiting study participants.

With respect to study participants, to ensure the quality of the data, the study required 10-12 BDOs per site who were willing participants. Preferably, these BDOs would have less than 6 months of experience and represent a range of performance (i.e., both high and low performers). This would create a sample that most closely resembled that of the job candidate pool. In addition, it was important to have the participation of 2 SPOT Transportation Security Managers (STSMs) per site. AIR asked for STSMs who had the opportunity to observe BDOs' performance, could participate in a rater training session on the use of the BDO Job Performance Measure (JPM), and could provide independent performance ratings for BDOs participating in the study. In addition, the airports designated points-of-contact (POCs) to lead the recruitment effort at their airports and correspond directly with AIR.

Following the initial teleconferences, AIR provided each site with read-ahead materials for the POCs to disseminate to potential participants during the recruitment process (see Appendix J). These materials described the study and participants' role in the pilot test. In turn, each airport provided a list of BDO volunteers and information about their work schedules. AIR used this information to create draft agendas to ensure the most efficient use of BDO and BDO manager time and limit disruption of security operations and the critical mission of passenger screening. Then, AIR and the airport POCs coordinated efforts to finalize the agendas before the data collection. This process resulted in two-, three-, or four-day site visits.

Training

Prior to the pilot test, AIR conducted a series of internal and external training sessions, including Administrator training and Resource Person training. Each is described briefly below.

Administrator Training

First, AIR completed internal training sessions during which Administrators practiced using the protocols and administration guides for each selection measure. These sessions were conducted to familiarize members of the research team with the assessment materials to ensure accurate test administration during the site visits. During each session, AIR researchers practiced

administering each selection measure. These practice sessions included reading each script aloud, distributing test materials, and operating the video-based equipment (e.g., Passenger Observation Assessment video recordings).

Resource Person Training

Second, AIR trained two members of the BDO National Deployment (NDO) Team to serve as the Resource Person for the Role-Play Exercise. These BDOs were chosen to serve as the Resource Person for two reasons. First, as members of the NDO Team, these BDOs were intimately familiar with the standard operating procedures for engaging passengers to elicit information. This experience would allow the trainees to better understand the Role-Play Exercise and the role of the Resource Person. Second, these BDOs were identified as capable of providing a consistent performance across administrations.

The three-hour training session was conducted at AIR's office in Washington, DC on February 17, 2011. Two AIR research team members facilitated the training using a semi-structured protocol (see Appendix K). Representatives from TSA's OSO also attended. To begin the session, the Resource Persons completed a background questionnaire, which provided information necessary to document the expertise of the participants. Both Resource Persons were experienced BDOs. On average, they had worked for TSA for 7.21 years (SD = 1.47) and had been in their current position as a BDO for 5.04 years (SD = 2.30).

After providing demographic information, the Resource Persons received an overview of the Role-Play Exercise, including its purpose and structure. Next, the instructions for the Role-Play Exercise were reviewed, including specific instructions about how to perform as a Resource Person. For example, the Resource Persons were told that they would be playing a prescribed, fictitious passenger and that, in order to standardize the administration of the assessment across all participants, it would be important for them to remain in character throughout each administration.

Other instructions included:

- Speak in a clear and audible manner;
- Avoid providing too much or too little information in response to a job candidate's question;
- Remain in character throughout the entire conversation;
- Act as natural as possible throughout this conversation (e.g., avoid being overly dramatic or theatrical);

- Try to limit your talking so the participant has enough time to ask his or her questions; and
- Remain impartial and unbiased to all participants throughout the process.

As discussed in Chapter III, two different scripts were developed for the Role-Play Exercise. The Resource Persons were given background information on these characters and were asked to review and study it to prepare for their role as Resource Person. Nonetheless, Resource Persons were instructed that they did not need to commit the information to memory because they would receive a one-page summary of the characters' information that they could refer to during the training and pilot test.

The training concluded with two types of practice exercises. First, Resource Persons reviewed previously recorded examples of effective and ineffective performance. After each video, the trainers highlighted key learning points (e.g., providing appropriate amounts of information in response to a participant's question). Second, trainees practiced role-plays by assuming the role of the Resources Person. After each practice session, they were provided feedback about their performance.

Assigning Test Forms

As mentioned in Chapter III, AIR developed two to four separate forms for each of the newly developed selection measures. Prior to each site visit, AIR assigned each pilot test participant a specific form of each assessment to complete. This allowed AIR to standardize the form assignment process prior to each visit and to obtain similar sample sizes for each form of each assessment, as shown in Table 10.

Table 10. Pilot Test Participant Form Assignment

Selection Measure	Form (Version)	Number of BDOs Who Completed Each Form
Passenger Observation	A (1)	13
	A (2)	14
	B (1)	16
	B (2)	15
Mental Math	A	30
	B	28
Writing Knowledge	A	30
	B	28
Visual Recall	A	20
	B	20
	C	18

Sensitive Security Information (SSI)

Structured Interview	A	15
	B	12
	C	14
	D	14
Role-Play Exercise	A	30
	B	25

When assigning forms, AIR worked to counterbalance the combination of forms. That is, although the assessments were always completed in the same order, various combinations of forms were implemented. For example, some BDOs completed the same form across the selection measures (e.g., all Form As), but most completed different forms for each measure (e.g., Form A – Version 1 for the Passenger Observation Assessment, Form B for the Mental Math Assessment, Form A Writing Knowledge Assessment, Form B Visual Recall). By varying the presentation of the combinations of forms, AIR ensured that the forms were distributed across airports and shifts. This limited the effects of a location bias, which could influence the results of the pilot test.

On-Site Administration

Best practices for test administration dictate the use of standardized communications (e.g., instructions) between Administrators and job candidates (AERA, 1999). A set of standardized instructions helps ensure the accuracy of each administration, the consistency across participants and sites, and the comparability of test scores across participants. As a result, AIR standardized the test administration processes at each site visit by using administration guides for each assessment and by following a semi-structured pilot test protocol. Below, these materials are briefly discussed followed by a description of the administration process.

Pilot Test Protocol

The Pilot Test Protocol (see Appendix L) contained information on how to prepare for and administer each phase of the pilot test. This document described the equipment, assessment materials (e.g., user's manuals, response booklets, media files), and other materials (e.g., on-site agenda) needed to conduct the pilot test. It also included scripts for conducting the group introduction session, administering the background questionnaires, proctoring the Problem Solving measures, and facilitating the group feedback sessions. Finally, it included a list of frequently asked questions (FAQs) (e.g., purpose of the pilot test, intended use of the test scores).

Assessment Administration Guides

An administration guide was developed for each selection measure. Each guide is described below.

Work Sample Test

The Work Sample Test Administration Guide contained specific information required for administering the Passenger Observation, Mental Math, Writing Knowledge, and Visual Recall Assessments. The guide began by describing general instructions for proctoring the Work Sample Test, such as a list of necessary materials, the role of the Administrator, test security information, how to handle cheating or dishonest behavior, and establishing a proper testing environment. Similar to the Pilot Test Protocol, the Work Sample Test Administration Guide provided the Administrators with a standardized script to follow while administering these four assessments. The scripts were provided to enhance the consistency of each testing session. The guide also included a list of FAQs that participants may ask. For example, some of these questions addressed how BDOs should properly change responses to items to receive credit (e.g., striking out the original answer with an —XI) and how test administrations would protect confidential information (e.g., name, airport location).

Structured Interview and Role-Play Exercise

The Structured Interview and the Role-Play Exercise were also administered using a standardized administration guide, which provided Administrators with information on how to prepare for and administer these assessments. The guide listed each of the materials required and provided instructions for preparing a proper testing environment. It also included instructions for preparing additional personnel (i.e., Assessors and Resource Persons). These particular instructions were important because these assessments require, at a minimum, two Assessors and a trained Resource Person.

The guide also included standardized scripts for administering both assessments, the Structured Interview items, the Role-Play Exercise questions, notes pages, rating scales, and instructions for how to evaluate participants' performance. Information provided to each test-taker at the beginning of the session (labeled as the Candidate Packet) is provided at the end of the guide. This packet allowed participants to follow along by reading the instructions while the Administrator read them aloud. To ensure test security, candidate packets were collected immediately after the session had concluded.

Modifications to Administration Guides

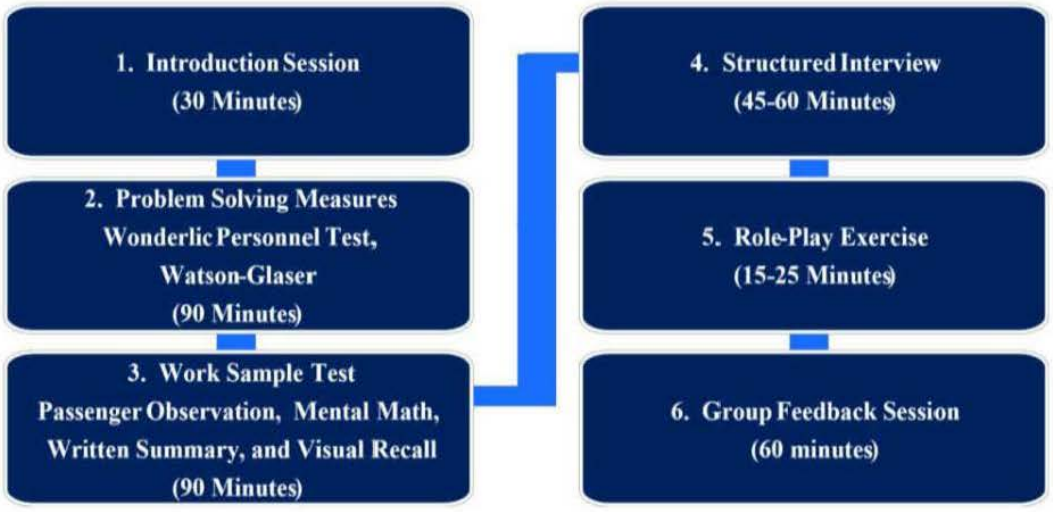
During the first two site visits, AIR made minor adjustments to the scripts in the Pilot Test Protocol, the Work Sample Test Administration Guide, and the Structured Interview and RolePlay Administration Guide. These changes primarily consisted of minor modifications to the script to improve the dialogue between the Administrators and participants. In addition, the instructions for the Role-Play Exercise were significantly modified after the first site visit. The instructions for this exercise originally informed test-takers that they should not ask the assigned questions directly. This proved to be too difficult for many participants because they were then unduly preoccupied with identifying alternative ways to phrase the questions. Furthermore, this instruction appeared to inhibit the natural flow of the conversation during the Role-Play

Exercise. AIR removed this language and instructed participants to naturally embed the questions within the conversation.

On-Site Administration Process

One day prior to data collection, AIR met with the site POC to become familiar with the facilities, set up the testing equipment, and review the agenda. Then, the data collection occurred over two to four days depending on the number of BDOs, the start and end times for each day, and other important considerations. Figure 4 shows the order in which data collection activities were conducted at each site and the amount of time typically required for each activity. Following the exhibit, these activities are briefly described.

Figure 4. On-site Administration Process



Introduction Session

Each site visit began with an introduction session with large groups of BDOs (e.g., 7-10) and STSMs. Separate introduction sessions were typically conducted for each shift (i.e., AM and PM shifts). During this session, AIR informed BDOs about the purpose of the study, their role in the study, and the agenda for the site visit. AIR also reviewed the processes for ensuring the confidentiality of participants' responses and the importance of maintaining the security of the test materials. Before concluding, AIR reminded the participants about the voluntary nature of the study, provided an opportunity to withdraw from the pilot test, and answered questions.

Problem Solving Measures

Administration of the Problem Solving measures was completed by one AIR Administrator. STSMs, other TSA representatives, and other AIR personnel were dismissed from the room.[25] First, BDOs completed the Wonderlic Personnel Test. Following the test publisher's administration procedure, BDOs were given 12 minutes to complete 50 items. Next, participants were asked to complete the Watson-Glaser Critical Thinking Appraisal. For this measure, BDOs were given 30 minutes to answer 40 items. At the conclusion of the session, AIR reviewed the agenda and reminded participants of their next scheduled sessions.

Work Sample Test

The Work Sample Test was administered by one AIR researcher. Each of the four assessments in the Work Sample Test was administered in the following order: the Passenger Observation Assessment, the Mental Math Assessment, the Writing Knowledge Assessment, and the Visual

[25] At this time, AIR conducted a demonstration of the Work Sample Test and Role-Play Exercise for STSMs and other TSA representatives. During these demonstrations, STSMs and TSA representatives provided general reactions about the clarity and job-relevance of the assessments.

Recall Assessment. For the first three site visits, the Work Sample Test was administered to participants individually. For the last two sites, however, AIR pilot tested administering the Work Sample Tests to groups of three to four BDOs. Because this was the proposed method of administration for these assessments, it was important to determine if this would be an effective approach. This process proved to be much more efficient and significantly reduced the amount of administration time required for each site visit. Furthermore, BDOs commented that the group administration did not have any detrimental effect on their performance.

Structured Interview and Role-Play Exercise

The Structured Interview and Role-Play Exercise were administered to BDOs individually. Prior to beginning this assessment, AIR requested permission to record the session. The audio recordings were collected to inform subsequent changes and revisions to these assessments. For example, AIR reviewed the Role-Play Exercise performances to determine if the content of the Resource Person's script needed to be modified. All efforts were made to ensure confidentiality during the session (e.g., by not referring to the BDO by name).

The Structured Interview and Role-Play Exercise required two to three Assessors. During the first four site visits, TSA personnel (i.e., STSMs [$n = 5$] and/or human resource (HR) representatives [$n = 2$]) served as Assessors for these sessions. Because government personnel would be administering the operational versions of these assessments, this approach increased the realism of the administration process. During subsequent site visits, BDOs were given the option to have AIR serve as Assessors. This option was provided for two reasons. First, STSMs were not always available to serve as Assessors during each site visit. Second, some BDOs indicated that the presence of their STSMs led to heightened test anxiety. Thus, AIR sought to reduce these adverse reactions and, consequently, ensure the quality of the data.

Participant Feedback

Following each assessment, BDOs completed the feedback questionnaire described in Chapter III. Questionnaires typically took 5-10 minutes to complete. In addition, AIR invited BDOs to participate in group feedback sessions during the last day of the site visits. Sessions typically included four to eight BDOs. During these group feedback sessions, AIR used a semi-structured

protocol (see Appendix L) to facilitate a discussion regarding the difficulty of the measures, the clarity of the instructions, and recommendations for improving the measures. These sessions were designed to gather qualitative feedback that would supplement the information captured by the feedback questionnaires. For example, these sessions provided BDOs with an opportunity to discuss the assessments with their colleagues as well as AIR. Each approach to collecting feedback data (i.e., the questionnaire and group sessions) provided unique information. The feedback questionnaires allowed BDOs to provide their individual reactions, whereas the feedback sessions provided a forum for participants to share and expand on ideas.

AIR used audio recorders to document the group feedback sessions. This ensured that specific recommendations were not lost during the note taking process, and they served as a resource when implementing changes to the assessment and administration guide following the pilot test.
Permission was obtained from each of the participants prior to beginning the session and the Administrators followed specific instructions (e.g., by not referring to the BDOs by name during the session) to maintain their confidentiality.

TSA Observation

During the first site visit, representatives from OSO completed the two Problem Solving measures, observed a demonstration of the Work Sample Test, and attended a Structured Interview/Role-Play Exercise session. These activities allowed OSO to provide feedback on the selection measures and administration processes. For example, OSO representatives were able to assess the fidelity of the instruments and determine whether the administration processes were similar to TSA's existing selection system. As mentioned earlier, because OSO representatives may have impacted the BDO's performance during the Structured Interview/Role-Play Exercise session, these ratings were removed from the dataset.

Protocol for Collecting Job Performance Data

This section describes the process AIR used to collect criterion (job performance) data. These data consisted of ratings from TSA's existing Performance Accountability and Standards System (PASS) and the BDO JPM.

PASS Data

AIR received PASS data from TSA for each BDO who participated in the pilot study. These data were from the previous performance period (i.e., 2009-2010) and consisted of scores on both the Performance Skills Evaluation and Job Knowledge Test. Composite competency

ratings, consisting of the sum of individual ratings for each competency (e.g., Interpersonal Skills, Teamwork), were also provided.

BDO Job Performance Ratings

This section describes the process used to collect job performance ratings using the BDO Job Performance Measure. As detailed below, managers received training on how to use the measure and were then provided instructions for returning the completed ratings directly to AIR.

Rater Training Program

Before providing performance ratings, all STSMs completed a rater training session. AIR developed the rater training program on the basis of best practices outlined in the professional literature (e.g., Bernardin & Buckley, 1981; Woehr & Huffcutt, 1994) for developing and implementing training on how to complete valid and reliable performance appraisals. In particular, optimal training includes rater error training and frame-of-reference training.

Rater error training focuses on minimizing the impact of rater errors and biases by describing their occurrence (Woehr & Huffcutt, 1994). Rater errors and biases are natural ways of thinking that can limit the effectiveness of performance ratings (e.g., Landy & Farr, 1980; Woehr &

Huffcutt, 1994).[26] Rater error training also provides information about the ideal distribution of performance ratings (i.e., a range of scores with equal representation in each performance category) and describes strategies for trainees to use to avoid these rating errors and biases.

Frame-of-reference training focuses on describing to trainees the multidimensional nature of job performance (Bernardin & Buckley, 1981). Multidimensional job performance refers to the idea that job performance consists of several separate, but related, elements of performance. For example, BDOs are required to observe passenger behaviors, engage passengers in conversation, review travel documents, search accessible property, and communicate with other security personnel. Another focus of frame-of-reference training is to calibrate performance ratings both within and across raters. This is typically accomplished by having raters practice completing ratings using a prototypical example, reviewing their ratings, and discussing discrepancies.

In keeping with the literature, AIR's training included elements of rater error training and frameof-reference training. These two approaches were combined to maximize improvements in rater accuracy and reliability. This hybrid training program consisted of:

- providing information about the multidimensional nature of job performance,
- reviewing information on rater errors and biases and strategies for avoiding them, and
- calibrating performance ratings within and across raters (through frame-of-reference training).

The training had five overarching goals, including:

- familiarizing STSMs (raters) with the purpose of the pilot test,
- providing an overview of the development of the BDO JPM,
- improving the interrater and intrarater reliability of the STSMs' performance ratings, •
  answering questions from raters about the pilot test or their rating task, and
- emphasizing the importance of accurate ratings for research purposes.

Prior to each training session, AIR sent the STSMs a copy of the performance measure and instruction guide. In addition, several training aids were sent to the STSMs prior to the rater training. These training aids included:

- a Microsoft PowerPoint® presentation to guide the STSMs through the training;

[26] Rating error and biases have been extensively discussed in the literature (e.g., Landy & Farr, 1980; Woehr & Huffcutt, 1994). Examples of rater errors include leniency/severity (i.e., providing consistently high or low ratings) and central tendency (i.e., consistently providing moderate ratings). An example of a rater bias is the similar-to-me bias (i.e., providing higher ratings to employees who share similar characteristics with the rater).

- a handout of common rater errors and biases, with definitions and strategies for avoiding them; and
- a set of vignettes that described a hypothetical instance of BDO job performance for use in calibrating performance ratings.

Rater trainings were held remotely, via telephone and web conferencing, and typically lasted two hours. Table 11 presents the dates of each training session. Each training session was facilitated by one or two trainers who followed a semi-structured protocol (displayed in Appendix M).

Table 11. BDO JPM Rater Training Dates

Dates of Trainings
February 16, 2011
February 18, 2011
February 21, 2011
February 25, 2011
March 2, 2011
March 11, 2011

The training began with an overview of the research effort, the purpose of the pilot test, the role of BDOs and STSMs in the pilot test, and the importance of their participation. During the training, the facilitators explained that all performance ratings made by STSMs as part of the pilot test would be used for research purposes only and would not affect administrative decisions (e.g., to determine promotions and pay raises). Following best practices for providing training,

AIR described the processes for ensuring the confidentiality of STSMs' performance ratings. Prior research suggests that ratings are less inflated (i.e., positively biased or negatively skewed) when ratings are confidential and made for research (rather than administrative) purposes (Landy & Farr, 1980).

The training session continued with an overview of the BDO JPM, a discussion of the multidimensionality of job performance, and common rater errors and biases. The discussion concerning the multidimensionality of job performance was intended to help the STSMs understand that job performance is complex and usually cannot be evaluated using a single overall rating. Additionally, by describing job performance as a complex construct, AIR emphasized the point that the performance dimensions represent qualitatively different sets of behaviors.

The training session also provided an opportunity for the raters to practice using the measure and allowed them to calibrate their performance ratings using a frame-of-reference exercise. First, the STSMs read nine vignettes (one for each performance dimension). The vignettes described a specific level of BDO performance on a specific performance dimension. Three vignettes were written to reflect high performance, three were written to reflect adequate performance, and the remaining three vignettes were written to reflect poor performance. STSMs then individually provided ratings for each of the fictional BDO's performance on a particular dimension, completing the same process for each of the nine vignettes. Next, the STSMs reviewed their ratings as a group. The trainer focused on any discrepancy between the STSM ratings and the actual performance described in each vignette and provided recommendations for conceptualizing performance using the BDO JPM.

During the final portion of the training, AIR explained the process for completing and returning the performance ratings. Two STSMs from each airport were asked to independently rate BDO participants' performance in terms of how well they performed and conducted SPOT, wrote reports, and demonstrated knowledge of the SPOT standard operating procedure. STSMs were instructed to base their performance ratings on the last three months of performance. In addition to providing performance ratings, raters were asked to complete three items that assessed their opportunity to observe each ratee perform specific duties using a five-point scale (1=*almost never had the opportunity to observe* to 5 = *very frequently had the opportunity to observe*). Then, STSMs were instructed to record the scores within an encrypted Microsoft Excel® spreadsheet and returned it directly to AIR via email.

At the conclusion of the session, the trainees completed a training evaluation form. This form collected information about trainees' reactions to the training (e.g., —this training session enhanced my performance rating skillsl) and solicited suggestions for improving the training. The ratings were completed in a separate file and returned to an AIR team member who did not facilitate the rater trainings.

Pilot Test Results

This section of the validation report describes the results of the pilot test. Data analyses were conducted to determine which measures were functioning appropriately and to identify the revisions that needed to be made. The types of analyses conducted include: descriptive and inferential analyses, reliability and agreement analyses, and item-level analyses. Brief descriptions for each of these analyses are provided below.

Descriptive and Inferential Analyses

AIR conducted descriptive analyses to examine the distribution of scores for each measure. These descriptive analyses included:

- ***N*** – Number of BDOs who are included in a specific analysis.
- **Mean** – Arithmetic average.
- **Median** – Midpoint of a frequency distribution of scores.
- **Standard Deviation** – Extent to which individual scores, on average, vary from the average score of a distribution.
- **Range (Possible)** – Lowest and highest possible scores for a measure.
- **Minimum** – Lowest observed value for a distribution of scores.
- **Maximum** – Highest observed value for a distribution of scores.
- **Skewness** – Measure of symmetry for a distribution of scores. Positively skewed distributions have few values in the upper end of a distribution. Negatively skewed distributions have few values in the lower end of a distribution.
- **Kurtosis** – Measure of symmetry for a distribution of scores. Low values of kurtosis indicate the distribution is peaked in the center. High values of kurtosis suggest the distribution of few scores in the center of the distribution.

For inferential analyses, AIR calculated correlation coefficients (*r*). A correlation coefficient is a measure of association between two variables. Coefficients range from -1.0 to 1.0 with values further from 0 indicating a stronger level of association. Positive values indicate that as one scale increases in value the other scale increases. Negative values indicate that as one scale increases in value the other decreases. The meaning of these coefficients will vary depending on the analysis. When correlating selection measures with job-relevant outcomes, coefficients greater than .21 are likely to provide useful information for hiring decisions (Department of Labor, 2000).

Reliability and Agreement Analyses

In addition to descriptive and inferential analyses, AIR also calculated the amount of error for each of the selection measures. Error was estimated by examining four different analyses. These included:

- **Cronbach's Alpha (α)** – Measure of internal consistency or how much items measure a common concept. Standards from the literature suggest that a Cronbach's Alpha of .70 or greater indicates adequate internal consistency (Nunnally & Bernstein, 1994).
- **Standard Error of Measurement (SEM)** – Measure that accounts for inherent error in test scores. A given score could actually fall within a range that is created by adding and subtracting the SEM from that score. Higher SEM values lead to wider ranges and less precision than do lower SEM values.
- **Kappa (κ)** – Measure of rater agreement that corrects for chance. Researchers suggest that kappa values greater than .40 can represent adequate agreement (Landis & Koch, 1977).
- **Weighted Kappa** – Measure of relative agreement that accounts for the size of the discrepancy between ratings given by two raters. Researchers suggest that weighted kappa values greater than .40 can represent adequate agreement (Landis & Koch, 1977).

Item-Level Analyses

In addition to conducting analyses for total scores (i.e., scale-level analyses), AIR also conducted analyses that pertained to individual items. These analyses allowed AIR to identify poor performing items that could be removed to improve the overall measure. These analyses included:

- **Item Difficulty** – Percentage of people who answered an item correctly. Researchers suggest identifying items that provide a moderate spread of difficulty (Anastasi, 1988). In order to identify items with a range of difficulty, an item was considered too difficult if fewer than 20% of the test-takers answered it correctly and an item was considered too easy if more than 95% of the test-takers answered it correctly.
- **Item Discrimination** – Extent to which success on a single item corresponds to success on the whole test. Item discrimination is measured by correlating scores on a single item with the total test score while excluding that particular item. Items with low item-to-total correlations will likely reduce the overall reliability, or internal consistency, of the measure (Haladyna, 2004). In order to ensure reliability, items with item-to-total correlations of less than 0.10 were considered for revision or removal.
- **Item-Distractor** – Usually conducted for multiple-choice measures, the extent to which incorrect responses on a single item corresponds to performance on the total test. Poor item distractors will limit the ability of a measure to differentiate between test-takers

(Haladyna, 2004). Item distractor analyses can be conducted using corrected item-tototal correlations, where good items have very low or negative values. Item distractors with correlations greater than or equal to 0.05 were considered for revision or removal.

Results are presented below by measure. The results from the descriptive and reliability/agreement analyses are provided for each measure. Inferential statistics and item-level results, however, are provided for some measures but not others. Inferential statistics are presented for the Problem Solving measures only because the analyses were conducted for the purpose of examining the relationships between the Problem Solving measures and the criterion measures. Such analyses for the other measures were not possible because of the small samples sizes for each form. Such small sample sizes would have yielded unstable inferential statistics that would have been difficult to interpret. On the other hand, item-level results are presented for only those measures that were developed during the research study. These analyses were conducted to determine whether the test items were functioning properly and are not relevant for the criteria. An overview of the item-level results are presented below with more details presented in Appendix N.

Problem Solving Measures

This section presents descriptive statistics and reliability estimates for the Wonderlic Personnel Test and the Watson-Glaser Critical Thinking Measure. The inferential statistics, as evidence of construct and criterion-related validity, are presented next. Because the Wonderlic Personnel Test and the Watson-Glaser Critical Thinking Appraisal were procured, rather than developed, for this study, AIR did not intend to modify these tests' structure. Thus, item-level analyses were not conducted.

Descriptive Statistics and Reliability

Descriptive and reliability statistics for both Problem Solving measures are presented in Table 12. In this pilot test, average scores on the Wonderlic Personnel Test ($M = 20.85$, $SD = 5.93$) were comparable to those from a general population of test-takers ($M = 21.58$, $SD = 7.10$) (Wonderlic, Inc., 2002). Average scores on the Watson-Glaser Critical Thinking Appraisal ($M = 24.03$, $SD = 6.84$) were slightly lower than those observed for test-takers in entry-level positions across industries ($M = 27.70$, $SD = 5.90$) (Pearson, Inc., 2004). The Watson-Glaser Critical Thinking Appraisal was normally distributed, but the Wonderlic Personnel Test had a peaked distribution with scores accumulating toward the lower end of the distribution. That is, there was significant kurtosis and positive skewness.

Table 12. Descriptive and Reliability Statistics for Problem Solving Measures

	Wonderlic Personnel Test	Watson-Glaser Critical Thinking Appraisal
N	60	60
Mean	20.85	24.03
Median	20	24
Standard Deviation	5.93	6.84
Range (Possible)	0 - 50	0 - 40
Minimum (Observed)	12	10
Maximum (Observed)	40	37
Skewness	.30	-.43
Kurtosis	2.05	.67
Cronbach's Alpha	.82	.84
Standard Error of Measurement	2.52	2.74

Both measures showed adequate levels of internal consistency with Cronbach's Alphas exceeding .80. This suggests that scores for these measures are reliable. The Standard Error of Measurement for both measures suggested a moderate level of precision. For example, if the Wonderlic Personnel Test was administered to the same sample 100 times, the average score would fall between 18.33 and 23.37. For a similar scenario, the Watson-Glaser Critical Thinking Appraisal's average score would fall between 21.29 and 26.77. These ranges suggest that although BDOs' scores on these two tests are likely to vary, they are relatively consistent.

Inferential Statistics

The inferential statistics consisted of examining the correlation among the Problem Solving measures and between the Problem Solving measures and job performance. These statistics were calculated to provide some evidence of construct and criterion-related validity.

Construct Validity

AIR examined the correlation between the two Problem Solving measures to assess the construct validity of the measures. The Wonderlic Personnel Test and the Watson-Glaser Critical Thinking Appraisal were moderately correlated ($r = .68$, $p < .01$). This relationship was expected because both measures assess similar aspects of Problem Solving.

Criterion-Related Validity

The correlations between the Problem Solving measures and the BDO Job Performance Measure provided some evidence of criterion-related validity (see Table 13). Both Problem Solving measures showed similar patterns of association with the BDO JPM. The Wonderlic Personnel Test was found to be significantly correlated with four dimensions, whereas the Watson-Glaser Critical Thinking Appraisal was significantly correlated with three. The size of

the correlations is comparable to those commonly found in the literature for Problem Solving measures (Schmidt & Hunter, 1998).

Table 13. Criterion-Related Validity for Predicting BDO Job Performance Measure Ratings

	Wonderlic Personnel Test	Watson-Glaser Critical Thinking Appraisal
1. Observation and Assessment of Passenger Behavior	-.05	-.04
2. Assignment and Calculation of Points	.10	.17
3. Referral of Passengers	.09	.17
4. Interaction with Passengers	.02	-.05
5. Cooperation and Communication with SPOT Teammates	.14	.14
6. Interaction with Other Security Personnel	.31*	.34*
7. Documentation	.28*	.24
8. Search of Accessible Property and Review of Travel Documentation	.33*	.38**
9. Preparedness and Dutifulness	.31*	.32*

Note: N = 52. * $p < .05$, ** $p < .01$

Neither Problem Solving measure demonstrated a significant positive correlation with the PASS measures (see Table 14). The non-significant findings may be attributed to the lack of variability and range restriction observed with these criteria. As mentioned in Chapter III, BDOs must obtain a certain score on each of these measures to retain their position. Thus, BDOs who scored low on these measures were unlikely to be included as part of the pilot test. Alternatively, given the relatively small sample sizes, the observed non-significant and negative correlations may be attributed to sampling error. Specifically, these relationships may be unduly influenced by unusual patterns of scores.

Table 14. Criterion-Related Validity for PASS Measures

	Wonderlic Personnel Test	Watson-Glaser Critical Thinking Appraisal
Performance Skill Evaluation - 1st Half	-.13	-.08
Performance Skill Evaluation - 2nd Half	-.05	-.15
Competency Evaluations	-0.33*	-.23
Job Knowledge Test	.20	.20

Note. N = 38-40. * $p < .05$

The statistically significant negative correlation between the Wonderlic Personnel Test and the competency evaluations may be due to the content of these measures. For example, the Wonderlic Personnel Test emphasizes problem solving, whereas the competency evaluations include ratings of interpersonal skills and oral communication. These constructs are often unrelated and may account for this divergent relationship (Hausknecht, Trevor, & Farr, 2002; Neuman & Wright, 1999).

Work Sample Test

This section describes the results of descriptive, reliability, and item-level analyses for the Work Sample Test. Results are presented separately for the Passenger Observation, Mental Math, Writing Knowledge, and Visual Recall Assessments.

Passenger Observation Assessment

Most BDOs performed well on the Passenger Observation Assessment (see Table 15). The highest possible score for this measure was 64 and average scores ranged from 55.00 to 59.13 across the four forms. In terms of variability, there was a relatively narrow range of scores for each of the forms (SD = 1.50 to 2.99). In fact, the highest range across the four forms was 10. These narrow ranges of scores limit the usefulness of the measure because it does not differentiate among test-takers.

Table 15. Descriptive and Reliability Statistics for the Passenger Observation Assessment

	Form A V1	Form A V2	Form B V1	Form B V2
N	13	14	16	15
Mean	57.92	55.00	58.56	59.13
Median	57.00	55.50	59.00	59.00
Standard Deviation	2.87	2.99	1.50	2.23
Range (Possible)	0 - 64	0 - 64	0 - 64	0 - 64
Minimum	54	50	55	55
Maximum	63	60	60	63
Skewness	.52	-.20	-.73	-.37
Kurtosis	-1.25	-.02	-.23	-.35
Cronbach's Alpha	.55	.52	.00	.36
Standard Error of Measurement	1.94	2.07	1.51	1.79

All of the forms showed less than desirable levels of internal consistency with Cronbach's Alphas less than .70. However, the Cronbach's Alpha for Form B V1 was especially low. This very low level of internal consistency may be attributed to the restricted range of scores for this

form. Specifically, scores ranged from 55 to 60. Such range restriction limits the amount of variability observed at the item level and can deflate Cronbach's Alpha estimates (Nunnally & Bernstein, 1994).

In addition to scale-level analyses, AIR examined the item-difficulties for each behavior and appearance factor (see Table 16). Three behaviors and appearance factors were flagged for being too easy. Specifically these three factors were all answered correctly more than 95% of the time. AIR examined each of these items to determine if any should be removed.

Table 16. Average Item Difficulty for the Passenger Observation Assessment[22]

Behavior or Appearance Factor	Mean Item Difficulty
Behavior or Appearance Factor #1	.84
Behavior or Appearance Factor #2	.87
Behavior or Appearance Factor #3	.89
Behavior or Appearance Factor #4	.91
Behavior or Appearance Factor #5	.95
Behavior or Appearance Factor #6	1.00
Behavior or Appearance Factor #7	.97
Behavior or Appearance Factor #8	.81

Despite having acceptable difficulty, Behavior or Appearance Factor #8 was identified during multiple feedback sessions as problematic because this appearance factor was difficult to see. Based on this feedback, AIR reviewed these items to determine if they should be removed from the measure. A description of the changes made to each of the selection measures is presented at the end of this chapter.

Mental Math Assessment

Overall, scores on both forms of the Mental Math Assessment were elevated (see Table 17). Out of a possible 20 points, the average score for both forms was greater than 19. Furthermore, there was relatively little variability for scores on both forms (SD = 1.06 to 1.14). The distributions for both forms reflected these elevated levels of performance. Both forms were

[22] The names of the specific behaviors and appearance factors have been removed to maintain test security.

negatively skewed with scores clustering around the high end of the distribution. During the group feedback sessions, the pilot test participants commented that this assessment was too easy; this feedback mirrors the results presented in Table 17.

Table 17. Descriptive and Reliability Statistics for the Mental Math Assessment

	Form A	Form B
N	30	28
Mean	19.47	19.32
Median	20.00	20.00
Standard. Deviation	1.14	1.06
Range (Possible)	0 - 20	0 - 20
Minimum	16	16
Maximum	20	20
Skewness	-2.10	-1.73
Kurtosis	3.21	2.72
Cronbach's Alpha	.63	.46
Standard Error of Measurement	.69	.78

Both forms of the Mental Math Assessment had internal consistency values that fell below the .70 threshold. As with Form B Version 1 of the Passenger Observation Assessment, this may be partially attributed to the limited degree of variability rather than divergent items. Specifically, because the items did not vary (i.e., a high percentage of items were answered correctly), the items were less likely to be intercorrelated (Nunnally & Bernstein, 1994). This may have deflated the Cronbach's Alpha values for both forms.

AIR also examined the average item difficulty for the Mental Math Assessment (see Table 18). These values provided additional evidence suggesting that this assessment was not challenging. Specifically, on average across items, 97% of participants answered each item correctly.

Table 18. Average Item Difficulty for the Mental Math Assessment

Mental Math Assessment	Mean Item Difficulty
Form A	.97
Form B	.97

Writing Knowledge Assessment

Both forms of the Writing Knowledge Assessment were moderately difficult with means of 12.97 and 15.56 out of a possible 25 points (see Table 19). Both forms of this measure also showed a wide range of scores. Specifically, Form A had a range of 15 and Form B had a range of 19. Scores from both forms were normally distributed.

Table 19. Descriptive and Reliability Statistics for the Writing Knowledge Assessment

	Form A	Form B
N	30	27
Mean	12.97	15.56
Median	12.50	16.00
Standard Deviation	4.35	4.71
Range (Possible)	0 - 25	0 - 25
Minimum	6	6
Maximum	21	25
Skewness	.22	-.28
Kurtosis	-.83	-.38
Cronbach's Alpha	.86	.88
Standard Error of Measurement	1.66	1.62

The Writing Knowledge Assessment also demonstrated adequate levels of reliability. Form A had a Cronbach's Alpha of .86 and Form B had an alpha value of .88. Furthermore, the low Standard Error of Measurement suggested that these measures provided precise estimates of a test-taker's knowledge of writing rules.

The items on the Writing Knowledge Assessment appeared to be moderately difficult (see Table 20). On average across items, 52% and 60% of respondents answered each item correctly. This suggests that the items will differentiate among test-takers.

Table 20. Average Item Difficulty for the Writing Knowledge Assessment

Writing Knowledge Assessment	Mean Item Difficulty
Form A	.52
Form B	.60

Across both forms, five items were identified as being too challenging. That is, less than 20% of BDOs answered the item correctly. These items were reviewed and potential revisions were considered.

Visual Recall Assessment

Because the number of items for each form differed ($n = 26 - 28$), scores on the Visual Recall Assessment are presented as percentages. In other words, these values represent the total number correct responses divided by the total number of items. Scores on the Visual Recall Assessment were moderately difficult with mean values of 75%, 75%, and 67% across the three forms (see Table 21). Form A demonstrated a wide range of scores (27% to 100%), whereas the distributions for the other forms were slightly more constricted (59% to 85% and 48% to 85%). However, Form A was slightly negatively skewed with scores peaking at the high end of the distribution, whereas Forms B and C were normally distributed.

Table 21. Descriptive and Reliability Statistics for the Visual Recall Assessment

	Form A	Form B	Form C
N	20	20	18
Mean	75%	75%	67%
Median	77%	78%	69%
Standard Deviation	15%	7%	9%
Range (Possible)	0 - 100%	0 - 100%	0 - 100%
Minimum	27%	59%	48%
Maximum	100%	85%	86%
Skewness	-1.65	-.57	-.05
Kurtosis	5.07	-.45	.49
Cronbach's Alpha	.75	.00	.30
Standard Error of Measurement	1.94	1.97	2.16

Form A demonstrated an adequate level of reliability with a Cronbach's Alpha of .75. The Cronbach Alpha for Form C was less than desirable. However, after removing items that showed low item-to-total correlations, this value increased to .68. The Cronbach's Alpha for Form B was very low and could not be increased by removing problematic items.

AIR attributed this primarily to an image that was qualitatively different than those used for other forms. Specifically, the image included multiple screening lines and additional detail in the background. Items that referred to different screening lines, or the additional background content, may not have functioned in a similar fashion as those that focused on a single screening line or the foreground of an image.

All three forms showed adequate levels of item difficulty (see Table 22). On average, across items, 70% to 75% of participants answered the items correctly. Although these levels are a bit higher than a desirable level of 50%, they are also far below the established threshold of 95%.

Table 22. Average Item Difficulty for the Visual Recall Assessment

Visual Recall Assessment	Mean Item Difficulty
Form A	.75
Form B	.75
Form C	.70

Item-distractor analyses identified items that featured response options that were either (1) misleading or (2) not chosen by any BDOs. For example, some items asked for the number of objects in the image and provided numeric responses options (e.g., 1, 2, 3, and 4). For some items, one or two of the options may not have been selected by any BDOs. These types of items were reviewed and revised accordingly prior to the validation study.

Structured Interview

This section presents results for the Structured Interview. Descriptive statistics are presented for each form and each competency. These results are followed by interrater agreement statistics.

Descriptive Statistics

Given the small samples sizes for each form (see Table 23), AIR focused on median performance as opposed to average performance for the Structured Interview. Average ratings are susceptible to outliers (i.e., very high or very low performers). Median ratings for most competencies were comparable across forms. Differences in median ratings across forms for each competency were less than 1.00. For example, the median rating for the Multitasking questions ranged from 3.50 to 4.50. These median values suggest that most BDOs performed well on these questions, which is to be expected given that all pilot test participants had been previously screened using TSA's existing Structured Interview. Nevertheless, there was some variability observed with these items. Across all competencies, standard deviations ranged from .60 to 1.38. Across the four forms, ratings for responses to the Decisiveness and Teamwork items were a bit lower, whereas responses to the Honesty-Integrity items were elevated.

~~Sensitive Security Information (SSI)~~

Table 23. Descriptive Statistics for the Structured Interview

	Attention to Detail				Decisiveness				Flexibility				Interpersonal Skills				Multitasking			
Form	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
N	13	11	12	14	12	12	13	14	12	12	13	14	13	11	13	14	13	10	12	14
Mean	3.62	4.36	3.58	3.86	3.33	4.00	2.92	3.93	3.75	4.25	3.23	3.64	3.38	3.18	3.23	4.00	3.92	4.20	3.42	4.21
Median	3.00	4.00	4.00	4.00	3.50	4.00	3.00	4.00	4.00	4.50	3.00	4.00	3.00	3.00	4.00	4.00	4.00	4.50	3.50	4.50
Std. Dev.	.96	.67	1.38	1.10	1.37	.95	1.38	1.21	.87	.87	.73	1.22	.87	1.25	1.24	1.04	.86	.92	1.16	.97
Range (Poss.)	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5
Min	2	3	1	2	1	2	1	2	2	3	2	1	2	1	1	2	3	3	1	2
Max	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5
Skew	.28	-.59	-.58	-.49	-.22	-.76	-.06	-.76	-.44	-.57	-.39	-.69	.87	-.04	-.83	-.48	.16	-.47	-.59	-1.07
Kurtosis	-.89	-.29	-.83	-1.00	-1.23	.16	-.84	-.95	.23	-1.45	-.76	.10	.35	-.47	-.15	-1.10	-1.68	-1.81	.36	.37
	Self-Management				Teamwork				Honesty-Integrity				Oral Communication							
Form	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D				
N	13	11	13	14	13	12	13	14	13	11	13	14	13	12	13	14				
Mean	3.85	4.18	3.08	3.86	3.38	3.75	3.00	4.07	3.85	4.27	4.69	4.21	3.38	4.00	3.46	4.14				
Median	4.00	4.00	3.00	4.00	3.00	3.50	3.00	4.50	4.00	5.00	5.00	5.00	4.00	4.50	3.00	4.00				
Std. Dev.	.69	.60	1.12	1.29	.96	.87	.91	1.14	1.34	1.01	.85	1.12	1.12	1.21	1.27	.86				
Range (Poss.)	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5	1-5				
Min	3	3	1	1	2	3	1	2	1	2	2	2	2	2	2	3				
Max	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5				
Skew	.20	-.03	.25	-.94	.39	.57	-.78	-.88	-.89	-1.37	-3.08	-1.25	-.08	-.75	.10	-.31				
Kurtosis	-.50	.41	.47	.12	-.44	-1.45	.44	-.62	-.15	1.32	9.72	.28	-1.39	-1.05	-1.73	-1.64				

Interrater Agreement

Using this standard, most pairs of raters demonstrated moderate levels of exact agreement (i.e., choosing the same rating) and relative agreement (i.e., choosing similar ratings). This pattern was consistent whether the rating pairs consisted solely of AIR personnel, solely of STSMs, or a combination of the two.

Table 24. Kappa and Weighted Kappa Estimates for the Structured Interview

Rater Pair	N	Kappa	Weighted Kappa
AIR Raters			
Rater 1 & Rater 2	18	0.27*	0.26*
Rater 1 & Rater 3	35	0.14	0.30**
Rater 2 & Rater 3	62	0.17*	0.40**
Rater 2 & Rater 5	23	0.07	0.09
Rater 2 & Rater 6	9	0.63**	0.67**
Rater 5 & Rater 6	6	0.23	0.45*
STSM Raters			
Rater 7 & Rater 8	54	0.35*	0.40**
Rater 7 & Rater 9	18	0.18	0.26*
Rater 7 & Rater 11	36	0.77**	0.81**
Rater 12 & Rater 13	62	0.22**	0.49**
Rater 12 & Rater 14	18	0.30*	0.45**
Rater 15 & Rater 16	108	0.43**	0.60**
AIR & STSM Raters			
Rater 1 & Rater 15	36	0.35**	0.56**
Rater 1 & Rater 16	36	0.32**	0.52**
Rater 3 & Rater 15	45	0.25**	0.39**
Rater 3 & Rater 16	45	0.25**	0.39**
Rater 5 & Rater 12	18	0.03	0.34*
Rater 6 & Rater 18	18	0.09	0.13

Note. N = Number of shared observations. * $p < .10$, ** $p < .01$.

Kappa values ranged from .03 to .77, whereas weighted kappa values ranged from .09 to .81. This wide variability may be attributed to the range of shared observations for rater pairs ($N = 6 – 108$). In order to calculate stable estimates of interrater agreement, AIR examined weighted averages using only the rating pairs that had at least 10 shared observations. For AIR raters, the average kappa value was .16 while the weighted kappa value was .31. For STSM raters, the average kappa value was .25 while the weighted kappa value was .42. For AIR/STSM rater pairs, the average kappa value was .39 while the weighted kappa value was .54.

Across all the different types of rater pairs, 10 demonstrated adequate levels of agreement. Specifically, the weighted kappa values for these pairs equaled or exceeded .40. These rater pairs consisted of all three types of pairings (i.e., AIR, STSMS, and a combination). This suggests that different types of Assessors are likely to provide consistent ratings for the Structured Interview.

Role-Play Exercise

This section presents results for the Role-Play Exercise. Descriptive statistics are presented for each form and each evaluation criteria. These results are followed by interrater agreement statistics.

Descriptive Statistics

Scores on the Role-Play Exercise demonstrated adequate levels of variability (see Table 25). Specifically, average consensus ratings for the Elicitation of Information (M = 3.41 – 3.63, SD = 1.01 – 1.15) and Active Listening (M = 3.30 – 3.40, SD = 1.15 – 1.20) rating scales were slightly above the mid-point of the scale. Furthermore, raters used the entire scale when providing ratings.

Table 25. Descriptive Statistics for the Role-Play Exercise

	Elicitation of Information		Active Listening		Summary Statements	
Form	A	B	A	B	A	B
N	27	24	27	25	30	25
Mean	3.41	3.63	3.30	3.40	2.65	2.39
Median	3.00	3.00	3.00	3.00	3.00	3.00
Std. Dev.	1.15	1.01	1.20	1.15	0.72	0.85
Range (Possible)	1 - 5	1 - 5	1 - 5	1 - 5	0 - 3	0 - 3
Minimum	1	2	1	1	0	0
Maximum	5	5	5	5	3	3

Skewness	-.08	.31	-.19	-.18	-2.42	-1.38
Kurtosis	-.83	-1.25	-.75	-.82	5.98	1.23

In addition to these two rating scales, BDOs also provided answers to the questions they were tasked with asking the passenger. These answers, or Summary Statements, were then scored. Based on the mean value (2.39 – 2.65 out of 3.00) for these Summary Statements, one can conclude that most BDOs were able to elicit a majority of the requisite information from the fictitious passenger. However there was some variability as seen in the standard deviations for both forms (0.72 to 0.85) and the range of scores (0 to 3).

Interrater Agreement

As with the Structured Interview, AIR calculated measures of exact agreement (i.e., choosing the same rating) and relative agreement (i.e., choosing similar ratings) for the Role-Play Exercise (see Table 26). Specifically, AIR calculated kappa and weighted kappa values for each rater pair (i.e., AIR, STSMs, or a combination of the two).

Table 26. Kappa and Weighted Kappa Estimates for the Role-Play Exercise

Rater Pair	N	Kappa	Weighted Kappa
AIR Raters			
Rater 1 & Rater 2	10	0.39	0.66
Rater 1 & Rater 3	20	0.78	0.91
Rater 2 & Rater 3	40	0.58	0.80
Rater 2 & Rater 5	15	0.65	0.86
Rater 2 & Rater 6	5	0.72	0.90
Rater 5 & Rater 6	5	1.00	1.00
AIR & STSM Raters			
Rater 1 & Rater 15	20	0.64	0.75
Rater 1 & Rater 16	20	0.64	0.79
Rater 3 & Rater 15	25	0.64	0.80

Rater 3 & Rater 16	25	0.64	0.81
Rater 5 & Rater 12	10	0.50	0.71
Rater 6 & Rater 18	10	0.50	0.78
AIR Raters			
Rater 1 & Rater 2	10	0.39	0.66
Rater 1 & Rater 3	20	0.78	0.91
Rater 2 & Rater 3	40	0.58	0.80
Rater 2 & Rater 5	15	0.65	0.86
Rater 2 & Rater 6	5	0.72	0.90
Rater 5 & Rater 6	5	1.00	1.00

Note. N = Number of shared observations. All values were significant at the $p < .01$ level.

Overall, raters typically demonstrated higher levels of agreement for the Role-Play Exercise than for the Structured Interview. Specifically, kappa values ranged from .39 to 1.00, whereas weighted kappa values ranged from .66 to 1.00.

This range of kappa values may be partially attributed to differences in the number of shared observations. In order to calculate stable estimates of interrater agreement, AIR examined weighted averages using only the rating pairs that had at least 10 observations. For AIR raters, the average kappa value was .62 while the weighted kappa value was .82. For STSM raters, the average kappa value was .61 while the weighted kappa value was .78. For AIR/STSM rater pairs, the average kappa value was .65 while the weighted kappa value was .82.

Across the 18 different rater pairs, 17 demonstrated adequate levels of agreement. Specifically, both kappa and weighted kappa exceeded the threshold of .40 (Landis & Koch, 1977).

BDO Job Performance Measure

This section presents the descriptive, inferential, and agreement statistics for the pilot test of the BDO Job Performance Measure and the Opportunity to Observe items.

Descriptive Statistics

The BDO Job Performance Measure consists of nine performance dimensions. Each performance dimension was rated on a five-point scale, such that the minimum rating for any given performance dimension was a 1 and the maximum rating a 5. Behavioral examples of performance were provided at the 1st, 3rd, and 5th scale point for each performance dimension.

Ten STSMs, or six rater pairs, provided performance ratings using the BDO JPM. All performance raters were STSMs (n = 10). On average, these raters were 37.4 years old (SD = 7.82 years) and had worked for TSA for 7.2 years (SD = 2.71 years). Nine of the ten subject matter experts (SMEs) were male, and eight of the ten SMEs identified themselves as White.

Two sets of performance ratings were received for 67 BDOs.[37] All of the performance ratings were included in the pilot test analyses. Descriptive statistics for each of the nine performance dimensions are presented in Table 27, below. The mean ratings for each performance dimension ranged from 3.65 (*Interaction with Passengers*) to 4.39 (*Preparedness and Dutifulness*). Across all performance dimensions, the mean ratings were slightly negatively skewed, with the highest skewness for Dimension 8 (*Search of Accessible Property and Review of Travel Documentation*) and Dimension 9 (*Preparedness and Dutifulness*). Overall, the range of scores for each

performance dimension (provided by the difference between the minimum and maximum rating for each performance dimension) indicate that raters used most of the scale points to make their ratings, but were less likely to use the lower values (i.e., 1 and 2) of the scale.

Although the BDO JPM included separate dimensions of performance, each of these categories pertained to the BDO job. Thus, to create a parsimonious measure of BDO job performance, AIR computed a composite BDO JPM rating. Specifically, AIR averaged the separate ratings for each of the nine dimensions to create a single value for each BDO on the BDO JPM. The descriptive statistics for this composite scale are also presented in Table 27, below. The internal consistency of this composite measure was very high (α = .93).

[37] As mentioned previously, only 60 BDOs completed the selection measures. These additional ratings represent BDOs who were scheduled to participate but did not take part in the on-site data collection process.

Table 27. Descriptive Statistics for the BDO Job Performance Measure

	Observation and Assessment of Passenger Behavior	Assignment and Calculation of Points	Referral of Passengers	Interaction with Passengers	Cooperation and Communication with SPOT Teammates
N_{38}	67	67	67	67	67
Mean	3.81	4.01	4.00	3.65	3.75
Median	4.00	4.00	4.00	4.00	4.00
Variance	1.08	.74	1.08	.90	.64
Std. Dev.	1.04	.86	1.04	.95	.80
Minimum	1.50^	2.00	1.00	1.50^	2.00
Maximum	5.00	5.00	5.00	5.00	5.00
Skewness	-.56	-.50	-.84*	-.37	-.27
Kurtosis	-.84	-.75	-.20	-.79	-.63
	Interaction with Other Security Personnel	**Documentation**	**Search of Accessible Property and Review of Travel Documentation**	**Preparedness and Dutifulness**	**Composite BDO JPM**
N	67	67	67	67	67
Mean	4.05	3.90	4.32	4.39	3.98
Median	4.00	4.00	4.67^	4.67^	4.33
Variance	.71	.96	.64	.58	.52
Std. Dev.	.84	.98	.80	.76	.72
Minimum	2.00	1.00	3.00	2.00	2.44
Maximum	5.00	5.00	5.00	5.00	5.00
Skewness	-.42	-.65	-.69	-1.18*	-.71*
Kurtosis	-.96	-.24	-1.13	.61	-.63

Note. * $p < .05$; ^ Median and minimum scores are not whole numbers because performance scores were aggregated across raters. [38] The sample size refers to the number of BDOs that were rated for each dimension.

In addition to providing performance ratings, raters completed three items that assessed their opportunity to observe each ratee. Raters responded to each item using a five-point scale (1=*almost never had the opportunity to observe* to 5 = *very frequently had the opportunity to observe*). The descriptive statistics for each of the Opportunity to Observe items are presented in Table 28, below. The mean and median values for these items indicate that, overall, most raters had a significant opportunity to observe ratees prior to making performance ratings. However, the range of ratings for these items, and specifically the minimum ratings, indicate that some raters may have had very little opportunity to observe the performance of the BDOs.

Table 28. Descriptive Statistics for the BDO Manager Opportunity to Observe Measure

	Performance Conducting SPOT	Written Reports and Other Documentation	Knowledge of SPOT and the SPOT Standard Operating Procedures (SOP)	Composite Opportunity to Observe Measure
N	67	67	67	67
Mean	3.98	4.03	3.87	3.96
Median	4.00	4.00	4.00	4.00
Variance	.61	.66	.59	.43
Std. Dev.	.78	.81	.77	.66
Minimum	2.00	1.00	2.00	1.67
Maximum	5.00	5.00	5.00	5.00
Skewness	-.65	-1.06*	-.66	-.88*
Kurtosis	.00	1.83*	-.02	1.59*

Note. * $p < .05$.

A composite of these Opportunity to Observe items was created by averaging across all three items. These results are also presented in Table 28. The internal consistency of this composite measure was acceptable for research purposes (α= .79). Overall, this composite measure reflects the conclusions drawn from the individual Opportunity to Observe items. Specifically, although most raters had a significant opportunity to observe performance, there was also a wide range such that some raters had very little opportunity to observe the BDOs they were rating.

Inferential Statistics

Table 29, below, presents the intercorrelations among the performance dimensions of the BDO JPM and the Opportunity to Observe Items. Performance dimensions generally showed moderate to strong intercorrelations, ranging from .23 to .83. Not surprisingly, the highest correlations were found among performance dimensions that encompass similar behaviors. For example, Dimension 1 (Observation and Assessment of Passenger Behavior) and Dimension 2 (Assignment and Calculation of Points) were strongly correlated (r = .80). Weaker correlations were found among performance dimensions with more divergent behaviors. For instance, Dimension 5 (Cooperation and Communication with SPOT Teammates) and Dimension 7

(Documentation) were weakly correlated (r = .23). Overall, these results indicate that although some performance dimensions are highly correlated, each dimension accounts for a unique portion (i.e., variance) of the overall performance domain.

Table 29. Intercorrelations Among Dimensions of the BDO Job Performance Measure and the Opportunity to Observe Items

BDO Job Performance Measure - Dimensions	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Observation and Assessment of Passenger Behavior													
2. Assignment and Calculation of Points	.80												
3. Referral of Passengers	.81	.83											
4. Interaction with Passengers	.78	.72	.67										
5. Cooperation and Communication with SPOT Teammates	.65	.53	.54	.61									
6. Interaction with Other Security Personnel	.60	.64	.73	.54	.51								
7. Documentation	.41	.53	.59	.30	.23	.59							
8. Search of Accessible Property and Review of Travel Documentation	.64	.68	.75	.50	.45	.81	.58						
9. Preparedness and Dutifulness	.52	.55	.66	.34	.40	.77	.56	.77					
10. Performance Conducting SPOT*	.24	.19	.18	.12	.11	-.06	.28	.04	.05				
11. Written Reports and Other Documentation*	.37	.21	.35	.18	.34	.42	.36	.52	.42	.34			
12. Knowledge of SPOT and the SPOT SOP*	.17	.09	.18	.07	.14	.04	.33	.14	.09	.78	.55		
13. BDO JPM Composite	.87	.88	.92	.77	.68	.85	.67	.85	.76	.17	.43	.18	
14. Opportunity to Observe Composite	.31	.19	.28	.15	.24	.17	.39	.28	.22	.84	.76	.92	.31

Note. * Denotes an Opportunity to Observe item.

Agreement Statistics

Rater agreement is presented in Table 30, below. All pairs of raters demonstrated at least moderate levels of exact agreement (i.e., weighted kappa) for their performance ratings, and all kappa and weighted kappa values were significant at $p < .0001$. Specifically, kappa values ranged from .27 to .60, whereas weighted kappa ranged from .39 to .73. With regard to the recommended threshold of .40 (Landis & Koch, 1977), one of the six rater pairs exceeded this threshold for the kappa values, and five of six exceeded this threshold for the weighted kappa values. Overall, this indicates acceptable levels of interrater agreement between pairs of raters.

Table 30. Kappa and Weighted Kappa Values for Performance Ratings

Rater Pair	N	Kappa	Weighted Kappa
Rater 2 & Rater 9	54	0.60	0.73
Rater 3 & Rater 10	54	0.38	0.56
Rater 5 & Rater 6	99	0.37	0.51
Rater 6 & Rater 7	99	0.32	0.55
Rater 4 & Rater 11	162	0.27	0.39
Rater 5 & Rater 7	99	0.27	0.48

Note. N = Number of shared observations between raters across performance dimensions. All kappa and weighted kappa values are significant at $p < .0001$.

Feedback Questionnaire

This section presents results that pertain to the job-relatedness subscales of the feedback questionnaire. (Complete results from the feedback questionnaire are presented in Appendix O.) Because these subscales provide a direct estimate of face validity,[39] AIR emphasized these ratings while revising the selection measures following the pilot test. Specifically, descriptive statistics for each subscale across the eight assessments are presented. Subscales were derived by calculating the average for each item within the subscale. AIR also conducted tests of difference to determine if any selection measures received significantly higher ratings than other measures in the experimental battery. This section concludes with a brief review of the comments that were provided by BDOs.

[39] Face validity represents the degree to which BDOs perceive a correspondence (i.e., relatedness) between the selection measures and their job.

Descriptive Statistics

As part of the feedback questionnaire, BDOs rated each selection measure on two aspects of jobrelatedness. First, BDOs considered the extent to which the measures appeared to assess content that was relevant to the job (i.e., content job-relatedness) (see Table 31). Overall, the two Problem Solving measures received the lowest ratings (M = 1.80 and 2.41) for content jobrelatedness. The Passenger Observation Assessment and Role-Play Exercise were two of the highest rated measures (M = 4.09 and 4.14, respectively). Nonetheless, the Mental Math Assessment and Visual Recall Assessment were also seen as measuring job-relevant constructs (M = 3.88 and 3.91, respectively).

Table 31. Descriptive Statistics for the Content Job-Relatedness Subscales

	Passenger Observation	Mental Math	Writing Knowledge	Visual Recall
N	58	58	58	58
Mean	4.09	3.88	3.47	3.91
Median	4.00	4.00	3.50	4.00
Standard Deviation	0.80	0.92	0.91	0.81
Minimum	1.00	2.00	1.00	2.00
Maximum	5.00	5.00	5.00	5.00
	Wonderlic Personnel Test	**Watson-Glaser Critical Thinking Appraisal**	**Structured Interview**	**Role-Play Exercise**
N	60	60	54	55
Mean	1.80	2.41	3.24	4.14
Median	2.00	2.00	3.50	4.00
Standard Deviation	0.73	0.97	0.91	0.76
Minimum	1.00	1.00	2.00	2.00

Maximum	5.00	4.00	5.00	5.00

Second, BDOs considered the extent to which higher scores on a measure were likely to correspond to better job performance (i.e., predictive job-relatedness) (see Table 32). Similar to content job-relatedness, BDOs rated the Problem Solving measures as least likely to correspond to job performance (M = 2.43 and 2.88). The Passenger Observation Assessment and Role-Play Exercise (M = 3.86 and 3.99, respectively) were the two highest rated measures on this subscale, although only by a small margin.

Table 32. Descriptive Statistics for the Predictive Job-Relatedness Subscale

	Passenger Observation	Mental Math	Writing Knowledge	Visual Recall
N	58	58	58	58
Mean	3.86	3.51	3.30	3.62
Median	4.00	3.50	3.25	4.00
Standard Deviation	0.67	0.92	0.88	0.81
Minimum	1.00	1.00	1.00	1.00
Maximum	5.00	5.00	5.00	5.00
	Wonderlic Personnel Test	**Watson-Glaser Critical Thinking Appraisal**	**Structured Interview**	**Role-Play Exercise**
N	60	60	54	55
Mean	2.43	2.88	3.36	3.99
Median	2.50	3.00	3.50	4.00
Standard Deviation	0.76	0.96	0.84	0.78
Minimum	1.00	1.00	1.00	2.00
Maximum	4.00	5.00	5.00	5.00

Ratings from both the content and predictive job-relatedness subscales suggest that the Work Sample Test and Role-Play Exercise were seen as highly related to the BDO job. The Structured Interview received moderate ratings of job-relatedness. The Problem Solving measures were seen as the least relevant to the BDO job.

Comparison of Measures

To determine if the selection measures were significantly different on any of the subscales of the feedback questionnaire, AIR conducted two one-way Analyses of Variance (ANOVAs). For each ANOVA, the independent variable was the type of selection measure (e.g., Wonderlic

Personnel Test, Passenger Observation Assessment). The dependent variable was the average ratings for each of the selection measures on a particular feedback subscale (e.g., predictive jobrelatedness). The omnibus results for each of these ANOVAs are presented in Table 33, which includes effect size estimates (η^2). The effect size estimates indicate the amount of variance accounted for in the subscales that can be attributed to the differences in the predictors. Higher η^2 suggest larger differences in levels of job-relatedness across assessments.

Table 33. *F* Statistics for ANOVAs Comparing Differences in Subscales Across the Selection Measures

Subscale	Omnibus *F* Results	Eta-squared (η^2)
Content Job-Relatedness	$F(7, 453) = 58.25$, $p < .001$	.47
Predictive Job-Relatedness	$F(7, 453) = 22.43$, $p < .001$	.26

Overall, both ANOVAs indicated that there was a significant difference between the measures for each subscale. The largest differences between the selection measures appeared to exist on the content job-relatedness subscale. Slightly smaller differences were observed for the predictive job-relatedness subscale. In order to isolate the differences between the selection measures for each subscale, AIR conducted post-hoc tests of difference. Specifically, AIR used a Bonferroni post-hoc test. This measure conservatively estimates significant differences by adjusting for the number of comparisons. Based on the results of the Bonferonni post-hoc tests, the selection measures were rank ordered for each subscale on the feedback questionnaire. The results of the post-hoc tests are presented in Table 34.

Table 34. Relative Ordering of the Selection Measures Based on Post-Hoc Differences Across Subscales

Selection Measure	Feedback Questionnaire Subscale	
	Content Job-Relatedness	Predictive JobRelatedness
Wonderlic Personnel Test	8	8
Watson-Glaser Critical Thinking Appraisal	8	7
Passenger Observation	1	1
Mental Math	1	2
Writing Knowledge	2	3
Visual Recall	1	2
Structured Interview	2	3
Role-Play Exercise	1	1

The Passenger Observation, Visual Recall, and Mental Math Assessments, as well as the RolePlay Exercise, were some of the highest rated measures across the two subscales. On the other hand, the Wonderlic Personnel Test and the Watson-Glaser Critical Thinking Appraisal received the lowest ratings of job-relatedness.

Qualitative Review of the Responses to Open-ended Questions and Comments

Each feedback questionnaire included a section for BDOs to provide comments regarding the clarity of the instructions, the difficulty of the measures, and other recommendations for improving the assessments. With the exception of the Role-Play Exercise, BDOs indicated that the instructions for each assessment were very clear and straight forward. For the Role-Play Exercise, a handful of BDOs recommended shortening the instructions to ensure candidates were engaged throughout the administration process. This is evidenced by the following quote,

—Initial instructions are very extensive. A candidate may be overwhelmed by the entire process...‖

BDOs also provided comments regarding the difficulty of each selection measure. First, although many BDOs found the questions on the Wonderlic Personnel Test to be easy, they found the overall test to be difficult due to the amount of time provided to complete the measure (i.e., 12 minutes). For example, one BDO commented, —I felt most of the questions were easy but not enough time [was given] to complete the test.‖

For the Work Sample Test, BDOs indicated that the levels of difficulty varied across the assessments. In general, BDOs indicated that there were too many passengers included in the Passenger Observation Assessment. For example, one BDO indicated, —Too many people to follow at once. Confusing for someone who does not [have] observation techniques.‖ BDOs also indicated they found it challenging to ignore the SPOT behaviors they were trained to observe. For example, one BDO commented, —I found learning new behaviors in just a few seconds [and] trying not to use ingrained behaviors learned was a little confusing.‖

Most BDOs found the Mental Math Assessment to be easy. Consistently, BDOs recommended adding negative numbers to the items to increase both the realism and the difficulty of the assessment. For example, one BDO commented, —To make this even more relevant to the BDO job, add problems with simple subtraction as well.‖

In general, BDOs found the Writing Knowledge Assessment, the Visual Recall Assessment, and the Structured Interview to be moderately difficult and related to the BDO job. In response to the Writing Knowledge Assessment, most BDOs provided feedback similar to a participant who said, —Written reports play an important role in the job. Good to see an assessment that takes this into account.‖ As an example of the responses regarding the Visual Recall Assessment, one

BDO indicated, —Very closely related to the BDO job and I feel this would be a great test for a new hire.‖ Similarly, participants commented that the Structured Interview was moderately difficult and provided an opportunity to demonstrate their skills, saying, for example, —The questions asked provided the opportunity for me to demonstrate that I could process and recall information in a high pressure situation. This relates to the BDO casual conversation.‖

For the Role-Play Exercise, BDOs confirmed that they would be better prepared than untrained job candidates. One BDO commented, —This was a good exercise, however I could imagine how a person who is not yet a BDO would become nervous under such on-the-spot scenarios.‖

Finally, BDOs provided a number of general comments that were similar across sites. For instance, many BDOs commented on the poor resolution and brightness of the Passenger Observation Assessment video. This was primarily due to the use of a portable projector for displaying the videos. Additionally, many BDOs questioned the job-relatedness of the Writing Knowledge Assessment, suggesting that the assessment emphasized editing, which could be performed by the Spelling and Grammar function within Microsoft Word®.

Final Predictor Measures for Validation Study

On the basis of the results of the pilot test, the feedback received from BDOs, and AIR experiences during the pilot test, AIR identified a number of potential changes that needed to be considered before assessing the selection (predictor) and criterion measures in the validation study. These changes were reviewed during a briefing held in June 2011 with representatives from TSA's Office of Human Capital (OHC) and OSO. This section describes the changes that were made to each of the selection measures following the pilot test.

Problem Solving Measures

AIR retained both Problem Solving measures for the validation study. These measures demonstrated adequate levels of variability and internal consistency during the pilot test. Both the Wonderlic Personnel Test and the Watson-Glaser Critical Thinking Appraisal were retained in the experimental selection battery primarily to serve as measures of construct validity.

The Watson-Glaser Critical Thinking Appraisal was also retained to serve as an alternative assessment in the event that the other newly developed selection measures demonstrated inadequate psychometric properties in the validation study.

Work Sample Test

In addition to retaining the Problem Solving measures, AIR modified the four assessments within the Work Sample Test. This section describes these changes.

Passenger Observation Assessment

Following the pilot test, AIR decided to reduce the number of forms for the Passenger Observation Assessment to address poor functioning items and improve the variability of the total scores. The best functioning passengers (i.e., those who did not display tricky behaviors or appearance factors and were properly spaced throughout the video) were retained from each version within a form. For example, the best functioning passengers from Form A Version 1 and Form A Version 2 were combined into a single form (i.e., Form A). A similar process was conducted for the two versions of Form B. This process was informed by item-level analyses and feedback from BDOs. For example, AIR retained those passengers that displayed behaviors and appearance factors with adequate levels of item difficulty. Passengers who were identified as tricky or overly difficult by the BDOs were removed.

Second, AIR identified two behaviors and appearance factors for removal: *Wearing Glasses* and *Shaking Hands with Another Passenger*. AIR removed the appearance factor *Wearing Glasses* because many BDOs indicated it was difficult to see if the passengers were wearing glasses, especially when the glasses had light-colored frames. AIR removed the behavior *Shaking Hands with Another Passenger* because it did not show any variability across the forms; this item was answered correctly by all of the BDOs. Furthermore, there was an instance in one of the video recordings when a passenger acknowledges another person in-line with an informal handshake, similar to a high-five. Some BDOs commented that this behavior could be seen as a handshake by some participants, which would make the item less objective.

Third, AIR modified the Passenger Observation Assessment to ensure that job candidates could more easily identify the appropriate passengers. For example, AIR modified the written descriptions of the passengers to make each passenger more distinguishable. Also, the amount of time that the tone sounded and the identification boxes appeared above the passengers head was increased to three seconds. This helped ensure that job candidates could identify passengers even if they had taken their attention away from the screen in order to record their responses. The identification boxes were also repositioned for some passengers so that it was clearer which passenger was identified. Finally, the numbering system was changed to reflect which of the requisite passengers was identified (i.e., passenger one through eight) rather than the passenger's relative order in the entire queue (i.e., the passenger's position in the security screening line from the first passenger).

Fourth, AIR implemented a group-based administration that relied on computer monitors, rather than a projector. This was done for two reasons. First, a number of BDOs indicated that it was difficult to see the details of the video recording. The BDOs indicated that they could not identify certain behaviors and appearance factors, especially for those passengers in the second row of the security screen line. Second, this modification allowed AIR to more efficiently

administer the assessment. During the two site visits to Boston's Logan Airport (BOS) and Baltimore-Washington International (BWI), AIR pilot tested group administrations of this measure to determine whether there would be any significant impact on the testing process. The BDOs and the pilot data suggested that this modified approach did not hinder performance. Thus, AIR decided to implement the Passenger Observation Assessment on computer monitors in small groups (i.e., three to four BDOs) during the validation study.

Finally, AIR reduced the length of the instructions in the administration guide, while retaining the important content. Many BDOs commented during the pilot test that these instructions were too long and that it may be difficult for a job candidate to retain so much information. Thus, AIR revised these instructions so that they were shorter and emphasized the key pieces of information a job candidate would need to complete the measure.

Mental Math Assessment

Mental Math Assessment scores from the pilot test indicated that this measure was too easy. Many of the participants answered every item correctly. As a result, AIR modified this measure to increase its difficulty, while ensuring that the requisite tasks were job-relevant. First, AIR included negative numbers in the revised version. (b)(3):49 U.S.C. § 114(r)

(b)(3):49 U.S.C. § 114(r)

(b)(3):49 Second, AIR slightly increased the speed at which each value appeared on the screen. The amount of time was reduced from 1.5 seconds to 1 second. Finally, AIR removed items that consisted of three values or less. These items were answered correctly by nearly all of the BDOs during the pilot test. Thus, they did not provide any useful information in differentiating among participants.

In addition to increasing the difficulty of the measure, AIR also improved the clarity of the instructions. Specifically, AIR modified the initial instructions to clarify the design of the test and the structure of the test items for test-takers prior to completing the example item. AIR also added language in the administration script that explained the purpose of including behaviors and appearance factors on the slide (i.e., improve the realism of the test) and provided additional information following the practice items. **Writing Knowledge Assessment**

Following the pilot test, AIR modified the Writing Knowledge Assessment to ensure that testtakers did not edit or rewrite the passage, but instead identified and circled the errors. First, the name of the measure was changed to —Writing Knowledgel so that test-takers did not believe they had to draft or create a Written Summary. Also, instructions were added to the administration guide and the response booklet that emphasized the fact that test-takers were to focus only on identifying and circling errors. Finally, in the example item, AIR included circles that identified errors in the passage, rather than displaying the correct forms of the sentences.

Visual Recall Assessment

Following the pilot test, AIR reduced the number of Visual Recall Assessment forms from three to two (i.e., Forms A and C); Form B was removed due to its low levels of internal consistency. AIR also removed or edited items that were identified as tricky or misleading by item-level analyses or BDOs during the group feedback sessions. For example, items that referred to passengers who were —facing the camera‖ were removed for being too subjective. AIR also edited response options that were not selected by BDOs or showed negative-item total correlations.

In addition to revising the content of the measure, AIR also updated the instructions. For example, language was added to the script that emphasized that test-takers should focus on passengers *and* the entire background environment.

Finally, AIR developed a practice test for this measure. This item consisted of a single image and two multiple-choice items. Items were specifically developed to provide test-takers with an understanding of the type of information they should be memorizing (e.g., details of the passenger, objects in the background environment), without using items that were included on the actual test.

Structured Interview

Following the pilot test, TSA's Office of Human Capital recommended removing the Structured Interview from the validation study. There were a number of benefits and drawbacks to this approach. First, including the Structured Interview as part of the validation study would have provided a more comprehensive evaluation of the proposed selection system. For example, AIR could have collected additional data for each interview item to determine which ones demonstrated the most variability and elicited relevant responses. Also, AIR could have examined the degree of overlap among the newly developed selection measures and the Structured Interview. On the other hand, the Structured Interview was one of the most resource intensive assessments. It typically lasted 60 minutes and could only be administered on a oneonone basis. AIR estimated that including this instrument would have added one to two days of data collection per site visit. Given that the Structured Interview had been previously included as part of a TSA content-validation study, AIR agreed with OHC's recommendation and removed the assessment from the validation study.

On the basis of the results of the pilot test and feedback from BDOs, AIR modified the updated Structured Interview. Specifically, AIR revised the updated Structured Interview questions, and edited the rating scales, competency definitions, probes, and administration procedures. The following sections describe each of these changes.

Interview Items

After the pilot test, AIR identified the two best functioning interview items for each competency. These items were chosen based on the range of observed scores and a qualitative review of the types of responses that were typically elicited. Items that provided more variability were favored over those with a narrower range of scores because these items are more likely to provide useful information during the BDO hiring process. Likewise, AIR identified items that typically elicited responses that aligned with the intended competency. For example, Teamwork items that typically led BDOs to discuss how they facilitated a group's performance, rather than discuss their own activities, were retained.

AIR also sought to retain pairs of items that were parallel within a competency. For example, within the Multitasking competency, AIR retained two items that pertained to how a job candidate can synthesize multiple pieces of information. This increased parallelism will allow TSA to continue to provide Assessors with the option of selecting which interview item to administer.

Competency Probes, Definitions, and Rating Scales

In addition to refining the Structured Interview items, AIR also modified the Structured Interview probes. In the current Structured Interview, each interview question is associated with Follow-Up Probes. These probes can be used to prompt the job candidate to provide more information or clarification. The goal in using Follow-Up Probes is that the Interview Assessor will gain enough information to evaluate the job candidate's performance on the competency, while treating job candidates consistently throughout the hiring process. Currently, the same probes are used across each of the items. Specifically, the following probes are included for each competency:

- —Can you please provide some more information?‖
- —What specifically did you do?‖
- —What was the outcome of that situation?‖

During the pilot test, however, AIR determined that the development of more specific Follow-Up probes would allow for increased specificity in requesting additional information while maintaining the standardization desired in the use of Follow-Up Probes. As such, AIR recommended adding *item-specific* Follow-Up Probes. Example item-specific Follow-Up Probes for a newly developed question in the Teamwork competency are:

- —Can you please provide some more information about how you helped the team work together?‖, and

- —What specifically did you do to help the team work together?‖

In addition to these Follow-Up probes, AIR recommended including a new type of interview probe: Item Probes. During the pilot test, several participants indicated that they had no previous personal experience that pertained to the competency in question. To address these types of responses, AIR recommended adding Item Probes to help avoid instances in which a job candidate cannot be scored on a particular question. That is, if a job candidate is unable to provide a response to a given question (e.g., says, "That has never happened to me"), raters may try to elicit a response by using the Item Probes. Example Item Probes for a newly developed interview item assessing the Attention to Detail competency are:

- —Has there been a task at work or school that required you to follow specific instructions?‖, and
- —Has there been a task at home that required you to follow specific instructions?‖

Along with updating the Structured Interview probes, AIR also recommended modifying the competency definitions and rating scales to increase their ease of use by the Structured Interview Assessors. In regards to competency definitions, AIR recommended some minor revisions to increase their readability. Examples of revisions to the Multitasking definition include the following:

- The term —utilizing‖ was changed to —use,‖ and
- The phrase —informational input‖ was changed to —information.‖

AIR also determined that differentiating a competency's definition from the behavioral anchors contained in the rating scale may increase the ease with which raters could use both tools to reliably provide ratings. For example, the current Teamwork definition includes the following behavioral anchors that are also included in the *Exemplary* rating scale point: —Encourages and facilitates cooperation, pride, trust, and group identity; fosters commitment and team spirit; works with others to achieve goals.‖ AIR recommended revising the Teamwork competency's definition to avoid using the same behavioral anchor language in its definition and scale anchors. The proposed definition was as follows: —Works cooperatively with others to achieve shared goals; openly shares information, knowledge, and expertise with the team; puts team goals ahead of individual/personal goals.‖ The revised definition of Teamwork represents a summary of the competency without including specific anchors from any of the rating scale points.

Proposed changes to each competency's rating scale were intended to increase the ease with which Interview Assessors would be able to use the rating scales to provide reliable ratings. The types of changes proposed include:

- increasing the parallelism across the anchors within each rating scale,
- adding exemplars to assess responses for the newly developed interview questions,
- removing frequency-based qualifiers, and
- using a dichotomous rating scale for the Honesty-Integrity competency.

In the current version of the Structured Interview, a five-point rating scale is provided for each competency. Behavioral anchors are provided at the 5 (*Exemplary*), 3 (*Satisfactory*), and 1 (*Unsatisfactory*) scale points. The type of anchors provided at each scale point in the original rating scales varied within a competency. In other words, the types of criteria used to describe *Exemplary* performance may have differed from the criteria used to describe *Satisfactory* or *Unsatisfactory*. These differences made it difficult for the Interview Assessors to provide consistent ratings using the rating scales. Changes were proposed to create varying levels of the *same* behavioral anchors in each rating scale (i.e., very effective, somewhat effective, and ineffective examples of the same behavior).

For example, in the Multitasking competency, one anchor in the *Satisfactory* rating scale point specifically referenced management of conflicting priorities, whereas the other scale points did not explicitly include —time management.‖ As such, AIR recommended incorporating varying levels of time management in all three scale points.

AIR also recommended re-ordering the anchors within the *Exemplary*, *Satisfactory*, and *Unsatisfactory* scale points to present them in the same order for each competency. For example, anchors were edited and reordered in the Self Management competency such that varying degrees of the following behaviors are presented in the same order for each rating scale point: Goal setting, level of effort displayed, prioritization of work, on-time completion of tasks, and level of initiative displayed. In some cases, AIR recommended adding anchors that relate to the newly developed Structured Interview items. For example, in the Attention to Detail competency rating scale, an anchor regarding the job candidate‘s ability to provide clear instructions was added to each scale point to address a new item.

With the exception of the Oral Communication objective, which can be observed throughout a job candidate‘s interview, AIR recommended removing all behavioral anchors that assess the frequency of a behavior (e.g., always, consistently). Because the interview provides only a single example of a job candidate‘s behavior, and it is unlikely that they will refer to a series of behaviors, the frequency-related qualifiers are less helpful for differentiating a job candidate‘s performance between two scale points. The frequency qualifiers were replaced with —degree of effectiveness‖ qualifiers (e.g., very effective behavior, somewhat effective behavior, ineffective behavior).

For the new Honesty-Integrity items, AIR developed a five-point rating scale to evaluate a job candidate's response. This scale was similar in format to the existing ratings scales. Although the five-point rating scale allowed Assessors to evaluate a wide-range of performance, some commented that they had difficulty differentiating between *Satisfactory* and *Exemplary* responses with regard to a job candidate's Honesty/Integrity. Thus, AIR also developed a twopoint rating scale for TSA to consider implementing. This scale features behavioral anchors for *Unsatisfactory* and *Satisfactory* responses. This scale may be easier for Assessors to use while still enabling hiring officials to differentiate among job candidates.

Administration Procedures for Structured Interview

In addition to updating the items, competency definitions, probes, and rating scales, AIR recommended a few modifications to the administration process for this assessment. These include:

- Providing a Candidate Packet for each job candidate to review while the instructions are read aloud,
- Providing high-level information about the competencies assessed in the interview,
- Standardizing scoring procedures for questions that the job candidate does not answer,
- Adding key administration instructions provided in the Structured Interview Training to the Structured Interview Administration Guide,
- Labeling each competency in the administration guide so candidates cannot read the title, and
- Using testing signs (e.g., List of Prohibited Items, Testing in Progress).

Each of these modifications is described in more detail below.

During the pilot test, participants were provided with a —Candidate Packet‖ which contained a copy of the script that was read aloud by the Administrator. This process was implemented to accommodate job candidates who may be able to retain the instructions best by following along with the Administrator on a hard copy, rather than simply listening to them read aloud. AIR recommended that TSA implement this change in the administration of the Structured Interview.

In addition to the —Candidate Packet‖, AIR recommended that TSA provide job candidates with a high-level description of some of the competencies assessed during the interview. This recommendation is based on comments made by BDOs during the job analysis that they did not see the relationship between the interview questions and the job. Providing job candidates with an indication of the performance criteria may help address this issue and could improve the

perceived job-relatedness of the measure without impacting its ability to provide useful information during the hiring process (Klehe et al., 2008).

Following the pilot test, AIR determined that it may be advantageous to document items for which job candidates are unable to provide a response. Documenting which interview questions typically lead to non-responses may help inform future item revisions. As such, AIR recommended the addition of an —N/A‖ response option to the current five-point scale (1=*Unsatisfactory,* 5=*Satisfactory*). In addition, AIR recommended adding instructions to the Structured Interview Administration Guide to indicate how to score job candidates on interview items for which they do not provide a response (i.e., provide a not applicable or N/A option).

The original Structured Interview Administration Guide provides key guidance with regard to its administration, some of which is not currently included in the TSA's BDO Administration Guide. As a result, AIR recommended including the following instructions in TSA's BDO Administration Guide:

- —If a candidate has difficulty responding to a question, make a note, and return to the question at the end of the interview‖, and
- —Also, note that information obtained from any interview question can be considered when making a competency rating‖.

In addition, AIR recommended modifying the layout of this guide to list the competencies in a more discreet fashion. During the pilot test, AIR observed some instances where Interview Assessors had to cover their administration guides so that job candidates could not see which competency was associated with the question asked. To make it easier for Interview Assessors to conceal the name of the competency during the Structured Interview, AIR recommended reformatting the names of the competency on each rating scale page. Specifically, competency names were moved to the upper right hand corner and a lighter grey font color was used.

Finally, during the pilot test, AIR added testing administration signs as an appendix to the Revised Structured Interview Administration Guide to be used during the administration of each of the assessments. These signs included a:

- —Prohibited Items‖ sign, which listed pencils, calculators, blank sheets of paper, cell phones, books, newspapers, and electronic devices (e.g., iPod) as items from the testing room, and
- —Testing in Progress‖ sign, which was used to indicate that the room was in use.

These signs helped to ensure a proper testing environment during the pilot test and are recommended additions for the BDO Administration Guide.

Role-Play Exercise

Following the pilot test, AIR revised the Role-Play Exercise to clarify the instructions, differentiate the rating scales, update the Resource Person's script, and modify the alternate forms. Each of these changes is described in more detail in this section.

Many of the BDOs commented on their feedback questionnaires and during the feedback sessions that the instructions for the Role-Play Exercise were too long. AIR shortened the instructions, while retaining the key pieces of information. BDOs also commented that the image of the security officer that was included to provide additional contextual information during the instructions for external applicants may be distracting. Thus, AIR removed this image from the Role-Play Exercise for the validation study.

In addition to revising the instructions, AIR also modified the rating scales for the Role-Play Exercise. Both AIR and STSM Assessors commented that it was difficult to differentiate the two rating scales (i.e., Elicitation of Information and Active Listening). Thus, AIR repositioned certain anchors across the rating scales so that they were better aligned to their definitions. For instance, AIR moved the qualifier that addresses a job candidate's ability to embed questions within the conversation from Elicitation of Information to Active Listening. AIR also renamed the Active Listening rating scale so that it better reflected the underlying construct. Specifically, the rating scale was renamed Conversation Skills.

Following the pilot test, AIR also modified the Resource Person's script. These modifications were based on a qualitative review of the Role-Play Exercise performances. Specifically, AIR recorded each instance in which a BDO's question elicited an —I don't know‖ response from the Resource Person. Based on the frequency and type of question, AIR updated the Resource Person's script to include a response to the question. For example, AIR included information in the revised script that addressed the passenger's baggage, how frequently he or she travels, his or her familiarity with the screening process, and information about the weather at his or her destination. These modifications were made to help improve the naturalness of the conversation.

Finally, AIR modified how the alternate forms of the Role-Play Exercise were structured in order to increase their parallelism. Specifically, rather than use two separate character scripts as a foundation for the two forms, AIR developed a single character but created parallel Summary Statement questions. Thus, the forms differed not by the Resource Person's information, but by the questions that the job candidate was tasked with answering. As additional information was

added to the character's script, it became more likely that the Resource Person would have difficulty keeping the two fictitious characters distinct. This may have increased the likelihood that a Resource Person would provide inaccurate information that impacted a test-taker's performance. Thus, by creating a single character, AIR intended to improve the standardization for the administration of the alternate forms. In order to develop these parallel questions, AIR considered the type and amount of information each question required a test-taker to gather. For example, if one form asked job candidates to gather information about the Resource Person's trip to the airport, another form may require job candidates to elicit information about the passenger's trip to his or her hotel. For the validation study, AIR developed two sets of questions using these specifications.

Feedback Questionnaire

In addition to updating the selection measures, AIR also modified the feedback questionnaire. First, because the Structured Interview was not included in the validation study, AIR removed the Propriety of Questions subscale because these items pertained specifically to the HonestyIntegrity questions. Second, AIR revised some items based on BDOs' feedback. For example, the job-relatedness questions referred to the BDO job as a unitary concept (e.g., —A person who scored well on this component will be a good BDO.‖). However, BDOs indicated that some measures (e.g., the Passenger Observation Assessment) were only related to specific job duties (e.g., observing passenger behaviors). Thus, AIR modified the items on the content and predictive job-relatedness scales so that they distinguished between separate aspects of the job (e.g., —A person who scored well on this component will be good at some tasks needed for the BDO job.‖). Finally, AIR added more targeted open-ended questions to the general comment box. Specifically, BDOs in the validation study were asked —How difficult was this test/exercise/component?‖ and —How clear were the test/exercise/component instructions?‖ These questions were used to guide BDOs in providing their feedback, but they were still provided with a general comment box for use in providing recommendations for improving the measures or other relevant suggestions.

BDO Job Performance Measure

Based on the results of the pilot test, no revisions were made to the BDO JPM or rater training. However, AIR did include two additional Opportunity to Observe items. These items referred to a BDO interacting with passengers and security personnel as well as conducting referrals and passenger screening. These items were added to provide a more comprehensive evaluation of STSMs' opportunity to observe a BDO's performance.

In addition to the Opportunity to Observe items, AIR added three items to the training evaluation form. These additional items asked STSMs to provide feedback about the format of the BDO

JPM and the effectiveness of the frame-of-reference exercise. AIR included these items to conduct a more thorough evaluation of the rater training.

Sensitive Security Information (SSI)

CHAPTER V: VALIDATION STUDY

Validation Study Participants

At each site, participants were recruited through communications with their supervisor. Prior to agreeing to participate, BDOs were informed that their participation was voluntary and would not affect their standing with TSA. BDOs were also informed that their responses would be kept confidential and that results would be reported at a group-level so individual responses could not be identified. Each participating airport provided a list of volunteers for use in developing the data collection schedule. In total, 216 BDOs participated in the validation study. The final sample ($n = 214$), however, was slightly smaller.

During the validation study, AIR documented any and all administration issues (e.g., equipment failure, TSA observations) or environmental distractions (e.g., loud noises, interruptions) that occurred. For two participants, these occurrences were severe enough to warrant their removal from the data set entirely. For others, the distractions occurred only during one assessment. Table 35 below provides detail regarding the number of participants that were removed for each assessment. Each of these instances was qualitatively reviewed by AIR and decisions about their inclusion or exclusion were made by the research team.

Table 35. Number of Participants Removed for Each Assessment

Assessment	Number of Participants Removed
Wonderlic Personnel Test	12
Watson Glaser Critical Thinking Appraisal	13
Passenger Observation Assessment	6
Mental Math Assessment	11
Writing Knowledge Assessment	7
Visual Recall Assessment	19
Role-Play Exercise	17

Prior to analyzing participant data, AIR performed a number of data cleaning activities. This process consisted of quality control activities such as checking for data entry errors and examining out of range values. Additionally, AIR examined the data for each measure to check for excessive missing data and outliers, which can have a considerable negative impact on data analysis results.

Three categories of outlier data were assessed: univariate, bivariate, and multivariate outliers. Univariate outliers occur when there is an extreme value on either end of the distribution for a single variable. Bivariate and multivariate outliers are cases that have extreme values on two or more variables, respectively. AIR identified 17 univariate outliers, 31 bivariate outliers, and 1 multivariate outlier. For univariate outliers and bivariate outliers, AIR only removed the extreme values on a specific selection measure. For example, if a participant was identified as a univariate or bivariate outlier for the Passenger Observation Assessment, then only this score was removed, whereas his or her scores on the remaining selection measures were retained. AIR removed all of the selection measure scores for the one participant who was identified as a multivariate outlier.[23]

Table 36 provides frequencies and percentages for the primary airport, pay grade, and tenure for both the validation study and the BDO population. Participants had an average of 2.94 years tenure within the current position (SD = 1.41), which is quite similar to that of the BDO population (M = 2.93, SD = 1.17). Additionally, participants had been with TSA for an average of 6.81 years (SD = 2.17).

Table 36. Airport, Pay Grade, and Tenure for Validation Study Sample and BDO Population

Primary Airport	Sample (n = 214)		Population (n = 2935)	
	N	Percentage	N	Percentage
DCA – Washington, DC	13	6%	42	1%
DEN – Denver, CO	22	10%	59	2%
IAD – Washington, DC	15	7%	42	1%

[23] This participant was likely identified as a multivariate outlier due to his or her high performance ratings but low scores on multiple selection measures.

ILM – Wilmington, NC	4	2%	5	0%
JFK – New York, NY	23	11%	116	4%
LAS – Las Vegas, NV	29	14%	75	3%
LAX – Los Angeles, CA	25	12%	96	3%
MDW – Chicago, IL	17	8%	31	1%
PHL – Philadelphia, PA	17	8%	61	2%
PHX – Phoenix, AZ	23	11%	66	2%
RDU – Raleigh, NC	14	7%	25	1%
SJC – San Jose, CA	12	6%	24	1%
Pay Grade	**N**	**Percentage**	**N**	**Percentage**
F	164	77%	2035	69%
G	50	23%	629	21%
H	0	0%	222	8%
I	0	0%	49	2%
Tenure (Years)	**Mean**	**Std. Deviation**	**Mean**	**Std. Deviation**
In Current Position	2.94	1.41	2.93	1.17
With TSA	6.81	2.17	—	—

— Not Available

Table 37 presents the demographic characteristics of the validation study participants in comparison to the overall BDO population. The sample of BDOs who participated in the validation study was comparable to the BDO population in terms of gender, ethnicity, and race. Specifically, a majority of the sample was male (66%), non-Hispanic (79%), and white (49%); the BDOs in the population were mostly male (67%), non-Hispanic (83%), and white (55%). The average age of the sample ($M = 40.94$, $SD = 11.18$) was similar to the average age of BDOs in the overall population ($M = 42.59$, $SD = 11.13$).

Table 37. Demographic Data for Validation Study Sample and BDO Population

Demographics	Sample ($n = 214$)		Population ($n = 2935$)	
	N	Percentage	N	Percentage
Gender				
Female	72	34%	967	33%
Male	142	66%	1968	67%
Education Level				
High School/GED	22	10%	—	—
Some College	98	46%	—	—
Associate's Degree	28	13%	—	—
Bachelor's Degree	55	26%	—	—
Master's Degree	5	2%	—	—

Ethnicity				
Hispanic or Latino	43	20%	488	17%
Not Hispanic or Latino	169	79%	2448	83%
Other	2	1%	—	—
Race				
Asian	11	5%	142	5%
Black of African American	53	25%	655	22%
Native Hawaiian or Pacific Islander	8	4%	14	1%
Two or More Races	17	8%	4	<1%
White	105	49%	1607	55%
American Indian or Alaska Native	1	1%	25	1%
Other	10	5%	—	—

— Not Available Note. —Otherl includes participants who either wrote Hispanic (Dominican Republic), Indian, or Latino.

Ideally, the sample of BDOs would be similar to an applicant population to better simulate a typical hiring scenario. Given the voluntary nature of the study and the recent hiring freeze for the SPOT program, however, this was difficult to achieve. For instance, nearly a quarter of the participants were G-Band BDOs (23%) and the average tenure for the participants was 2.94 years (SD = 1.41). These figures suggest that the majority of BDOs who participated in the validation study were experienced job incumbents and are likely to perform differently on the selection measures than job candidates. AIR considered these differences while conducting the validation study analyses and developing recommendations.

Validation Study Method

This method section describes the procedures used for collecting data during the validation study, including the pre-site visit preparation activities, the on-site administration process, and procedures for collecting selection measures (predictor) and job performance (criteria) data. Generally, these procedures were the same as those used during the pilot test. Nonetheless, any changes that occurred between the pilot test and the validation studies are described in this section.

Pre-Site Visit Preparation

This section describes the steps AIR took to prepare for the site visits. These included identifying airports to participate, conducting trainings, and assigning test forms to participants. Each of these steps is described in more detail below.

Site Selection Process

AIR worked closely with TSA's Office of Security Operations (OSO) to identify and communicate with potential validation study sites, which were identified based on geographic location and airport size. In addition, sites who had participated in the pilot test were ineligible for inclusion in the validation study.[24] AIR collected data at 11 sites from early September 2011 to late November 2011, as outlined in Table 38.

Table 38. Location and Dates of Validation Study Site Visits

Primary Airport (Code)	Date of Visit
IAD – Washington, DC	September 13, 2012 – September 14, 2012
MDW – Chicago, IL	September 21, 2012 – September 23, 2012
RDU – Raleigh, NC	October 4, 2012 – October 6, 2012
PHX – Phoenix, AZ	September 28, 2012 – September 29, 2012
SJC – San Jose, CA	October 12, 2012 – October 13, 2012
LAX – Los Angeles, CA	October 18, 2012 – October 20, 2012
DEN – Denver, CO	October 25, 2012 – October 27, 2012
PHL – Philadelphia, PA	November 1, 2012 – November 3, 2012
LAS – Las Vegas, NV	November 9, 2012 – November 11, 2012
JFK – New York, NY	November 14, 2012 – November 17, 2012
DCA – Washington, DC	November 29, 2012 – November 30, 2012

Note. Candidates from Wilmington, NC (ILM) participated during the RDU site visit.

[24] These airports were excluded to ensure that the BDOs who participated in the validation study had not completed previous versions of the assessments. This limited the impact of confounding practice effects. Also, by recruiting airports who had not participated in the pilot test, AIR and TSA were able to include a more diverse group of BDOs throughout each phase of the study.

As was done for the pilot test, AIR conducted teleconferences with representatives from all of the participating airports before the site visits. These teleconferences were conducted using a semi-structured protocol (see Appendix P). During the teleconferences, AIR provided background information about the study, described the purpose of the validation study, outlined scheduling and room/equipment requirements, and provided criteria for recruiting study participants.

With respect to study participants, to ensure the quality of the data, the study required 20-30 BDOs per site who were willing participants. Preferably, these BDOs would have less than 6 months of experience and represent a range of performance (i.e., both high and low performers). This would create a sample that most closely resembled that of the job candidate pool. In addition, it was important to have the participation of 2 SPOT Transportation Security Managers (STSMs) per site. AIR asked for STSMs who had the opportunity to observe BDO's performance, could participate in a rater training course on use of the BDO Job Performance Measure (JPM), and could provide independent performance ratings for each participating BDO. In addition, each airport designated a points-of-contact (POC) to lead the recruitment effort at their airport and correspond directly with AIR.

Following the initial teleconferences, AIR provided each site with read-ahead materials for the POCs to disseminate to potential participants during the recruitment process (see Appendix Q). These materials described the study and participants' potential role in the validation study. In turn, each airport provided a list of BDO volunteers and information about their work schedules. AIR used this information to create draft agendas that would ensure the most efficient use of BDO and BDO manager time and limit disruption of security operations and the critical mission of passenger screening. Then, AIR and the airport POCs coordinated efforts to finalize the agendas before the data collection. This process resulted in two-. three-, four-, or five-day site visits.

Training

As with the pilot test, AIR conducted a series of training sessions prior to beginning data collection. These sessions included Administrator training, Assessor training, and Resource Person training. In general, these trainings were similar to those conducted during the pilot test. However, changes to the procedures used for the validation study are specified below.

Administrator Training

First, AIR completed internal training sessions during which Administrators practiced using the protocols and administration guides for each selection measure. These sessions were similar to those conducted during the pilot test. During each session, two or three AIR researchers practiced reading each script aloud, distributing test materials, and operating the video-based equipment (e.g., Passenger Observation Assessment video recordings).

Assessor Training

In addition to an Administrator training, AIR conducted an internal Assessor training for the Role-Play Exercise Assessors. This training was modeled after the BDO JPM rater training provided to STSMs during the pilot test (see Appendix R). This one-hour training session was designed with the following learning objectives:

- Understand the responsibilities of the Assessor during the Role-Play Exercise,
- Understand common rating errors and how to avoid them,
- Understand how to use the rating scales and Summary Statement Checklist,
- Understand how to rate a job candidate's performance using the Assessor Guide, and
- Understand how to make ratings so that they are fair, accurate, and consistent across raters.

To accomplish these goals, two AIR facilitators began by providing the Assessors with an overview of the Role-Play Exercise, including the purpose of the assessment, the administration process, and the rating materials. In addition, the facilitators discussed common rater errors,[25] in an effort to minimize the occurrence of these errors during data collection. Finally, the training concluded with a frame-of-reference exercise. This included having Assessors observe two previously recorded Role-Play Exercise performances. Assessors were then asked to rate the performances using the Elicitation of Information and Conversation Skills rating scales. Assessors shared their individual ratings and discussed any discrepancies.

Resource Person Training

AIR trained a member of the BDO National Deployment (NDO) Team to serve as the Resource Person for the Role-Play Exercise. The three-hour training session was conducted at AIR's office in Washington, DC on September 7, 2011. As with the pilot test, AIR facilitated the training using a semi-structured protocol (see Appendix S). Also, similar to the pilot test, the Resource Person completed a background questionnaire, received an overview of the Role-Play Exercise, and reviewed the instructions for the Role-Play Exercise that would be provided to BDO participants. AIR facilitators also provided specific instructions about how to perform as a Resource Person. These instructions were similar to those provided during the pilot test. Unlike the pilot-test, however, the Resource Person would only play the role of a single character (instead of two) during the validation study. AIR facilitators reviewed the background information on this character with the Resource Person. Finally, as with the pilot test, the training concluded by having the trainee review two examples of a Resource Person's

[25] Examples of rater errors are provided in the *Protocol for Collecting Job performance Data* section in Chapter IV of this report.

performance and practice as the Resource Person. After each activity, AIR facilitators provided feedback about the trainee's performance.

Assigning Test Forms

As mentioned in Chapter III, AIR developed multiple versions of the new assessments and chose the best forms of each assessment to use in the validation study. To ensure equal sample sizes across all forms, AIR created form schedules. This allowed AIR to obtain similar sample sizes for each form of each assessment. Table 39 presents the number of participants who completed each form.

Table 39. Validation Study Participant Count by Form

Assessment	Form	N
Passenger Observation Assessment	A	101
	B	112
Mental Math Assessment	A	112
	B	101
Writing Knowledge Assessment	A	112
	B	101
Visual Recall Assessment	A	110
	B	103
Role-Play Exercise	A	102
	B	109

On-Site Administration

As with the pilot test, AIR standardized the test administration processes at each site visit by using administration guides for each assessment and by following a semi-structured validation study protocol. These protocols and administration guides were similar to those used during the pilot test unless otherwise specified.

Validation Study Protocol

The Validation Study Protocol (see Appendix T) contained information for how to prepare for the validation study. The protocol outlined the equipment, test materials (e.g., user's manuals, response booklets, media files), and other materials needed to successfully conduct the validation study. The protocol also contained scripts for the group introduction session, background questionnaires, administering each of the Problem Solving measures, and the group feedback sessions. Finally, a list of frequently asked questions (FAQs) was provided (e.g., purpose of the validation study, intended use of the test scores).

Assessment Administration Guides

An administration guide was developed for each selection measure. Each guide is described below.

Work Sample Test

As with the pilot test, AIR proctored the Work Sample Test during the validation study using an administration guide. The guide was updated to reflect all changes made following the pilot test (see Chapter IV for a complete description of these changes). The guide provided general instructions for proctoring the Work Sample Test (e.g., list of necessary materials, establishing a proper testing environment), a standardized script for administering each of the four assessments, and a list of FAQs that participants may ask (e.g., how Administrators would protect confidential information).

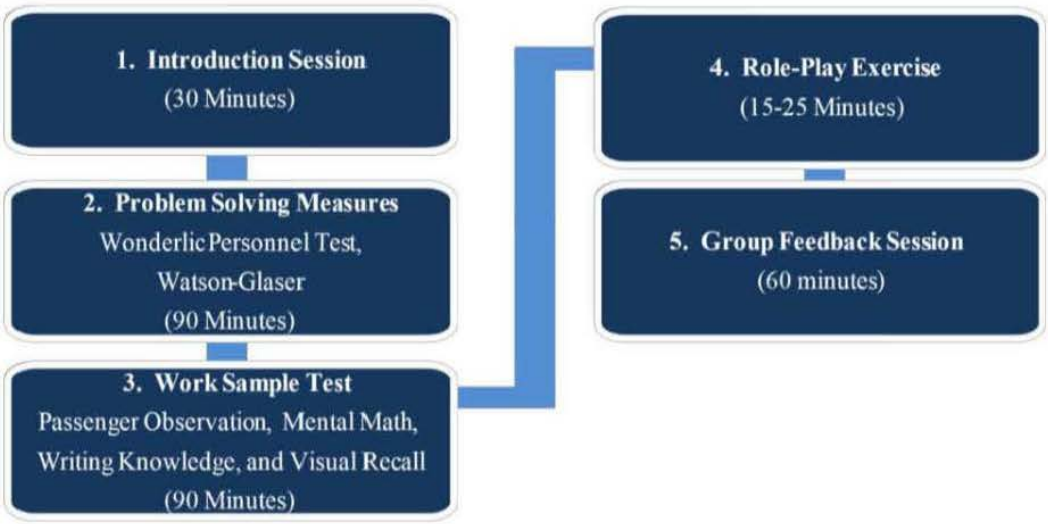
Role-Play Exercise

During the validation study, the Role-Play Exercise was also administered using a standardized administration guide, which provided Administrators with information on how to prepare for and administer this assessment. For example, the guide provided instructions for collecting the necessary materials, preparing a proper testing environment, and identifying additional personnel (i.e., Assessors and Resource Persons). The guide also included a standardized script for administering the Role-Play Exercise. Information provided to each test-taker at the beginning of the session (labeled as the Candidate Packet) was provided at the end of the guide.

Administration Process

Site visits for the validation study lasted from two to four days depending on the number of BDOs and other scheduling considerations. Figure 5 outlines the order in which the assessments were administered as well as the amount of time generally required for each. The day before each site visit began, AIR met with the POC to acquaint themselves with the facilities, set up equipment, and coordinate activities with the BDO supervisors.

Figure 5. Validation Study Administration Process



Introduction Session

As with the pilot test, each site visit began with an introduction session with large groups of BDOs (e.g., 7-12) and STSMs. Separate introduction sessions were typically conducted for each shift (i.e., AM and PM shifts). During this session, AIR informed BDOs about the purpose of the study, their role in the study, and the agenda for the site visit. AIR also reviewed the processes for ensuring the confidentiality of participants' responses and the importance of maintaining the security of the test materials. Before concluding, AIR reminded the participants about the voluntary nature of the study, provided them an opportunity to withdraw from the validation study,[31] and answered their questions.

Problem Solving Measures

Administration of the Problem Solving measures was completed by one AIR Administrator. STSMs, other TSA representatives, and other AIR personnel were dismissed from the room.[32] First, BDOs completed the Wonderlic Personnel Test. Following the test publisher's administration procedure, BDOs were given 12 minutes to complete 50 items. Next, participants were asked to complete the Watson-Glaser Critical Thinking Appraisal. For this measure, BDOs were given 30 minutes to answer 40 items. *Work Sample Test*

[31] None of the participants chose to withdraw.

[32] At this time, AIR conducted a demonstration of the Work Sample Test and Role-Play Exercise for STSMs and other TSA representatives. During these demonstrations, STSMs and TSA representatives provided general reactions about the clarity and job-relevance of the assessments.

The Work Sample Test was administered by one AIR researcher. Each of the four assessments in the Work Sample Test was administered in the following order: the Passenger Observation Assessment, the Mental Math Assessment, the Writing Knowledge Assessment, and the Visual Recall Assessment. AIR administered the Work Sample Test to groups of three to four BDOs. For the first site visit, the stimuli for the Passenger Observation and Mental Math Assessments were presented using computer monitors, whereas paper copies of the images for the Visual Recall Assessment were distributed using envelopes. However, AIR presented these images using a timed Microsoft PowerPoint® slideshow at the remaining sites. This approach improved the standardization of the measure by using automatic timing mechanisms. BDOs commented that this method of administration did not have any detrimental effect on their performance.

Role-Play Exercise

The Role-Play Exercise was administered to BDOs individually. Prior to beginning this session, AIR requested permission to record the session. The audio recordings were collected to inform subsequent changes and revisions to this assessment. For example, AIR reviewed the Role-Play Exercise performances to determine if the content of the Resource Person's script needed to be modified. All efforts were made to ensure confidentiality during the session (e.g., by not referring to the BDO by name). Two AIR Assessors evaluated each BDO's performance.

Participant Feedback

Similar to the pilot test, BDOs completed a feedback questionnaire following each assessment. As described in Chapter IV, AIR reduced the number of items on this questionnaire and added two open-response questions (see Appendix U). Questionnaires typically took 5-10 minutes to complete. In addition, AIR invited BDOs to participate in group feedback sessions during the last day of the site visits. Sessions typically included four to eight BDOs. During these sessions, AIR used a semi- structured protocol (see Appendix T) to facilitate a discussion regarding the difficulty of the measures, the clarity of the instructions, and recommendations for improving the measures. AIR used audio recorders to document the group feedback sessions. This ensured that specific recommendations were not lost during the note taking process, and they served as a resource when implementing changes to the assessment and administration guide following the validation study. Permission was obtained from each of the participants prior to beginning the session and the Administrators followed specific instructions (e.g., not referring to the BDOs by name during the session) to maintain their confidentiality. Again, each approach to collecting feedback data (i.e., administration of the feedback questionnaires and facilitation of group sessions) provided unique information. The feedback questionnaires allowed BDOs to provide

their individual reactions, whereas the feedback sessions provided a forum for participants and AIR to share and expand on ideas.

TSA Observation

Throughout the development of the selection measures, AIR sought feedback from multiple stakeholders. This was to ensure that AIR was fully capturing the aspects of the BDO job in the assessments as well as meeting the requirements of all interested parties. Prior to the final site visit, AIR obtained permission from one BDO to allow TSA representatives to observe the administration of the assessments. The representatives were simply observers in the process; however, this participant's data was ultimately flagged and removed from the final analyses.

Protocol for Collecting Job Performance Information

AIR followed procedures that were similar to the pilot test for collecting job performance data (see Chapter IV). Specifically, AIR received Performance Accountability and Standards System (PASS) data from TSA for each BDO who participated in the validation study. These data were from the previous performance period (i.e., 2011-2012) and consisted of ratings for the Performance Skills Evaluation, the Job Knowledge Test, and ratings of individual competencies (e.g., Interpersonal Skills, Team Work). Also, TSA provided a composite score that included each of these components.

For the BDO Job Performance Measure (BDO JPM), AIR identified STSMs at each airport to provide ratings for BDOs who participated in the validation study. STSMs were trained in how to use the BDO JPM during a series of teleconferences. Table 40 lists the dates of each training session and the number of STSMs who participated.

Table 40. BDO JPM Rater Training Dates and Number of Participants

Date of Training	Number of Participants
September 8, 2011	2
September 21, 2011	8
October 5, 2011	5
October 18, 2011	10
October 26, 2011	1
October 27, 2011	1

Trainings were conducted using the same protocol that was used during the pilot test. Specifically, AIR reviewed the BDO JPM, provided a description of the multidimensionality of

performance, reviewed common rater errors and biases, and facilitated a frame-of-reference exercise. Additional information about this training is provided in Chapter IV.

Validation Study Results

This section describes the results of the validation study. The section begins by presenting descriptive statistics and reliability estimates for each of the predictor (selection) and criterion (job performance) measures. Summary statistics for item-level analyses are also presented (more detailed information for these analyses is presented in Appendix V). Individual items that failed to reach the established thresholds for these analyses were reviewed to determine if they needed to be revised or removed. Next, estimates of validity (i.e., construct, criterion-related) are presented. Information concerning potential subgroup differences is presented next, followed by a discussion concerning important operational considerations (e.g., ease of implementation, resource requirements) for each measure.

Descriptive Statistics

This section presents descriptive statistics for each predictor and criterion measure. Intercorrelations among the subcomponents of the criterion measures (i.e., PASS and BDO JPM) are also presented.

Problem Solving Measures

Descriptive statistics for the Problem Solving measures are presented in Table 41. On average, BDOs scored 20.15 (SD = 5.73) on the Wonderlic Personnel Test. Based on data reported by the publisher (Wonderlic, Inc., 2002), this value is comparable to mean scores of applicants for law enforcement positions (M = 20.93, SD = 6.14). The mean value for BDOs on the Watson-Glaser Critical Thinking Appraisal was 23.66 (SD = 6.15). These values are comparable for applicants for non-management positions across industries (M = 27.70, SD = 5.90) (Pearson, Inc., 2004).

Table 41. Descriptive Statistics for the Problem Solving Measures

	Wonderlic Personnel Test		Watson-Glaser Critical Thinking Appraisal	
	BDO Sample	National Norms	BDO Sample	National Norms
N	202	1,854	201	332
Mean	20.15	20.93	23.66	27.70
Median	19.00	—	24.00	—
Standard Deviation	5.73	6.14	6.15	5.90
Range (Possible)	0 - 50	—	0 - 40	—
Minimum (Observed)	8.00	—	8.00	—

Maximum (Observed)	35.00	—	39.00	—
Skewness	0.21	—	-0.02	—
Kurtosis	-0.44	—	-0.39	—
Cronbach's Alpha	0.82	—	0.79	—
Standard Error of Measurement	2.43	—	2.82	—

— Not Available

A wide range of scores for both Problem Solving measures was observed. Specifically, scores on the Wonderlic Personnel Test ranged from 8 to 35, whereas scores on the Watson-Glaser Critical Thinking Appraisal ranged from 8 to 39. Based on the estimates of skewness and kurtosis, scores for both Problem Solving measures were normally distributed. Finally, both measures demonstrated adequate levels of internal consistency (i.e., alpha > .70).

Passenger Observation Assessment

Prior to conducting analyses for the Passenger Observation Assessment, AIR considered a number of scoring systems. Each system was evaluated based on four criteria. First, the scoring system had to reasonably represent the consequences of a BDO's job performance. For instance, overlooking SPOT behaviors and appearance factors can be a critical mistake and, therefore, making this error during the Passenger Observation Assessment should be appropriately reflected in the scoring system. Second, the scoring system had to provide adequate levels of variability. If the scores on the assessment did not vary across job incumbents, then it is unlikely that it would provide useful information during the hiring process. Third, the scoring system had to provide reliable scores. Scoring systems that consisted of items that measured the same construct (i.e., were more internally consistent) were more favorable than those that consisted of divergent items. Fourth, scoring systems were evaluated in terms of feasibility. Although it is possible that the Passenger Observation Assessment could be administered electronically, it is currently administered in a paper-and-pencil format. Thus, scoring systems that allowed Administrators to efficiently score a job candidate's performance were more favorable than less efficient approaches.

The Passenger Observation Assessment requires job candidates to observe eight passengers and indicate whether these passengers displayed six behaviors/appearance factors. Scoring for this measure could consist of a maximum of 48 items. The sections below describe the four major scoring systems that AIR evaluated while developing the Passenger Observation Assessment.

Initial Scoring System

Initially, AIR considered awarding a single point each time a test-taker correctly indicated that a passenger displayed, or did not display, a behavior/appearance factor. This scoring system is presented in Table 42.

Table 42. Initial Scoring System for the Passenger Observation Assessment

Candidate Response	Correct Response	
	Behavior/Appearance Factor is Present	Behavior/Appearance Factor is Absent
Behavior/Appearance Factor is Present	+1	0
Behavior/Appearance Factor is Absent	0	+1

Descriptive statistics and reliability estimates for both Form A and Form B are presented in Table 43.

Table 43. Descriptive Statistics for the Initial Scoring System for the Passenger Observation Assessment

	Form A	Form B
N	101	112
Mean	44.14	44.00
Median	44.00	45.00
Variance	4.52	7.03
Standard Deviation	2.13	2.65
Range (Possible)	0 - 48	0 - 48
Minimum	37.00	34.00
Maximum	48.00	48.00
Skewness	-.65	-1.25
Kurtosis	.37	1.89
Cronbach's Alpha	.48	.61
Standard Error of Measurement	1.41	1.66

Although this system reflects both the consequence of overlooking and incorrectly identifying a passenger's behaviors/appearance factors, it presents some challenges. First, it provides limited levels of variability. This can primarily be attributed to the equal weighting given to items (n = 9) where a candidate must respond to receive credit (i.e., a passenger displays a behavior/appearance factor) and items (n = 38) where a candidate can receive credit for abstaining from responding (i.e., not placing a checkmark when a passenger does not display a

Sensitive Security Information (SSI)

behavior/appearance factor). Many of the items in which a passenger does not display a behavior/appearance factor are answered correctly by 100% of the test-takers and, thus, these items would not help differentiate job candidates. Finally, the Administrators would need to score 48 separate items. This could lead to less efficiency and more scoring errors than other alternative scoring systems.

Revised Scoring System

Given the challenges with the initial scoring system, AIR considered an alternative approach. To improve the variability of test scores, this revised scoring system placed greater emphasis on the displayed behavior/appearance factor items. Specifically, test-takers received two points for correctly identifying a displayed behavior/appearance factor and one point for correctly identifying when a passenger did not display a behavior/appearance factor. This scoring system is presented in Table 44.

Table 44. Revised Scoring System for the Passenger Observation Assessment

Candidate Response	Correct Response	
	Behavior/Appearance Factor is Present	Behavior/Appearance Factor is Absent
Behavior/Appearance Factor is Present	+2	0
Behavior/Appearance Factor is Absent	0	+1

Descriptive statistics for both Form A and Form B are presented in Table 45.

Table 45. Descriptive Statistics for the Revised Scoring System for the Passenger Observation Assessment

	Form A	Form B
N	91	109
Mean	50.56	50.45
Median	51.00	51.00
Variance	9.96	12.88
Standard Deviation	3.16	3.59
Range (Possible)	0 - 57	0 - 57
Minimum	44.00	40.00

Maximum	57.00	57.00
Skewness	-.16	-.68
Kurtosis	-.55	-.05
Cronbach's Alpha	.46	.52
Standard Error of Measurement	2.32	2.49

This system helped improve the variability of the measures, but it still featured a number of items that did not contribute to the variability of the overall test score because they were answered correctly by 100% of the test-takers. Also, this system could require more time for Administrators to score a candidate's performance because they would need to verify a candidate's responses, determine what each item is worth (i.e., one or two points), and provide the appropriate score for each item based on a candidate's responses. Although the efficiency of this process could be improved through the use of a scoring key, this type of scoring system increases the likelihood of scoring errors.

Testlet-based Scoring System

Some researchers have suggested that it could be beneficial to group interrelated items and develop sub-scores (Wainer & Lewis, 1989). These groupings are often called testlets and can reduce the impact of item-order effects and help balance the content of items across a test. As a test of this approach, AIR rationally grouped the items across passengers and developed testlet scores. Specifically, each of the six items for a single passenger was summed to create a testlet sub-score. Grouping items across passengers, rather than across behaviors/appearance factors, was the most logical choice because the impact of contextual factors (e.g., the pace of the passenger screening line) would be the most consistent within a single passenger.

Minimum scores for each passenger testlet were 0 whereas maximum values could range from 6 to 9. Maximum values for each testlet depended on the number of behaviors/ appearance factors that a passenger displayed. Descriptive statistics for each passenger testlet are presented in Table 46.

Table 46. Descriptive Statistics for the Passenger-based Testlet Scoring System for the Passenger Observation Assessment

	Form A							
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**
N	95	95	95	95	95	95	95	95
Mean	6.81	5.88	6.87	5.79	6.38	6.22	5.66	6.78
Median	7.00	6.00	6.00	6.00	7.00	7.00	5.00	7.00

Sensitive Security Information (SSI)

Variance	.43	.12	1.43	.17	3.58	1.20	1.06	.49
Std. Deviation	.66	.35	1.20	.41	1.89	1.09	1.03	.70
Range (Possible)	0 - 7	0 - 6	0 - 8	0 - 6	0 - 9	0 - 7	0 - 7	0 - 7
Minimum	4.00	4.00	3.00	5.00	2.00	3.00	4.00	4.00
Maximum	7.00	6.00	8.00	6.00	9.00	7.00	7.00	7.00
Skewness	-3.46	-3.15	-.70	-1.44	-.27	-1.15	.36	-3.05
Kurtosis	10.93	10.11	.40	.08	-.67	.25	-1.45	8.05
Cronbach's Alpha	.32	.11	.19	-.19	.35	.17	.01	.28
Standard Error of Measurement	.54	.33	1.08	.45	1.52	.99	1.03	.59
	Form B							
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**
N	112	112	112	112	112	112	112	112
Mean	6.70	5.75	6.12	6.51	6.11	6.79	6.22	6.06
Median	6.00	6.00	7.00	7.00	7.00	7.00	7.00	6.00
Variance	2.05	0.26	1.36	0.83	1.36	0.34	1.27	1.57
Std. Deviation	1.43	0.51	1.17	0.91	1.17	0.59	1.13	1.25
Range (Possible)	0 - 8	0 - 6	0 - 7	0 - 7	0 - 7	0 - 7	0 - 7	0 - 8
Minimum	2.00	4.00	3.00	4.00	2.00	4.00	2.00	3.00
Maximum	8.00	6.00	7.00	7.00	7.00	7.00	7.00	8.00
Skewness	-0.93	-1.96	-1.02	-1.49	-0.91	-3.19	-1.29	-0.20
Kurtosis	0.50	3.07	-0.04	0.62	-0.15	10.29	0.98	0.25
Cronbach's Alpha	.15	.16	.24	.13	.24	.21	.25	.03
Standard Error of Measurement	1.32	.47	1.02	1.09	1.02	.52	.98	1.23

Note. P = Passenger.

Although a testlet-based scoring system may improve the consistency of content and reduce item-order effects, it presented some challenges in the case of the Passenger Observation Assessment. First, many testlets displayed levels of internal consistency that were much lower than those observed with other scoring approaches. In addition, many of the testlets did not show improved levels of variability. This is primarily due to the number of items within some

Sensitive Security Information (SSI)

testlets that were answered correctly by a large number of test-takers. Finally, the testlet-based scoring system would require a significant amount of time to score because Administrators would need to score each item and then develop subtotals for each passenger. This process also introduces additional instances where a scoring error may occur.

Displayed Behavior Scoring System

The final scoring system that AIR considered emphasized a job candidate's ability to observe a passenger's displayed behaviors and appearance factors. Specifically, this scoring system consisted of nine items in which a passenger displayed a behavior/appearance factor during the video recording. Each item was worth two points. Items in which a passenger did not display a behavior were not scored within this system. Descriptive statistics for both Form A and Form B are presented in Table 47.

Table 47. Descriptive Statistics for the Displayed Behaviors/Appearance Factors Scoring System for the Passenger Observation Assessment

	Form A	Form B
N	92	110
Mean	12.30	12.52
Median	12.00	14.00
Standard Deviation	3.09	3.01
Range (Possible)	0 - 18	0 - 18
Minimum	0.00	2.00
Maximum	18.00	18.00
Skewness	-.75	-.87
Kurtosis	1.75	.81
Cronbach's Alpha	0.44	0.31
Standard Error of Measurement	2.31	2.24

This system provided a number of benefits. First, this system emphasizes a candidate's ability to observe pre-determined behaviors/appearance factors. Because this is a critical component of the SPOT process, this seemed to be a reasonable representation of job-related consequences. Second, by awarding candidates two points for each displayed behavior/appearance factor, the system improved the variability of scores while removing items that do not provide any

differentiating information (i.e., those that are answered correctly by all test-takers). For example, using this scoring system, means ranged from 12.30 (*SD* = 3.09) to 12.52 (*SD* = 3.01) across the two forms. These mean values, in conjunction with the standard deviation, suggested an adequate level of variability. Third, the system resulted in estimates of internal consistency that were comparable to, if not better than, the alternative approaches. Nevertheless, both forms of the Passenger Observation Assessment showed low levels of internal consistency (0.31-0.44). The observed levels of internal consistency were similar to what has been seen with other highfidelity, work sample measures (Lievens & Coetsier, 2002). Also, this component of the Work Sample Test consisted of items with heterogeneous features (e.g., different passengers, behaviors). Thus, a more appropriate assessment of reliability may be a test-retest coefficient. Given the parameters of the current study, it was not feasible to assess this type of reliability.[26] Fourth, the system consists of fewer items than the other approaches. This improves the efficiency of the scoring process and limits the likelihood of scoring errors. Given these benefits, AIR used this scoring system while conducting subsequent analyses and when establishing the preliminary qualifying scores for the Passenger Observation Assessment.

Using the displayed-behaviors scoring system, both forms showed moderate levels of item difficulty (see Table 48). Specifically, on average across items, 68% to 70% of participants responded to the items correctly. These average item-difficulty values fell within the prescribed thresholds set for this study (i.e., less than .95, but greater than .20).

Table 48. Item-Level Statistics for the Passenger Observation Assessment

Passenger Observation Assessment	Mean Item Difficulty	Mean Corrected Item-Total Correlations
Form A	0.68	0.21
Form B	0.70	0.17

The average item-total correlations for both forms (.16 to .21) exceeded the established threshold for this study (i.e., .10) and suggested that most items were contributing to the total score. These results suggested that most items were functioning appropriately.

[26] Test-retest reliability could be assessed by administering the same test form to job applicants or job incumbents on multiple occasions. However, administering both forms of the assessment to participants over an extended period of time would have significantly increased the duration of each site visit.

Mental Math Assessment

Scores on the Mental Math Assessment were somewhat elevated (see Table 49). The average scores were 15.86 for Form A and 16.67 for Form B. However, based on the standard deviations (3.51, 2.58, respectively) there still appears to be an adequate level of variability across forms.

Table 49. Descriptive Statistics for the Mental Math Assessment

	Form A	Form B
N	111	92
Mean	15.86	16.67
Median	17.00	17.00
Standard Deviation	3.51	2.58
Range (Possible)	0 – 20	0 – 20
Minimum	7.00	10.00
Maximum	20.00	20.00
Skewness	-.98	-.85
Kurtosis	.21	.09
Cronbach's Alpha	0.79	0.82
Standard Error of Measurement	1.59	1.11

Internal consistency for both forms of the assessment exceeded the threshold of .70. Also, the standard errors of measurement for both forms (1.59 and 1.11) suggested adequate levels of precision. Specifically, if either form of the Mental Math Assessment was administered to a candidate 100 times, 65% of the time that individual's test score would fall within a range of approximately 3 points for Form A and 2 points for Form B. These narrow ranges suggest that scores on the Mental Math Assessment are precise and reliable.

In terms of item-level analyses, the average item difficulty for both forms (Form A – 80%, Form
B – 81%) indicated that, for the average item, most participants responded correctly (see Table 50). This level of item difficulty suggests that, on average, the items may not have been as challenging as those in other measures. This finding may be attributed to how often BDOs use the abilities assessed by this component. Specifically, BDOs compute mental arithmetic many times throughout each workday. Thus, it is likely that the items would be more challenging for a sample of job candidates. Nevertheless, these item-difficulty levels are still well below the established upper threshold (95%) for this study.

Table 50. Item-Level Statistics for the Mental Math Assessment

Mental Math Assessment	Mean Item Difficulty	Mean Corrected Item-Total Correlations
Form A	0.80	0.36
Form B	0.81	0.40

The average corrected item-total correlation for both forms (.36, .40) exceeded the established threshold (.10) for this study and indicated that a majority of items are contributing to the measure's total score. This result suggested that most items are functioning appropriately and can be retained.

Writing Knowledge Assessment

Means for Form A (M = 14.82) and B (M = 15.57) on the Writing Knowledge Assessment suggest that the measure was moderately difficult (see Table 51). Both forms of this measure also showed a wide range of scores. Specifically, scores from Form A ranged from 6 to 24, whereas scores on Form B ranged from 7 to 24. Also, based on the estimates of skewness and kurtosis, scores from both forms of the Writing Knowledge Assessment were normally distributed. Finally, both Writing Knowledge forms were adequately reliable with Cronbach's Alphas exceeding the common threshold of .70.

Table 51. Descriptive Statistics for the Writing Knowledge Assessment

	Form A	Form B
N	101	92
Mean	14.82	15.57
Median	15.00	15.00
Standard Deviation	4.37	3.50
Range (Possible	0 - 25	0 - 25
Minimum	6	7
Maximum	24	24
Skewness	.25	.13
Kurtosis	-.62	-.36
Cronbach's Alpha	.81	.76
Standard Error of Measurement	1.90	1.71

Sensitive Security Information (SSI)

The items on the Writing Knowledge Assessment appeared to be moderately difficult (see Table 52). On average, across items, 58% and 61% of respondents responded correctly. This finding suggested that most items were neither too easy nor too difficult.

Table 52. Item-Level Statistics for the Writing Knowledge Assessment

Writing Knowledge Assessment	Mean Item Difficulty	Mean Corrected Item-Total Correlations
Form A	0.58	0.34
Form B	0.61	0.29

The average corrected item-total correlation for both forms (.34, .29) indicates that a majority of items are contributing to the measure's total score. This suggests that most items are functioning appropriately and can be retained.

Visual Recall Assessment

The Visual Recall Assessment consists of a series of multiple-choice items. Based on the results of the validation study, AIR identified items that were (1) tricky, (2) overly difficult, or (3) did not contribute to the total score (i.e., demonstrated low item-total correlations). Given these criteria, AIR removed 6 items from Form A and 10 items from Form B. This rate of item retention is common during the development of multiple-choice measures and is why AIR created a large item bank during the initial phases of development (Haladyna, 2004). The items that were removed are listed in Table 53 below.

Table 53. Items Removed from the Revised Visual Recall Assessment

Form A	Form B
Image 1 Item 1	Image 1 Item 1
Image 1 Item 4	Image 1 Item 3
Image 1 Item 9	Image 1 Item 5
Image 2 Item 5	Image 1 Item 6
Image 2 Item 6	Image 1 Item 9
Image 2 Item 11	Image 1 Item 10
	Image 1 Item 13
	Image 1 Item 15
	Image 2 Item 5
	Image 2 Item 12

If TSA uses the Visual Recall Assessment in future studies, it may be helpful to include these items for experimental purposes. Specifically, TSA could collect additional data for each item and determine which items are functioning appropriately with a larger sample of test-takers.

For the revised version of the Visual Recall Assessment, Form A consisted of 18 items whereas Form B consisted of 17 items. Each correct item is worth 1point. Total scores represent a percentage of the number of items a BDO responded to correctly (i.e., the number answered correctly divided by the total score) and could range from 0 to 1.00. Descriptive statistics for this assessment are presented in Table 54. Both forms of the revised Visual Recall Assessment were moderately difficult with mean total scores of .74 out of 1.00. Also, both forms demonstrated adequate levels of variability with scores ranging from .26 to 1.00 on Form A and .18 to 1.00 on Form B. Scores for Form A were normally distributed, whereas scores on Form B were negatively skewed with most values clustering at the higher end of the distribution.

Table 54. Descriptive Statistics for the Visual Recall Assessment

	Form A	Form B
N	99	95
Mean	.74	.74
Median	.74	.76
Standard Deviation	.13	.16
Range (Possible)	0 - 1.00	0 - 1.00
Minimum	.26	.18
Maximum	1.00	1.00
Skewness	-.51	-1.00
Kurtosis	.95	1.39
Cronbach's Alpha	.53	.63
Standard Error of Measurement	.09	.10

The internal consistency for both Visual Recall forms (α= .53 and .63) was less than the common threshold for establishing reliability (i.e., α= .70). Visual Recall items were designed to capture a variety of components of the images (e.g., passengers versus background). This may have contributed to the moderate levels of internal consistency as the items were less likely to coalesce around a single topic. Also, as with the Passenger Observation Assessment, the Visual Recall

Assessment relies on high-fidelity stimuli. Researchers suggest that these types of measures may be less likely to be internally consistent and their reliability may best be assessed using test-retest estimates (Lievens & Coetsier, 2002; Lievens & Patterson, 2011). Because such a design requires both forms to be administered to a single BDO, this approach would have required a significant amount of additional resources. In addition, because incumbents would have been exposed to both forms of the measure, this would have limited the Visual Recall Assessment's test security.

All three forms showed an adequate level of item difficulty (see Table 55). On average, across items, approximately three quarters of participants (73% and 75%) responded to the item correctly. Although these levels are a bit higher than a desirable level of 50%, they are also far below the established threshold of 95%. Nevertheless, individual items that were identified as too easy or too difficult were reviewed and considered for revision or removal.

Table 55. Item-Level Statistics for the Visual Recall Assessment

Visual Recall Assessment	Mean Item Difficulty	Mean Corrected ItemTotal Correlations
Form A	0.75	0.18
Form B	0.73	0.24

The average corrected item-total correlation for both forms (.18, .24) indicates that a majority of items are contributing to the measure's total score. These values are above the threshold of .10 established for this study.

Item distractor analyses identified items that feature incorrect response options that were either (1) misleading or (2) not chosen by any participants. Items that were flagged during these analyses were considered for removal or revised accordingly.

Role-Play Exercise

AIR considered a number of approaches for evaluating a BDO's performance on the Role-Play Exercise. Scores on the Role-Play Exercise could be comprised of three components: two behavioral rating scales (i.e., Elicitation of Information and Conversation Skills) and an objective measure (i.e., the Summary Statement Checklist). Based on the results of the validation study, two scoring components were excluded from the composite score: the Conversation Skills rating scale and the third question from the Summary Statement Checklist.

First, the Conversation Skills rating scale did not appear to contribute unique predictive information. In other words, a participant's rating on this scale did not correspond with total scores on the PASS (b)(3):49 U.S.C. § 114 [27] or average ratings on the BDO Job Performance Measure (b)(3):49 U.S.C. § 114 The Elicitation of Information rating scale, on the other hand, was a significant predictor of both PASS Total Scores (b)(3):49 U.S.C. § 114 and average ratings on the BDO JPM (b)(3):49 U.S.C. § 114(r) Overall, the two rating scales were moderately correlated (b)(3):49 U.S.C. § 114 suggesting that they were unique, but shared some predictive variance. Thus, the observed differences in criterion-related validity may be attributed to the different content of the rating scales. Specifically, the Conversation Skills rating scale focused on less observable aspects of the test-takers' performance (e.g., a participant's demeanor, rapport building) whereas the Elicitation of Information rating scale emphasized more objective behavior (e.g., questioning techniques, presence of extended silences). If TSA chooses to include the Role-Play Exercise in future studies or selection systems, it may be helpful to include both rating scales for experimental purposes. Specifically, TSA could include the Conversation Skills rating scale, collect additional data with a larger sample of job candidates or incumbents, and determine if this measure demonstrates more acceptable levels of validity. Given its emphasis on a candidate's ability to engage a passenger in a non-threatening manner, this scale would likely provide content-relevant information. However, it would be prudent to gather additional psychometric information before including it as part of the Role-Play Exercise's total score.

In addition to removing the Conversation Skills rating scale for the analyses, AIR removed the third question on the Summary Statement Checklist from the Role-Play Exercise total score. This question was removed because it functioned differently across forms. Specifically, 88% of participants responded to the item correctly on Form A whereas only 54% answered it correctly on Form B. The number of participants that responded to each question correctly on each form is presented in Table 56.

Table 56. Number of Participants who Correctly Responded to the Summary Statement Checklist Questions

Summary Statement Checklist	Form A		Form B	
	N	Percentage	N	Percentage

[27] Ranges represent correlation coefficients across Forms A and B.

Question 1	81	86.2%	90	90.9%
Question 2	81	86.2%	86	86.0%
Question 3	83	88.3%	54	54.0%

The differences observed in Question 3 of the Summary Statement Checklist across forms could be attributed to the type of information candidates are required to elicit for each. Specifically, Question 3 on Form B may have been especially challenging because it required job candidates to elicit and retain two distinct pieces of information. Though Question 3 on Form A also required job candidates to elicit two pieces of information, the two-part answer required of Form A was less distinct, and thus may have been easier to remember as a set, than the two-part answer required for Form A.[28] Because these form differences could influence a candidate's performance and subsequent hiring decisions, AIR chose to remove the third question from the total score for both forms based on these results. In the post-validation version of the Role-Play Exercise, AIR modified the Resource Person's script in order to make the forms of this measure more parallel. Specifically, rather than require candidates for Form B of the exercise to identify two activities, they only need to identify one. Ideally, this will make this item less challenging and more similar to its counterpart on Form A.

Given these considerations, a candidate's score on the Role-Play Exercise consisted of two components: the Elicitation of Information rating scale and the first two questions from the Summary Statement Checklist. Total scores for this exercise range from 0 to 7. Descriptive statistics for the revised Role-Play Exercise scoring system are presented in Table 57. Although most BDOs scored highly on the Role-Play Exercise, with mean ratings of 5.75 on Form A (*SD* = 1.15) and 5.81 Form B (*SD* = 1.16), scores showed adequate levels of variability. Ratings for each form depicted a wide range of performance, with scores ranging from 3 to 7 on Form A and from 2 to 7 on Form B.

Table 57. Descriptive Statistics for the Role-Play Exercise

	Form A	Form B

[28] More specific information about the content of the questions has not been provided in this technical report to maintain the security of the test materials.

N	92	100
Mean	5.75	5.81
Median	6.00	6.00
Standard Deviation	1.15	1.16
Range (Possible)	0 - 7	0 - 7
Minimum	3.00	2.00
Maximum	7.00	7.00

Elevated scores were expected for this exercise because BDOs perform duties associated with this measure on a daily basis. For example, BDOs regularly engage passengers in conversation while passengers progress through the screening line. These conversations, like the Role-Play Exercise, are meant to elicit specific pieces of information in a non-threatening, casual manner.

For the Role-Play Exercise, AIR calculated measures of interrater agreement, including both kappa and weighted kappa estimates. These values indicate the degree to which raters provide the same rating above and beyond what one would expect purely by chance. Kappa values represent exact agreement (i.e., choosing the same rating), whereas weighted kappa values represent relative agreement (i.e., choosing similar ratings). Higher kappa values indicate higher levels of agreement. Kappa and weighted kappa values for different rater pairs are presented in Table 58 by each form of the Role-Play Exercise.

Table 58. Kappa and Weighted Kappa Estimates for the Role-Play Exercise

Rater Pair	Form A			Form B		
	Kappa	Weighted Kappa	N	Kappa	Weighted Kappa	N
Rater 1 & Rater 2	0.11	0.34	10			
Rater 1 & Rater 4	0.00	0.00	6	-0.19	-0.21	10
Rater 2 & Rater 3	0.77**	0.71*	6	0.22	0.36*	20
Rater 2 & Rater 4	0.00	0.00	2	-0.25	0.00	6
Rater 2 & Rater 5	-0.25	-0.13	6	0.11	0.32*	24
Rater 2 & Rater 7	0.04	-0.05	6			
Rater 2 & Rater 8	0.28**	0.46**	26	0.30*	0.51**	24
Rater 3 & Rater 8	0.19	0.49**	28	0.33**	0.57**	34
Rater 4 & Rater 5	0.14	0.16	6			
Rater 4& Rater 7	0.50	0.50	4	0.33	0.50	2

Sensitive Security Information (SSI)

Rater 4 & Rater 8	0.30	0.41	8	0.00	0.00	2
Rater 5 & Rater 7	0.43	0.50	6	-0.44	-0.50	6
Rater 6 & Rater 7	0.38**	0.50**	20	0.13	0.32**	32
Rater 6 & Rater 9	0.32**	0.51**	44	0.21	0.43**	20
Rater 7 & Rater 8	0.00	0.00	2	0.02	0.36	8
Rater 7 & Rater 9	0.51**	0.37*	8	0.41*	0.45*	12

Note. N = Number of shared observations. * $p < .05$, ** $p < .01$.

As seen in Table 58, the kappa and weighted kappa values varied across the rater pairs and forms of the Role-Play Exercise. However, this variability may be attributed to sampling error and the low number of shared observations for many of the rater pairs. Specifically, of the 32 possible combinations of raters (i.e., 16 rater pairs for each form), only 11 pairs had more than 10 shared observations. To obtain stable estimates of interrater agreement, AIR calculated interrater agreement averages using only the rater pairs that had at least 10 shared observations. For example, Rater 2 and Rater 8 were included in the analyses because they shared 26 observations on Form A and 24 observations on Form B. As can be seen in Table 59, within and across forms, average weighted kappa values were greater than or equal to .43.

Table 59. Interrater Agreement for the Role-Play Exercise

Rater Pair	Average Kappa	Average Weighted Kappa	N
Across Forms	.22	.45	13
Form A	.27	.49	5
Form B	.17	.43	8

Note. N = Number of shared observations. * $p < .05$, ** $p < .01$.

Typically, kappa values of .40 represent moderate levels of agreement (Landis & Koch, 1977). Given the complexity of the Role-Play Exercise performances, the weighted kappa values suggest that raters demonstrated adequate levels of agreement. **PASS**

Descriptive statistics for the PASS measures are presented in Table 60. (b)(3):49 U.S.C. § 114(r)

(b)(3):49 U.S.C. § 114(r)

52

(b)(3):49 U.S.C. § 114(r)

Table 60. Descriptive Statistics for the PASS Measures

(b)(3):49 U.S.C. § 114(r)

Table 61. Intercorrelations among the PASS Measures

(b)(3):49 U.S.C. § 114(r)

BDO Job Performance Measure (JPM)

This section presents the descriptive statistics for the validation study BDO Job Performance Measure developed for this study and the accompanying Opportunity to Observe items.

Descriptive Statistics

Nearly all of the performance raters (i.e., provided performance ratings of BDOs) were STSMs (total $n = 34$), [29] were an average of 47.6 years of age ($SD = 11.23$ years), and had worked for TSA for an average of 8.2 years ($SD = 1.69$ years). Twenty of the raters were male, and twentyfive of the raters identified themselves as White.

The descriptive statistics for each of the nine dimensions of the BDO Job Performance Measure is presented in Table 62. Each performance dimension was rated on a five-point scale, such that

[29] One rater indicated that he or she was an AFSD and another indicated that he or she was an ETSO, G-Band. Because AIR specifically recruited STSMs to provide performance ratings, these responses were likely errors.

the minimum rating for any given performance dimension was a one and the maximum rating a five. The mean ratings for each performance ranged from 3.96 (*Documentation*) to 4.41 (*Preparedness and Dutifulness*). Across all performance dimensions the means were slightly negatively skewed, with the greatest skewness (and mean value across all raters) for Dimension 9 (*Preparedness and Dutifulness*). The range of scores for each performance dimension indicate that raters used most of the scale points to make their ratings, but were inclined to make slightly higher (more positive) ratings of performance while infrequently utilizing lower scale points. Overall, the descriptive statistics for the BDO JPM were similar across the pilot and validation studies. As was done with the pilot test data, the composite BDO JPM was created by averaging across all nine performance dimensions. The descriptive statistics for this composite are also presented in Table 62. The internal consistency of this composite measure (α= .92) was both high and similar to that observed in the pilot test.

Table 62. Descriptive Statistics for the BDO Job Performance Measure

	Observation and Assessment of Passenger Behavior	Assignment and Calculation of Points	Referral of Passengers	Interaction with Passengers	Cooperation and Communication with SPOT Teammates
N	189	187	180	182	189
Mean	4.05	4.20	4.00	4.04	4.17
Median	4.00	4.00	4.00	4.00	4.00
Variance	.59	.59	.76	.56	.69
Standard Deviation	.77	.77	.87	.75	.83
Minimum	2.00	2.00	1.00	2.00	2.00
Maximum	5.00	5.00	5.00	5.00	5.00
Skewness	-.20	-.61*	-.38*	-.40*	-.65*
Kurtosis	-1.01*	-.52	-.93*	-.50	-.64
	Interaction with Other Security Personnel	**Documentation**	**Search of Accessible Property and Review of Travel Documentation**	**Preparedness and Dutifulness**	**Composite BDO JPM**
N	189	187	181	192	192
Mean	4.07	3.96	4.23	4.41	4.11
Median	4.00	4.00	4.00	4.50	4.17

Variance	.61	.64	.50	.52	.37
Standard Deviation	.78	.80	.71	.72	.60
Minimum	2.00	1.00	2.00	1.50	2.61
Maximum	5.00	5.00	5.00	5.00	5.00
Skewness	-.37*	-.32	-.51*	-.94*	-.38*
Kurtosis	-.81*	-.60	-.66	-.17	-.79*

For the pilot test, raters were asked to complete three items which assessed their —Opportunity to Observel each ratee's job performance. For the validation study, two additional items were added (—Interaction with passengers and other security personnell and —Performance referring a passenger and conducting passenger screeningl) to more fully reflect each of the nine dimensions of the BDO JPM. As with the pilot study, raters responded to each item using a fivepoint scale such that the minimum rating for any given item was a one (*almost never had the opportunity to observe*) and the maximum rating a five (*very frequently had the opportunity to observe*).

The descriptive statistics for each of the Opportunity to Observe items, as well as a composite measure of these items, are presented in Table 63, below. Similar to the pilot test, the mean and median values for these items indicate that, overall, most raters had a significant opportunity to observe ratees prior to making performance ratings. However, the range of ratings for these items, and specifically the minimum ratings, indicate that some raters may have had very little opportunity to observe the performance of some BDOs. The internal consistency among these dimensions (α= .92) was similar to the internal consistency for the pilot test (α= .93).

Table 63. Descriptive Statistics for the BDO Manager Opportunity to Observe Measure

	Performance Conducting SPOT	Written Reports and Other Documentation	Knowledge of SPOT and the SPOT Standard Operating Procedure (SOP)	Interaction with Passengers and Other Security Personnel	Performance Referring a Passenger and Conducting Passenger Screening	Composite Opportunity to Observe Measure
N	192	192	192	192	189	192
Mean	4.27	4.18	4.21	4.18	3.83	4.13
Median	4.50	4.50	4.00	4.00	4.00	4.20
Variance	.55	.81	.48	.61	.76	.45
Standard Deviation	.74	.90	.69	.78	.87	.67

Sensitive Security Information (SSI)

Minimum	2.00	1.00	2.50	2.00	1.00	2.20
Maximum	5.00	5.00	5.00	5.00	5.00	5.00
Skewness	-.75*	-1.17*	-.39*	-.75*	-.75*	-.74*
Kurtosis	-.26	1.54*	-.74*	-.14	.65	-.119

Note. * $p < .05$.

Table 65, below, presents the intercorrelations among the nine performance dimensions of the BDO JPM, the five Opportunity to Observe items, the composite BDO JPM, and the composite measure of Opportunity to Observe items. Performance dimensions generally showed moderate to strong intercorrelations which ranged from .34 to .78. The highest correlations were found among performance dimensions that encompass similar behaviors. For example, Dimension 1 (*Observation and Assessment of Passenger Behavior*) and Dimension 2 (*Assignment and Calculation of Points*) were strongly correlated ($r = .78$). Weaker correlations were found among performance dimensions with more divergent behaviors. For instance, Dimension 4 (*Interaction with Passengers*) and Dimension 9 (*Preparedness and Dutifulness*) were weakly

Sensitive Security Information (SSI)

correlated (r = .38). Similar to the results of the pilot study, these validation study results indicate that although some performance dimensions are highly correlated, each dimension accounts for a unique portion (i.e., variance) of the overall performance domain. This is also evident by examining the correlations between each performance dimension and the BDO JPM composite – these correlations ranged from .87 (Dimension 1, *Observation and Assessment of Passenger Behavior*) to .66 (Dimension 9, *Preparedness and Dutifulness*).

Information concerning rater agreement is presented in Table 64, below. Most pairs of raters (13 of 17) demonstrated at least moderate levels of exact agreement (i.e., weighted kappa) for their performance ratings, and many of the kappa and weighted kappa values were significant at $p < .05$. Kappa values ranged from -.08 to 1.00 whereas weighted kappa ranged from -.03 to 1.00. Generally, very high interrater agreement was observed for pairs of raters that only provided ratings for a small number of BDOs.

Table 64. Kappa and Weighted Kappa Values for Job Performance Ratings

Rater Pair	N	Kappa	Weighted Kappa
Rater 1 & Rater 25	81	-.08	.02
Rater 3 & Rater 9	54	-.03	-.03
Rater 4 & Rater 11	45	.03	.16
Rater 4 & Rater 14	27	.24*	.26**
Rater 5 & Rater 7	63	1.00**	1.00**
Rater 5 & Rater 23	72	1.00**	1.00**
Rater 9 & Rater 27	18	.60*	.60*
Rater 9 & Rater 33	9	1.00**	1.00**
Rater 11 & Rater 14	54	.45**	.58**
Rater 11 & Rater 21	45	.18	.35**
Rater 12 & Rater 28	315	.13**	.25**
Rater 13 & Rater 31	71	.26**	.44**
Rater 16 & Rater 33	63	.14	.18*
Rater 16 & Rater 35	72	.08	.10
Rater 18 & Rater 34	72	.35**	.36**
Rater 24 & Rater 29	45	.96**	.96**
Rater 26 & Rater 32	72	.35**	.41**

Note. N = Number of shared observations, * p <.05, ** p <.01.

Overall, interrater agreement was similar to the agreement observed during the pilot test, although some pairs of raters failed to meet statistically significant agreement levels. With regard to the recommended threshold of .40 (Landis & Koch, 1977), six of the seventeen rater pairs exceeded this threshold for the weighted kappa values, and eight of the seventeen rater pairs exceeded this threshold for the weighted kappa values. Using this threshold, rater agreement was moderate across all pairs of raters.

Table 65. Intercorrelations among Dimensions of the BDO Job Performance Measure and the Opportunity to Observe Items

BDO Job Performance Measure - Dimensions	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Observation and Assessment of Passenger Behavior															
2. Assignment and Calculation of Points	.78														
3. Referral of Passengers	.75	.72													
4. Interaction with Passengers	.60	.57	.60												
5. Cooperation and Communication with SPOT Teammates	.50	.46	.34	.43											
6. Interaction with Other Security Personnel	.61	.58	.53	.74	.60										
7. Documentation	.59	.58	.52	.46	.50	.56									
8. Search of Accessible Property and Review of Travel Documentation	.64	.64	.58	.52	.48	.57	.60								
9. Preparedness and Dutifulness	.54	.50	.49	.38	.51	.41	.48	.47							
10. Performance Conducting SPOT*	.32	.24	.34	.25	.18	.26	.24	.39	.31						
11. Written Reports and Other Documentation*	.35	.28	.36	.28	.10	.26	.29	.30	.25	.56					
12. Knowledge of SPOT/SPOT Standard Operating Procedure (SOP)*	.36	.27	.26	.18	.23	.19	.25	.32	.29	.67	.66				
13. Interaction with Passengers and Other Security Personnel (e.g., LEOs)*	.35	.25	.30	.30	.25	.31	.25	.40	.40	.78	.59	.73			
14. Performance Referring a Passenger and Conducting Passenger Screening*	.33	.20	.31	.24	.27	.21	.20	.31	.28	.60	.62	.67	.68		
15. BDO JPM Composite	.87	.83	.80	.77	.71	.81	.75	.78	.66	.37	.37	.36	.44	.35	
16. Opportunity to Observe Composite	.43	.32	.40	.35	.28	.31	.33	.41	.34	.84	.82	.87	.88	.84	.45

Note. N = 182-192, ** p <.01. * Denotes an Opportunity to Observe item.

Overall Performance Composite

Average ratings across the dimensions of the BDO Job Performance Measure and the PASS Total Score were moderately correlated ($r = .45$, $p < .001$). Given the modest amount of overlap between these two measures, it is likely that they are assessing two distinct aspects of performance. Thus, AIR developed a composite performance measure using these two criterion measures for exploratory analyses. Because the PASS Total Score uses a much larger distribution (i.e., 0 to 105) than the BDO JPM (i.e., 1 to 5), AIR standardized the two criteria by calculating z-scores. This transformed both measures so they used the same scale (i.e., a mean of 0 and a standard deviation of 1). This also helped ensure that the Overall Performance Composite would not be overly influenced by the composite with the greater variability (i.e., the PASS Total Score). AIR then averaged the two z-scores in order to calculate the Overall Performance Composite. Descriptive statistics for the Overall Performance Composite are presented in Table 66.

Table 66. Descriptive Statistics for the Overall Performance Composite

	Overall Performance Composite
N	191
Mean	.02
Median	.09
Standard Deviation	.86
Minimum	-2.60
Maximum	1.55
Skewness	-.53
Kurtosis	-.03

The Overall Performance Composite was slightly negatively skewed with scores clustering around the higher end of the distribution. Because this is a sample of job incumbents who have previously been screened and undergone on-the-job training, it is expected that most performance ratings would be higher than average. Nevertheless, the Overall Performance Composite demonstrated adequate levels of variability ($SD = .86$) with scores ranging from -2.60 to 1.55.

Validity Evidence

The validity of a selection measure is based on evidence demonstrating the appropriateness of making inferences about a candidate's suitability for a job using test scores (Guion, 1998; Putka & Sackett, 2010). Although validity is best considered a unitary concept, researchers often specify three different types. These include: content, construct, and criterion-related validity.

The following sections present a brief definition of each type of validity evidence, followed by results from the validation study corresponding to each type of evidence.

Sensitive Security Information (SSI)

Content Validity

Content validity refers to whether the selection measure assesses important aspects of the job (Stelly & Goldstein, 2007). Content validity may also refer to the method in which the test assesses these important aspects of the job. For example, the type of stimuli and administration procedures (e.g., time limits) should be appropriate for the job (Guion, 1998). This section describes two forms of evidence of content validity – the job analysis and ratings of jobrelatedness by subject matter experts (SMEs).

Job Analysis

Researchers recommend conducting a job analysis in order to collect evidence of content validity (McPhail & Stelly, 2010; Outtz, 2010). Thus, a primary form of evidence of content validity for the current study comes from the job analysis conducted in 2010 (AIR, 2010a). This job analysis identified the knowledge, skills, abilities, and other characteristics (KSAOs) that a job incumbent would need for successful BDO job performance. This information served as the foundation for identifying and developing job-relevant selection measures. Specifically, AIR identified the preemployment KSAOs that were the most critical for the BDO job and reviewed the literature to identify alternative approaches for measuring these KSAOs (AIR, 2010b). Furthermore, AIR reviewed the selection measures with SMEs who were familiar with the job throughout the development process to ensure that the stimuli and administration procedures coincided with the BDO job.

Feedback Questionnaire

In addition to evidence from the job analysis, AIR collected evidence of face validity using ratings from the feedback questionnaire. Face validity refers to whether test-takers view the measure as valid. The questionnaire included six subscales (see Appendix W for results from four other subscales). Two of these subscales measured predictive and content job-relatedness using a five-point agreement scale. Predictive job-relatedness items asked BDOs how likely the

Sensitive Security Information (SSI)

selection measure would predict successful job performance. For example, one item stated, —A person who scored well on this component will be good at some tasks needed for the BDO job.‖ Content job-relatedness items asked BDOs how closely the content of that selection measure aligned with the tasks BDOs perform. For example, one item stated, —The content of this component was clearly related to certain parts of the BDO job.‖

Means and standard deviations for both subscales are presented in Table 67. Rank orderings based on significant differences between selection measures for both subscales follows.

Table 67. Predictive and Content Job-relatedness Ratings for the Selection Measures

Selection Measure	Mean	Std. Dev.
Predictive Job-Relatedness		
Role-Play Exercise	4.46	.58
Passenger Observation	4.39	.58
Visual Recall	4.23	.62
Writing Knowledge	4.12	.64
Mental Math	4.06	.75
Watson-Glaser Critical Thinking Appraisal	3.58	.78
Wonderlic Personnel Test	3.21	.93
Content Job-Relatedness		
Role-Play Exercise	4.51	.64
Passenger Observation	4.48	.56
Visual Recall	4.22	.65
Mental Math	4.10	.74
Writing Knowledge	4.04	.67
Watson-Glaser Critical Thinking Appraisal	3.18	.88
Wonderlic Personnel Test	2.61	.93

The pattern was the same for both types of job-relatedness subscales. The Role-Play Exercise and the Passenger Observation Assessment received the highest mean ratings, although the other components of the Work Sample Test also received high ratings. The Watson-Glaser Critical Thinking Appraisal and the Wonderlic Personnel Test received the lowest ratings.

Two one-way ANOVAs were conducted to determine whether there were significant differences between the measures in terms of both predictive and content job-relatedness. The results of

these one-way ANOVAs are presented in Table 68. For each ANOVA, the independent variable was a categorization of the type of selection measure (e.g., the Wonderlic Personnel Test versus the Passenger Observation Assessment). The dependent variable was the average ratings for each of the selection measures on a particular feedback subscale (e.g., predictive jobrelatedness). Table 68 also includes effect size estimates (η^2). These estimates indicate the amount of variance accounted for in the subscales that can be attributed to the differences in the predictors. Higher η^2 suggest that a large difference between assessments in levels of jobrelatedness.

Table 68. *F* Statistics for ANOVAs Comparing Differences in Subscales Across Selection Measures

Subscale	Omnibus *F* Results	Eta-squared (η^2)
Content Job-Relatedness	$F(6, 1378) = 179.52$, $p < .001$	.44
Predictive Job-Relatedness	$F(6, 1378) = 79.70$, $p < .001$	.51

Selection measures were then rank ordered based on whether they were significantly different from one another on Bonferoni post-hoc tests (Table 69).

Table 69. Rankings of Selection Measures Based on Job-Relatedness Ratings

Selection Measures	Job-Relatedness	
	Predictive	Content
Role-Play Exercise	1	1
Passenger Observation	1	1
Mental Math	2	2
Writing Knowledge	2	2
Visual Recall	2	2
Watson-Glaser Critical Thinking Appraisal	6	6
Wonderlic Personnel Test	7	7

As shown in Table 69 above, the Role-Play Exercise and Passenger Observation Assessment were rated as significantly higher than any other selection measure included in the experimental selection battery. The second grouping of selection measures consisted of the remaining components of the Work Sample Test. These measures were rated significantly lower than the Role-Play Exercise and the Passenger Observation Assessment, but higher than the Problem Solving measures. Finally, the Problem Solving measures (i.e., the Wonderlic Personnel Test and the Watson-Glaser Critical Thinking Appraisal) were significantly lower than any of the

other selection measures included in the validation study. These trends are similar to those reported in the personnel selection literature for similar types of measures (Hausknecht et al., 2004).

As described above, each feedback questionnaire also included a section for BDOs to provide comments regarding the clarity of the instructions, the difficulty of the measures, and any other recommendations for improving the assessments. Because the assessments closely resembled those in the pilot test in both content and format, AIR expected the response patterns of the qualitative feedback data to resemble the responses from the pilot test. This was largely the case. For example, in terms of the clarity of the instructions, as with the pilot test, validation study participants consistently indicated that the instructions for each assessment were very clear and straight forward.

In addition to the clarity of instructions, BDOs also provided comments regarding the difficulty of each measure. Participants' feedback regarding the difficulty of the Wonderlic Personnel Test closely resembled the feedback provided in the pilot test. Although many BDOs found the questions on the Wonderlic Personnel Test easy, many still commented that the overall test was difficult because of the time limit. For example, one BDO commented, —This test was not difficult in regard to the content. Only the time given to complete it makes it challenging.‖ On the other hand, participants generally felt that the Watson-Glaser Critical Thinking Appraisal was less difficult than the Wonderlic Personnel Test and slightly more reflective of the BDO job. This was evidenced in the following quote, —[This test was] not very difficult; it was more relevant than the [Wonderlic Personnel Test] to the job of a BDO.‖

With regard to the Work Sample Test, BDOs perceived differing levels of difficulty across its components. For example, some BDOs indicated that the Passenger Observation Assessment required them to be responsible for observing too many passengers at one time. As one BDO indicated, —(b)(3):49 U.S.C. § 114(r) ‖ Other participants commented on the difficulty that arose as a result of their BDO training, —The task was somewhat difficult because I already have SPOT training and naturally observed passengers for behaviors beyond those given in this assessment.‖ That being said, most BDOs commented that the Passenger Observation Assessment reached the proper level of difficulty. This was reflected in responses such as, —I thought it was just the right amount of skill testing and will help determine if a person is good at multi-tasking and pays good attention to detail.‖

Because changes were made to increase the difficulty of the Mental Math Assessment, AIR anticipated responses indicating higher levels of difficulty than those from the pilot test. Specifically, as mentioned above, AIR decreased the amount of time for which point values appeared on the screen to one second. Furthermore, negative point values were added which required participants to perform subtraction as well as addition. Consequently, a large number of BDOs perceived the Mental Math Assessment as fairly difficult. For example, one BDO

indicated, —Somewhat difficult due to the speed of the point values appearing on the screen.‖ Nonetheless, BDOs generally indicated that the assessment was fair and an appropriate speed given the requirements of the job. This sentiment is reflected in the following comments, —The difficulty was appropriate for the job‖ and —This [Mental Math Assessment] was accurate in the sense that things happen very quickly and it is important to be at a correct total.‖

With regard to the Writing Knowledge Assessment, some BDOs indicated that this measure was moderately difficult, but did not measure critical aspects of the job. For example, one BDO commented, —SPOT Referral reports are completed by [two] team members jointly…This is completed at the end of [the] shift or during a quiet period - not rushed- with _spell check' etc. It is more important to understand behaviors and how to _resolve' them.‖

For the Visual Recall Assessment, BDOs indicated that this measure was moderately difficult and related to the BDO job. For instance, one BDO commented, —Though challenging, I can see how this component relates to my everyday function as a BDO.‖

In terms of the Role-Play Exercise, BDOs confirmed that they would be significantly better prepared to complete this assessment than would be untrained job candidates, providing several comments such as, —This exercise was not that difficult, but I can see it as being difficult for others who have no previous experience.‖ Nonetheless, most participants felt that the Role-Play Exercise more closely resembled the BDO job than any other component as reflected in the following quote, —Of all the assessments I feel this was the most relevant to a BDO's everyday activities.‖

Finally, BDOs indicated that both the Passenger Observation and the Visual Recall Assessments would provide job candidates with a realistic preview of the BDO job. For instance, numerous BDOs responded similarly to the following quote, —This component [Visual Recall Assessment] is well-suited for informing candidates of what the job will entail.‖

Construct Validity

In addition to content validity, AIR collected evidence of construct validity. Construct validity refers to whether a selection measure assesses what it was designed to measure. This is often evaluated by examining the degree to which a selection measure converges and/or diverges from other measures in the ways one would expect (e.g., strongly, weakly) based on underlying abilities or existing literature (McPhail & Stelly, 2010; Landon & Arvey, 2007). Table 70 depicts the intercorrelations among the selection measures.

Table 70. Intercorrelations among the Selection Measures

Selection Measures	1	2	3	4	5	6	7	8	9	10

1. Wonderlic Personnel Test										
2. Watson-Glaser Critical Thinking Appraisal	.59**									
3. Passenger Observation – Form A	.17	.08								
4. Passenger Observation – Form B	.18	.12								
5. Mental Math – Form A	.56**	.41**	-.12	.21						
6. Mental Math – Form B	.32**	.16	.06	-.13						
7. Writing Knowledge – Form A	.48**	.41**	.00	.18	.46**	.25				
8. Writing Knowledge – Form B	.50**	.26*	.16	-.01	.37*	.09				
9. Visual Recall – Form A	.14	.11	.00	.28	.31*	.11	.15	.29		
10. Visual Recall – Form B	.18	.04	.10	.27*	.11	.04	.33	.18		
11. Role-Play Exercise – Form A	.18	.20	-.02	.12	.14	.03	.25	.07	.13	.19
12. Role-Play Exercise – Form B	.33**	.33**	-.10	-.07	.41**	.24	.30	-.06	.22	-.16

Note. N = 81-201, * p <.05, ** p <.01.

As one would expect, the Problem Solving measures (i.e., the Wonderlic Personnel Test and the Watson-Glaser Critical Thinking Appraisal) are highly correlated (r = .59, p < .01). In addition, both measures are moderately correlated to the Mental Math Assessment (r = .32 - .56, p < .01), Writing Knowledge Assessment (r = .48 - .50, p < .01), and Form B of the Role-Play Exercise (r =.33, p < .01). These relationships are logical given that performance on these measures rely on one's mathematical, verbal, or logical problem solving abilities. Also, weaker correlations are seen between the Problem Solving measures and the Passenger Observation and the Visual Recall Assessments. This pattern was expected because performance on the Passenger Observation Assessment and the Visual Recall Assessment is more likely to rely on perception and recall abilities than on problem solving abilities.

Regarding correlations among the measures on the Work Sample Test and Role-Play Exercise, patterns are, in general, similar to what was expected based on underlying constructs. For example, the Passenger Observation Assessment is weakly related to the Mental Math Assessment (r = -.13 - .21, p > .05); one measure assesses abilities related to observation (e.g., sustained attention, selective attention), whereas the other assesses one's ability to compute mental arithmetic. The Passenger Observation Assessment also diverges from the Writing Knowledge Assessment (r = -.01 - .18, p > .05). Also, scores on the Passenger Observation Assessment did not overlap with the Role-Play Exercise (r = -.02 - .12, p > .05). This relationship was expected given that the Passenger Observation Assessment emphasizes observation abilities, whereas the Role-Play Exercise focuses on Oral Communication and Critical Thinking. Finally, Passenger Observation - Form A diverged from the Visual Recall Assessment (r = .00 to .10), whereas Form B was moderately correlated with this measure (r =

.27 - .28). However, only the relationship between Passenger Observation – Form B and Visual Recall - Form B reached significance ($p < .05$). AIR had anticipated a small degree of overlap between the Passenger Observation Assessment and the Visual Recall Assessment. The moderate correlations that were observed, to some extent, may be attributed to the fact that both measures require some level of observation and recall abilities.

The Mental Math Assessment and Writing Knowledge Assessment, on the other hand, were moderately correlated. In particular, Form A of the Mental Math Assessment overlapped with both forms of the Writing Knowledge Assessment ($r = .37$ - $.46$, $p < .05$). This is likely because of the reliance on one's problem solving ability for both measures. Scores on the Mental Math Assessment ($r = .03$ - $.41$) and Writing Knowledge Assessment ($r = .07$ to $.30$, $p < .05$) also tended to diverge from the Role-Play Exercise. Although, the relationship between Mental Math – Form A and the Role-Play Exercise reached significance ($r = .41$, $p > .05$). AIR anticipated that these measures would diverge. Both measures emphasize problem solving ability, however the Mental Math Assessment and the Writing Knowledge Assessment assess quantitative and verbal reasoning, respectively, whereas the Role-Play Exercise emphasizes critical thinking.

Finally, the Visual Recall Assessment was moderately correlated with the Mental Math Assessment ($r = .04$ - $.31$) and the Writing Knowledge Assessment ($r = .15$ - $.33$). All of these relationships failed to reach statistical significance, except for the correlation between Mental Math - Form A and Visual Recall - Form A ($r = .31$, $p < .05$). This pattern of relationship may be attributed to the shared emphasis on underlying general cognitive ability (i.e., verbal, quantitative, and recall) (Hough, Oswald, & Ployhart, 2001). The Visual Recall Assessment also showed a minimal level of overlap with the Role-Play Exercise ($r = -.16$ to $.22$), although none of these correlations were significant. AIR had anticipated that the Visual Recall Assessment and the Role-Play Exercise would diverge because one measure emphasizes recall abilities, whereas the other focuses on critical thinking and oral communication.

Criterion-Related Validity

Criterion-related validity typically refers to whether scores on a selection measure correspond to ratings of job performance (McPhail & Stelly, 2010; Putka & Sackett, 2010). In the current study, AIR used three measures of job performance: PASS Total Scores, average ratings across dimensions on the BDO Job Performance Measure, and an Overall Performance Composite. This section first describes the process that was used to correct the criterion-related validity coefficients for range restriction in the predictor and unreliability in the criteria. Next, uncorrected criterion-related validity coefficients are presented along with coefficients that have been corrected for indirect range restriction. Finally, coefficients that have been corrected using alternative approaches are also presented.

Correcting Criterion-Related Validity Coefficients

Sensitive Security Information (SSI)

The ability of selection measures to predict relevant job outcomes can be quantified using criterion-related validity coefficients. These coefficients are usually attenuated by measurement or methodological factors and represent lower-bound estimates of criterion-related validity. For instance, if the measure used to assess job performance (i.e., the criteria) does not accurately depict how an employee performs, it may appear that the selection measure is not an effective predictor. Also, if the sample that is used to validate the selection measure deviates from the intended applicant pool, then the ability of these measures to predict relevant job outcomes may be obscured. Researchers (Hunter, Schmidt, & Le, 2006; Sackett & Yang, 2000) have proposed a number of ways to correct for these types of artifacts. Corrected validity coefficients provide clearer estimates of a measure's predictive capabilities and allow policy makers to make more informed decisions.

Lower levels of reliability in either the selection measure or job performance measure will attenuate criterion-related validity coefficients. Most researchers, however, suggest only correcting for lower levels of reliability in job performance measures (Van Iddekinge & Ployhart, 2008). Correcting for measurement error in selection measures would provide criterion-related validity correlations that deviate from operational, or real-world, circumstances (Hunter et al., 2006).

The estimate of reliability that one uses to correct for measurement error in the criterion directly influences the magnitude of the corrected validity coefficient. Researchers typically correct job performance measures using either estimates of internal consistency or interrater reliability. Estimates of internal consistency indicate how well a group of items measure the same construct. However, because raters may struggle with distinguishing among multiple performance dimensions (e.g., Passenger Observation, Calculating Points, Documentation), job performance measures often demonstrate elevated levels of internal consistency. Using internally consistency estimates will tend to overestimate the job performance measure's reliability and underestimate corrected validity coefficients (Hunter et al., 2006). An alternative approach is to use measures of interrater reliability (i.e., the degree to which independent raters provide similar patterns of ratings for an employee). Researchers suggest estimating the average Pearson correlation or intraclass correlation coefficients (ICCs) between multiple raters and correcting for the criterionrelated validity coefficients based on the level of observed unreliability (Hunter et al., 2006).

In the current study, the average Pearson correlation for the BDO Job Performance Measure was .52. AIR used this estimate to correct the criterion-related coefficients between the selection measures and the BDO Job Performance Measure. Because the PASS measure relied on archival data and the Overall Performance Composite consisted of a combination of the other two criterion measures, no estimate of interrater reliability was available. Thus, these measures were only corrected for range restriction.

After correcting for unreliability in the job performance measure, researchers recommend that one should consider whether there is evidence of range restriction (Hunter et al., 2006). When validating selection measures, range restriction typically occurs because the sample of test-takers (most often job incumbents) has already been screened prior to being included in the study. These previous screening measures have likely removed many of the low-performing test-takers. For instance, prior to the current validation study, BDOs had successfully completed a Structured Interview and SPOT training. Furthermore, on average, BDOs who participated in the current study had 2.93 years of experience in the SPOT program (*SD* = 1.17) and received slightly elevated performance ratings. Thus, it is unlikely that, as a group, BDOs' scores on the selection measures and job performance ratings would show the same levels of variability as scores from a group of job applicants. This restricted variability can attenuate criterion-related validity coefficients and make it difficult to determine how effective the newly developed selection measures can predict job performance.

There are a number of approaches available for correcting for range restriction (Sackett & Yang, 2000). An important factor to consider when choosing an appropriate approach is determining whether range restriction can be attributed to direct or indirect factors. Direct range restriction occurs when the current sample of job incumbents has been previously screened on the same measures that are being validated. For example, this may occur when an organization validates a structured interview with a sample of incumbents who were previously screened using that same structured interview measure (Van Iddekinge & Ployhart, 2008). Indirect range restriction occurs when the measure being validated is associated with the selection measure used to screen the study's participants (i.e., incumbents) (Sackett & Yang, 2000). For instance, if an organization validates a problem solving measure and scores on the new measure are associated with ratings from the structured interview used to screen incumbents, then there will likely be indirect range restriction.

In the present study, scores on many of the selection measures are correlated with ratings from the existing Structured Interview (r = .05 -.26). Thus, based on current recommendations from the literature, the most appropriate approach for the present study would be to correct for indirect range restriction. Researchers have recently described a 7-step approach for addressing this form of range restriction (Hunter et al., 2006). This approach is presented in Table 71.

Table 71. Steps for Correcting for Indirect Range Restriction

Step	Description
1	Correcting for measurement error in the job performance measure
2	Calculating reliability of selection measure with incumbents[a]
3	Correcting for measurement error in the selection measure
4	Estimating reliability of the selection measure with the applicant population

5	Estimating indirect range restriction
6	Correcting for the effect of indirect range restriction
7	Reintroducing measurement error in the selection measure in order to estimate operational validity

Note. Adapted from (Hunter et al., 2006). [a] This value is typically known in concurrent validation studies.

This approach begins by removing measurement error and then applying the correction for direct range restriction (Steps 1-4). Indirect range restriction is then estimated and corrected (Steps 56). Finally, measurement error is reintroduced into the selection measure in order to provide an estimate of how well the assessment will predict performance when applied in an operational setting.

As seen in Step 5, correcting criterion-related validity coefficients for range restriction partially consists of adjusting the magnitude of the correlation based on the degree to which the sample that is used to validate the selection measure deviates from the intended applicant pool. One approach for estimating this degree of deviation is the ratio of the standard deviation of selection measure scores among the incumbent sample (SD_I) to these scores within an applicant population (SD_A) (Hunter et al., 2006). One challenge with this approach is identifying an appropriate estimate of variability for the applicant population (SD_A). Because data are typically collected using incumbents, precise values of variability within the applicant population are often not available.

Some researchers have derived estimates of SD_A using other available information (e.g., selection ratios). However, to date, this approach has only been applied to situations that represent direct range restriction (Sands, Alf, & Abrahams, 1978). Others have generated common estimates of variability based on large-scale selection studies (e.g., SD_{ratio} =.67 for cognitive ability measures) (Hunter et al., 2006). However, some have argued that these estimates do not take into account differences across specific occupations or the processes job applicants employ when choosing jobs (Schmitt, 2007). Given the challenges with using a common metric, and the limited applicability of estimates based on selection ratios, researchers (Sackett & Ostgaard, 1994) have recommended estimating the variability of test scores in the applicant population (SD_A) based on the amount of variability observed among applicants on relevant norm-referenced tests. The Wonderlic Personnel Test was a viable option in the present study for two reasons. First, it has been administered to over 100,000 applicants in various fields, which provides a more stable estimate of variability among job candiates. Second, it was moderately associated with each newly developed selection measure (r = .17 -.59). This level of overlap suggests that similar levels of variability may be seen among job applicants' scores on the newly developed selection measures.

Researchers (Sackett & Ostgaard, 1994) recommend reducing the observed standard deviation ratio for the overall sample of applicants by 10-20% (using a larger reduction for more complex

jobs) to produce a conservative estimate of SD_A for the job-specific sample. By dividing the standard deviation of the Wonderlic Personnel Test scores among incumbents ($SD_I = 5.73$) by the standard deviation for applicants' scores reported within the literature ($SD_A = 7.40$) (Sackett & Ostgaard, 1994), AIR computed a SD_{ratio} for scores on the Wonderlic Personnel Test:

Formula	Calculations
$SD_{ratio} = \frac{SD_I}{SD_A}$	$.77 = \frac{5.73}{7.40}$

By reducing the observed SD_A for the Wonderlic Personnel Test scores one can conservatively estimate what the ratio would be for the newly developed BDO selection measures. Because the BDO job consists of a variety of complex, unobservable tasks, AIR reduced the SD_A for the Wonderlic Personnel Test by 20% ($SD_A = 5.92$). By recalculating the SD_{ratio} using the revised SD_A, one gets a more conservative estimate.

Formula	Calculations
$SD_{ratio} = \frac{SD_I}{SD_A}$	$.97 = \frac{5.73}{5.92}$

By applying this same SD_{ratios} to the other selection measures in the experimental selection battery, one can conservatively estimate the degree of range restriction observed within the current sample. For instance, the formula depicts how the 20% SD_A was calculated for Passenger Observation – Form A.

Formula	Calculations
$SD_{ratio} = \frac{SD_I}{SD_A}$	$.97 = \frac{3.09}{SD_A} \rightarrow .97 = \frac{3.09}{3.19}$

Using this same approach for the other selection measures, AIR estimated the degree of range restriction for each selection measure. The degree of estimated range restriction was then incorporated into the indirect range restriction corrections in order to estimate operational validity estimates. These results are described in the next section.

Criterion-Related Validity Coefficients

Criterion-related validity coefficients for each of the selection measures and the composite scores are depicted in Table 72 (the criterion-related validity coefficients for subscales are presented in Appendix X). Correlation coefficients in parentheses have been corrected for indirect range restriction and/or unreliability in the criterion measure using the process described in the previous section.

Sensitive Security Information (SSI)

Researchers (e.g., DOL, 2000) have outlined rules-of-thumb for interpreting these coefficients. Generally, selection measures with moderate correlations (r = .21 to .35) are likely to provide useful information. Selection measures with lower correlations (r = .11 to .20) may provide useful information, but it will depend on other circumstances (e.g., the amount of overlap with other selection instruments in the selection battery).

Table 72. Corrected and Uncorrected Validity Criterion-Related Coefficients for the Selection Measures

Selection Measures	PASS Total Scores	BDO JPM - Average	Overall Performance Composite
Watson-Glaser Critical Thinking Appraisal	(b)(3):49 U.S.C. § 114(r)		
Passenger Observation – Form A			
Passenger Observation – Form B			
Mental Math – Form A			
Mental Math – Form B			
Writing Knowledge – Form A			
Writing Knowledge – Form B			
Visual Recall – Form A			
Visual Recall – Form B			
Role-Play Exercise – Form A			
Role-Play Exercise – Form B			

Note. N = 81-201. * p <.10, ** p <.05, **

The Watson-Glaser Critical Thinking Appraisal showed similar patterns of association with the PASS, the BDO Job Performance Measure, and the Overall Performance Measure. Both forms of the Passenger Observation Assessment and Role-Play Exercise were moderately associated with various outcome measures and are likely to provide useful information.

At least one form for the remaining selection measures (i.e., the Mental Math Assessment, the Writing Knowledge Assessment, and the Visual Recall Assessment) demonstrated adequate levels of criterion-related validity. Mental Math – Form A was highly related to the PASS Total Score and showed moderate levels of association with the BDO Job Performance Measure. Likewise, Writing Knowledge – Form B showed higher levels of association with the PASS Total Score than with the BDO Job Performance Measure. Visual Recall – Form A was moderately correlated with the BDO Job Performance Measure, whereas Visual Recall – Form B showed a higher level of association with the PASS.

Alternative Approaches for Correcting Criterion-Related Validity Coefficients

This section presents alternative approaches for correcting criterion-related validity coefficients. Although not all of these approaches are suitable for the current study, this section is intended to be informative and display the range of operational validity coefficients. Specifically, corrected coefficients have been adjusted using estimates of unreliability in the job performance measure (Hunter et al., 2006), direct range restriction estimates based on selection ratios (SR) (Sands et al., 1978), direct range restriction based on standard deviation ratios (SD_{ratio}) (Hunter et al., 2006), and/or indirect range restriction. The first table (Table 73) presents corrected and uncorrected coefficients for the selection measures and composite scores on PASS. The second table (Table 74) presents corrected and uncorrected coefficients for the selection measures and composite scores on the BDO Job Performance Measure.

Because estimates of reliability were not available for subcomponents on the PASS, AIR was unable to correct for unreliability in the performance measure. Nevertheless, these criterionrelated validity coefficients were corrected for direct and indirect range restriction. Based on the corrections using selection ratios, at least one form for each measure is very likely to provide useful information in situations where no more than a quarter of the applicants will be hired (i.e., the selection ratio is less than or equal to 25%). When considering PASS scores, corrections for indirect range restriction based on SD_{ratios} indicate that the Problem Solving measures are likely to provide useful information. However, both forms of the Passenger Observation Assessment and the Writing Knowledge Assessment, Mental Math – Form A, and Visual Recall – Form B are also likely to provide useful information and the Role-Play Exercise may provide useful information depending on the circumstances (e.g., whether some of the other measures are included in the selection battery or not).

Based on corrections using estimates of selection ratios, one of the most effective predictors of BDO JPM ratings was the Passenger Observation Assessment. This assessment is very likely to provide useful information in situations where no more than 25% of applicants will be hired. Other measures (e.g., the Visual Recall Assessment, the Role-Play Exercise) are also likely to provide useful information in these circumstances.

Table 73. Corrected Criterion-Related Validity Coefficients for Total PASS Points

(b)(3):49 U.S.C. § 114(r)

Table 74. Corrected Criterion-Related Validity Coefficients for Overall BDO Job Performance Measure Rating

(b)(3):49 U.S.C. § 114(r)

Coefficients corrected for indirect range restriction using SD_{ratios} also suggest that these same measures would likely provide useful information during the hiring process. Specifically, both forms of the Passenger Observation Assessment, the Visual Recall Assessment, and the RolePlay Exercise are likely to provide useful information. Form B of the Writing Knowledge Assessment is also likely to be a useful predictor of BDO Job Performance ratings.

Summary of Form Differences

As discussed in Chapter III, the different forms of the selection measures were developed in a parallel fashion, but were not formally equated. In other words, the measures were developed using identical test specifications and methodology. However, it was not feasible in the current study to include the same items on both forms of some of the measures. For example, the Passenger Observation Assessment and the Visual Recall Assessment consisted of stimuli that were distinct across forms (i.e., separate videos and images, respectively). Thus, generating items for these two measures that were identical was not possible. Likewise, administering both forms of each measure to the same participants would have required additional time on-site during the data collection process and reduced the test security of each measure.

Despite the efforts given to ensure that the forms were parallel, AIR identified instances in which the two forms of each selection measure diverged. For example, Mental Math - Form A displayed moderate levels of criterion-related validity across the three performance criteria (r = .20 - .30, $p < .05$). On the other hand, Mental Math - Form B demonstrated much lower levels of criterion-related validity (r = .06 - .09, $p > .05$). Similar patterns of form divergence were also observed for the other selection measures.

Researchers have suggested that sample sizes less than two hundred are susceptible to sampling error (Drasgow, Nye, & Tay, 2010). Given the relatively small sample sizes per form (e.g., n = 111 for Mental Math Form A, and n = 92 for Mental Math Form B), these differences may be attributed to sampling error. In other words, irrelevant characteristics of the sample may be accounting for the divergence patterns of criterion-related and construct validity between the two forms for each selection measure. AIR conducted exploratory analyses in order to identify differences in BDOs across the two forms that may have contributed to these form differences. For example, AIR examined whether BDOs who completed each form differed on their scores for the Problem Solving measures, organizational tenure, and the types of shifts they typically work. None of these analyses identified a subgroup of participants that were causing the form differences. AIR's recommendations in Chapter VII of this report have taken these form differences into account.

Subgroup Differences and Adverse Impact

Although AIR administered the selection measures to a diverse group of BDOs, the sample sizes for particular subgroups prohibited stable estimates of statistical adverse impact analyses (Hough et al., 2001). For example, of the 214 BDOs that participated in the study, only 11 identified themselves as Asian, 53 as Black or African American, and 43 as Hispanic or Latino. These sample sizes become significantly smaller if the separate forms of each selection measure are considered. For instance, Table 75 displays the demographics for BDOs who completed Forms A and B of the Mental Math Assessment.

Table 75. Demographics by Form for the Mental Math Assessment

Demographics	Mental Math Form A		Mental Math Form B	
	N	Percentage	N	Percentage
Gender				
Female	40	36	32	32
Male	72	64	69	68
Ethnicity				
Hispanic or Latino	25	22	18	18
Not Hispanic or Latino	86	77	82	81
Race				
Asian	3	3	8	8
Black of African American	25	22	27	27
Native Hawaiian or Pacific Islander	3	3	5	5
Two or More Races	12	11	5	5
White	57	51	48	48
American Indian or Alaska Native	—	—	1	1
Other	—	—	5	5

— Not Available

Across each form of the Mental Math Assessment, there are very few BDOs who identified themselves as a protected class. For example, less than 30 BDOs who identified themselves as

Hispanic or Latino completed either Form A (n = 25) or Form B (n = 18) of the measure. Similar levels were observed for Black or African Americans (n < 28). Even smaller values were observed for Asian test-takers (n < 10). Similar sample sizes for protected classes were observed for each of the selection measures.

Given these relatively small sample sizes, it is unlikely that adverse impact analyses would have been stable enough to provide useful information. Also, because the data are based on a voluntary sample of previously-screened and experienced BDOs, these analyses may not generalize to a sample of applicants (Tippins, 2010). For some of the analyses a single unusual value may have led to aberrant findings. Rather than conduct these analyses that may have led to inaccurate interpretations and conclusions, AIR identified relevant literature on the likelihood of subgroup differences for each selection measure. This information may be helpful to consider

while identifying which selection measures should be included in the revised selection system. AIR recommends that TSA continue to evaluate subgroup differences and adverse impact results for any selection system implemented for this or other jobs.

Problem Solving Measures

Because the Watson-Glaser Critical Thinking Appraisal and the Wonderlic Personnel Test are measures of problem solving, they are likely to lead to subgroup differences across races (Roth, Bevier, & Bobko, 2001). For instance, most White test-takers score, on average, a standard deviation higher than Black test-takers (Gottfredson, 2002; Outtz, 2002; Roth et al., 2001). It should be noted, however, that subgroup differences are greater when the tests are used for jobs that are higher in complexity than for jobs lower in complexity. Many researchers have concluded that these large subgroup differences do not prohibit the use of general cognitive ability measures, but they do require that the use of such tests be justified (Outtz, 2002; Roth et al., 2001). Typically this justification requires absolute job-relevance of the construct being measured, as indicated by a job analysis and the measure's ability to predict job performance in the job in question. Further justification comes from a measure's robust ability to predict job performance across a number of jobs and settings (Gottfredson, 2002; Schmidt & Hunter, 1998). In addition, comparisons between members of other subgroups (e.g., age, gender) often yield much smaller mean differences (Hough et al., 2001).

Work Sample Test and Role-Play Exercise

The Work Sample Test and the Role-Play Exercise were designed to emphasize specific abilities rather than general problem solving. Measures of specific ability tend to yield subgroup differences that are slightly smaller than those observed with a general problem solving measure (Roth et al., 2001). Both general problem solving measures and specific measures of cognitive ability (e.g., selective attention, sustained attention) may demonstrate low levels of subgroup differences between men and women (Bates & Lemay, 2004; Giambra & Quilter, 1989).

Recent estimates of racial subgroup differences for high-fidelity assessments have yielded moderate differences (d = .70) (Bobko, Roth, & Buster, 2005). The authors cautioned that this estimate was based on a broad sample of work sample exercises and that specific assessment techniques may yield lower mean subgroup differences. Subsequent analyses have also indicated that certain exercises (e.g., briefings, role-plays) may yield lower levels of subgroup differences than other exercises (e.g., in-basket, technical demonstrations) (Roth, Bobko, McFarland, & Buster, 2008). Likewise, including a performance simulation rather a paperandpencil exercise may reduce subgroup differences (Schmitt & Mills, 2001).

Operational Considerations

In addition to examining the evidence of reliability and validity for the selection measures, AIR also took into account operational considerations such as the ease of administration and the amount of resources required to administer and score each measure. These considerations are described below in more detail.

Ease of Administration

When considering the ease of administration, AIR evaluated the amount of training required for an Administrator to be able to proctor each assessment as well as the number of candidates that can complete a measure in a single administration. Each of the selection measures include an administration guide with instructions for establishing a proper testing environment, a script for proctoring each assessment, and answers to frequently asked questions. Individuals who administer the Role-Play Exercise will need to be trained on evaluating performances. This training would likely be similar to that provided for the Structured Interview Assessors.

In addition to the amount of training required for an Administrator, AIR also considered the number of candidates that could complete each selection measure during a single administration. If there is enough space to ensure a proper testing environment, the Watson-Glaser Critical Thinking Appraisal can be administered to a large number of candidates. The Work Sample Test can be administered to smaller groups of candidates (i.e., three to four people). Finally, the

Role-Play Exercise, similar to the Structured Interview, is best suited for a one-on-one administration.

Resource Requirements

AIR also considered the amount of resources, beyond identifying and training an Administrator that would be required for each selection measure. All of the measures, excluding the Writing Knowledge Assessment, will require some form of additional resources. For the Watson-Glaser Critical Thinking Appraisal, the additional resource is a recurring purchasing fee. The Passenger Observation, Mental Math, and Visual Recall Assessments require computer-based equipment because they rely on high-fidelity stimuli that are presented using computer monitors and audio equipment. Finally, because the Role-Play Exercise was designed to be administered immediately after the Structured Interview, it does not require additional Assessors. However, if implemented, TSA will need to identify and train one or more individuals to serve as the Resource Person. This/these individual(s) will need to be familiar with the BDO job and be able to provide a consistent performance across administrations.

Summary of Existing Evidence

This section provides an overview of the validity and other evidence that was collected during the validation study. Table 76 below briefly summarizes the evidence that was considered when making recommendations. Each column corresponds to a different piece of evidence and each row represents a separate selection measure. The color coding and lettering scheme is intended to provide a comparative review of the selection measures within each column with red circles with the letter —L‖ representing the lowest value, yellow circles with the letter —M‖ being moderate, and green circles with the letter —H‖ being the highest. Note that these summary findings are based on the results of the current study and AIR's experiences with these measures. These results may not reflect other pricing options, data collection methods, and administration processes that are not consistent with the parameters of this study.

Table 76. Overview of the Validation Study Findings[49]

Measure	Reliability	Content Validity	Face Validity	Construct Validity	Criterion-Related Validity	Ease of Admin	Resource Reqs.	Overall
Watson-Glaser	H	H	L	H	H	H	M	L
Passenger Observation	M	H	M	H	H	H	M	H
Mental Math	H	H	H	M	H	H	M	H
Writing Knowledge	H	M	H	M	H	H	H	L
Visual Recall	M	H	H	M	M	H	M	H
Role-Play Exercise	M	H	H	H	H	M	M	H

Note. Color coding scheme provides a comparative view within each column. Red and —Ll = lowest, Yellow and —Ml = moderate, Green and —Hl = highest. Admin = Administration. Reqs. = Requirements.

[49] These summary findings are based on the results of the current study and AIR's experiences with these measures. These results may not reflect other pricing options, data collection methods, and administration processes that are not consistent with the parameters of this study.

Reliability

Within the reliability column, the Watson-Glaser Critical Thinking Appraisal (α = .79), the Mental Math Assessment (α = .80 - .82), and the Writing Knowledge Assessment (α = .75 - 80) demonstrated the highest levels of reliability whereas the Passenger Observation Assessment (α = .31 - 45), the Visual Recall Assessment (α = .53 - .63), and the Role-Play Exercise (weighted κ = .32 - .57) were slightly lower. Although the coefficient alpha levels for the Passenger Observation and Visual Recall Assessments are less than the common threshold of .70, the observed levels of internal consistency are similar to what has been seen with other high-fidelity, work sample measures (Lievens & Coetsier, 2002). Thus, a more appropriate assessment of reliability may be a test-retest coefficient. Given the parameters of the current study, it was not feasible to assess this type of reliability. The kappa values observed for the Role-Play Exercise represent moderate agreement (Landis & Koch, 1977). Given the complexity of the performances that were being observed, this is an encouraging finding for the Role-Play Exercise.

Content Validity

Because all of the measures are based on a comprehensive job analysis, there is a substantial amount of evidence indicating that they assess job-relevant constructs. However, the Writing Knowledge Assessment was rated slightly lower in this column because there was some evidence suggesting that the underlying ability (i.e., Written Communication) may not be as critical as the abilities assessed by other measures (e.g., the Passenger Observation Assessment, the Role-Play Exercise). Furthermore, Written Communication may be able to be addressed during training. All of the measures, excluding the Watson-Glaser Critical Thinking Appraisal, received high ratings of job-relatedness from BDOs. In other words, BDOs were able to see the connection between the content of these measures and their job and also indicated that higher scores on the selection measures were indicative of better job performance.

Construct Validity

The Watson-Glaser Critical Thinking Appraisal, the Passenger Observation Assessment, and the Role-Play Exercise demonstrated high levels of construct validity. These measures converged and diverged from other measures in the validation study in ways that were expected based on their underlying constructs. For example, the Passenger Observation Assessment did not overlap with other measures in the study (r = -.12 - .28).[50] This was expected because one's ability to observe and identify behaviors should not correspond with the other abilities measured during the study (e.g., mental arithmetic, deductive reasoning, oral communication). The Mental Math Assessment, the Writing Knowledge Assessment, and the Visual Recall Assessment converged in ways that were expected, but did not diverge to the degree that AIR anticipated. For example, the Visual Recall Assessment overlapped with the Mental Math Assessment (r = .04-.31) and the Writing Knowledge Assessment (r = .15-.33). To some extent this could be expected because each of these measures assesses a specific cognitive ability and these abilities are often related (Carretta & Ree, 2000). However, individuals' ability to recall information is distinct from their

[50] A range of construct validity coefficients are presented for measures to represent both forms (i.e., Forms A and B).

arithmetic ability or writing abilities. Thus, these measures were expected to diverge to a greater degree. This overlap between the Visual Recall Assessment and other measures suggests it may not contribute unique information as part of the experimental selection battery.

Criterion-Related Validity

All of the measures demonstrated at least moderate levels of criterion-related validity. Specifically, at least one form, and in some instances, both forms of the selection measure, predicted PASS Total Scores or average ratings on the BDO JPM. Significant correlations were also observed between the selection measures and the Overall Performance composite. Some of the strongest predictors of PASS scores were the Mental Math – Form A (b)(3):49 the Writing Knowledge Assessment (b)(3):49 U.S.C. § 114(r) and Visual Recall – Form B (b)(3):49 For the BDO JPM, the strongest predictors were the Passenger Observation – Form A (b)(3):49 U.S.C. § 1 Writing Knowledge – Form B (b)(3):49 U.S.C. § and the Role-Play Exercise – Form B (b)(3):49 U.S.C. §

Operational Considerations

When taking into account operational considerations, AIR considered the number of applicants that could complete a measure in a single administration and the amount of training that it would take to prepare an Administrator to proctor a session. The Role-Play Exercise, because it is administered on a one-to-one basis and requires Administrators to provide performance ratings, was rated slightly lower than the other selection measures. All of the measures, excluding the Writing Knowledge Assessment, would require additional resources. The Watson-Glaser Critical Thinking Appraisal requires a recurring purchasing fee. The Passenger Observation, Mental Math, and Visual Recall Assessments require computer-based equipment (e.g., monitors, audio/visual cords) in order to be implemented. Finally, the Role-Play Exercise requires a trained Resource Person (i.e., role-player) who can provide a consistent performance as the passenger across administrations.

Recommended Selection Battery

AIR identified the Passenger Observation Assessment, the Mental Math Assessment, and the Role-Play Exercise as the most promising selection measures. Each of these measures assessed unique skills and abilities that are critical for BDO job performance. Also, all three measures predicted performance ratings on either the PASS or BDO JPM. Furthermore, because these measures were rated as highly job-relevant by BDOs and require applicants to perform tasks that are similar to those performed by BDOs, the system would provide a realistic job preview and may increase the number of applicants who opt out of the selection process, which can help TSA save valuable time and financial resources. Based on this evidence, AIR recommended that TSA

include these three measures as part of an experimental selection battery. During this experimental phase, TSA could collect additional data from job applicants, and continue to examine the feasibility of implementing each selection measure nationwide.

CHAPTER VI: POST-VALIDATION STUDY CHANGES

The previous chapter described the method and results from the validation study. This chapter describes the minor changes the AIR implemented following the validation study. These changes were based on the results of the validation study, feedback provided by BDOs, and AIR's experiences during the data collection process.

Overview and Purpose

Following the validation study, AIR compiled and reviewed qualitative and quantitative data collected from several sources on each of the assessments. These data were reviewed to identify any necessary changes to the assessments. Sources of data included results from the scale and item-level analyses, ratings from the feedback questionnaires, and audio recordings from the group feedback sessions and the Role-Play Exercise. Potential changes were prioritized and only those that were deemed necessary were implemented. Necessary changes were defined as minor modifications that would likely improve item- and scale-level statistics and increase the clarity of the assessment's instructions. AIR emphasized minor changes to ensure the fidelity between the measures that were included in the validation study and the final assessments that were provided to TSA. Significant changes to the assessments would otherwise limit the applicability of the validation study results.

Though only three of the measures from the validation study were included in the recommended selection battery (i.e., the Passenger Observation Assessment, the Mental Math Assessment, and the Role-Play Exercise), changes were made to all of the assessments so that these measures would be prepared for any possible future uses (i.e., use as training tools, performance measures, or inclusion in other research studies). Each post-validation study change implemented is described below.

First, AIR removed any language that was included within the Work Sample Test Administration Guide solely for the validation study. For example, AIR removed language that described a BDO's participation in the validation study (e.g., reminders of the voluntary nature of the study). In addition, AIR revised the phrasing in the Work Sample Test Administration Guide so that the assessments may be administered separately. For example, the term —component‖ was replaced with the term —assessment.‖ Also, language that linked one component to the other (e.g., —next, we will begin the Mental Math Assessment‖) was removed. This change was implemented to provide TSA with more flexibility in using each assessment and to accommodate AIR's recommendation that not every assessment from the Work Sample Test be included in future studies.

Next, changes were made to address quantitative and qualitative data collected regarding each assessment included in the Work Sample Test. The section below provides a brief description of each of the changes that were made.

Passenger Observation Assessment

Based on feedback from BDOs, changes were made to the Passenger Observation Assessment to increase the clarity of the instructions in two areas. First, instructions were added to emphasize the importance of recording only the behaviors that are visible—rather than inferring that behaviors have occurred. Though this instruction was provided in the script following the practice test, subject matter experts (SMEs) indicated that it would be helpful to provide additional guidance before the practice test begins. Thus, AIR included this information at both points in the instructions.

Second, scoring procedures were altered to account for changes in scoring made during the data analysis stage in order to maximize the reliability of the assessment and the efficiency of the scoring process (as described in Chapter V). Specifically, instructions to Assessors were altered to indicate that job candidates receive credit for each instance that they correctly identify the presence of a behavior or appearance factor—as opposed to the previous scoring methodology, which provided credit for identifying the presence *or absence* of a behavior or appearance factor. Instructions to job candidates regarding how their performance on the Passenger Observation Assessment is scored were also edited to account for this change in scoring procedures.

Mental Math Assessment

Based on AIR's experiences administering the Mental Math Assessment, instructions were added to the Administration Guide for the Work Sample Test Administrator to start the laptop and ensure the Microsoft PowerPoint® slideshow is functioning properly prior to a job candidate's

arrival. This addition was designed to help ensure that the necessary equipment and software are functioning appropriately before the Work Sample Test begins.

In addition, during group feedback sessions, some BDOs indicated that the last number for each item appeared on the screen for a shorter length of time than the numbers that preceded it. Based on this feedback, two AIR researchers independently reviewed the timing of each item. This review consisted of double-checking the animation mechanism within the Microsoft PowerPoint® slide and using a stopwatch to time the presentation of each number. Results of this review indicated that each number, including the final number, appears on the screen for exactly one second, as it is designed to do. AIR concluded that BDOs' perceptions may be attributed to the relative amount of time that the final number appeared on the screen. Specifically, all the preceding numbers remain on the screen as subsequent numbers appear. The final number, in comparison to the earlier numbers, remains on the screen for the shortest amount of time (although, it still appears for one second). This may have created the impression that the final number is shown for a shorter amount of time. Given the outcome of this review, AIR did not modify the timing of the Mental Math Assessment.

Writing Knowledge Assessment

During the administration of the Writing Knowledge Assessment, some BDOs had questions regarding stylistic issues. Specifically, they asked how to address stylistic grammar rules that could be considered correct by some, but not by others. For example, one rule that was cited was whether to include a comma before the word —and‖ in a series of items. To increase the clarity of the instructions with regard to this issue, instructions were added to the script to indicate that stylistic issues should not be considered errors.

Visual Recall Assessment

Based on feedback provided on the feedback questionnaires and during the group feedback sessions, AIR reviewed potentially problematic multiple-choice questions for the Visual Recall Assessment. Following this review, two revisions were made to Form B of the assessment. Specifically, in Image 1, some BDOs indicated that it was unclear which passenger was the focus of Item 14. AIR revised the item stem for this question to address possible confusion. For Image 2, some BDOs commented that one distractor in Question 12 could be considered the correct response. AIR revised this response option to ensure that the item would not be perceived as unfair or misleading.

Role-Play Exercise

For the Role-Play Exercise, as with the Work Sample Test, any language that had been previously included in the Role-Play Exercise Administration Guide specifically for the validation study was removed. For example, language that referred to ignoring signs of deception or other SPOT-related behaviors during the Role-Play Exercise, or information about using an audio recorder, was removed.

In addition to removing instructions that were specific to the validation study, AIR also modified the Role-Play Exercise based on item- and scale-level analyses, qualitative feedback, and AIR's experiences administering the assessment. Specifically, changes were made to the following components of the Role-Play Exercise:

- Instructions to Assessors,
- Instructions to job candidates,
- The Resource Person's character, The Summary Statement checklists, and
- The Role-Play Exercise rating scales.

Each of these changes is described in more detail in the following sections.

Instructions to Assessors

Based on administration issues that arose during the validation study, two additions were made to the instructions for Assessors in the Role-Play Exercise Administration Guide. These instructions were added to improve the standardization of this assessment. First, instructions were added for Assessors to confirm a job candidate's final response to questions from the Summary Statement Checklist. During the validation study, some BDOs initially provided more than one response to a question. Thus, AIR included language for the Assessors to confirm a job candidate's response if he or she provides more than one response. Second, instructions were added on how to score a job candidate's response to a Summary Statement question if the Resource Person provides inaccurate information during the Role-Play Exercise (e.g., indicates he/she is traveling to the wrong destination). Specifically, Assessors are instructed to award full credit to a job candidate if he/she repeats inaccurate information that was originally stated by the Resource Person.

Instructions to Job Candidates

The instructions that are provided to job candidates during the Role-Play Exercise are comprehensive. Specifically, these instructions review the scenario for the Role-Play Exercise, the parameters for the conversation, and the evaluation criteria. Due to the length of the instructions, the list of Summary Statement questions is presented early in the administration process and they repeated again at the end. Once all instructions are provided, job candidates are given one minute to review the Summary Statement questions before the exercise begins. During group feedback sessions, some BDOs admitted that once the Summary Statement questions were provided, they stopped listening to the instructions and began formulating their strategy for eliciting the necessary information. To help prevent this occurrence, instructions were added to the beginning of the script to inform job candidates that they will have time to review the Summary Statement questions again before the exercise begins.

Resource Person's Character

Following the validation study, changes were made to the Resource Person's character for two reasons. First, based on item-level difficulty statistics, a change was made to address observed differences in difficulty across forms. Specifically, one Summary Statement question in Form B of the Role-Play Exercise appeared to be more difficult than the others. AIR determined that this

question was especially challenging because it required job candidates to elicit and retain two distinct pieces of information whereas the other Summary Statement questions only required job candidates to elicit one piece of information. To address this issue, the problematic question was revised so that it too only required candidates to elicit one piece of information.

Second, AIR added information to the Resource Person's character to address participants' questions that inquired about information not included in the script. Including this information helped expand the breadth of the character and ensure that participant questions that were asked frequently during the Role-Play Exercise (e.g., where is [the resource person's] office located? who planned the trip?) could be addressed during future administrations. Following the validation study, AIR reviewed recordings of the Role-Play Exercise to document the types of

questions participants posed that were beyond the scope of the current character. Using these data, a few additional details were added to the Resource Person's character (e.g., location of the Resource Person's place of work, information about who planned the Resource Person's trip). This information was also incorporated into the Resource Person Training materials. Specifically, additional instructions were added that highlight the types of questions that were frequently asked during the pilot test and validation study.

Summary Statement Checklists

To help Assessors score job candidates' responses to the Summary Statement questions in a standardized manner, the Role-Play Exercise Administration Guide includes examples of acceptable and unacceptable responses in the Summary Statement Checklists. While administering the Role-Play Exercise during the validation study, AIR received some responses to the Summary Statement questions that were not included in the original checklists. To help increase the standardization of future administrations, AIR has included suggestions for how to score these types of responses and added examples of each to the Summary Statement Checklists.

CHAPTER VII: SUMMARY AND RECOMMENDED NEXT STEPS

This chapter reviews the work conducted for the current project, the important outcomes and products, and the key findings. Also, this chapter provides recommendations for next steps.

Review of Completed Work

The research consisted of multiple steps, each of which provided important information about the BDO job and the experimental selection measures. These steps included the following:

- Conducted and documented a job analysis of the BDO job (AIR, 2010a);
- Developed test specifications to determine the appropriate selection measures for assessing important pre-hire knowledge, skills, abilities, and other characteristics (KSAOs) (AIR, 2010b);
- Developed a Work Sample Test and a Role-Play Exercise to include in the experimental selection battery;
- Identified and procured two Problem Solving measures to include in the experimental selection battery;
- Updated Structured Interview items for inclusion in the experimental selection battery;
- Developed the BDO Job Performance Measure (JPM) and collected existing Performance Accountability and Standards System (PASS) ratings for the criteria in the validation study;
- Pilot tested the selection and criteria measures using AIR staff and at five airports prior to the validation study and analyzed these data to make key decisions regarding the validation study (e.g., procedures to follow, selection measure revisions, forms to administer);
- Conducted validation study at 11 sites and analyzed these data to determine the validity of each selection measure as a predictor of BDO job performance;
- Finalized the selection measures based on results of the validation study; and
- Developed scoring and standard setting recommendations.

The sections below describe important products that resulted from the validation study, as well as key findings.

Important Products

The research resulted in the development of several important products. These include:

- A comprehensive job analysis (AIR, 2010a), including lists of current job tasks, KSAOs and competencies, and tools and equipment, as well as information about the work environment. In addition to guiding the current research, this information can be used to inform training, career development, and other human capital initiatives and programs;
- Several experimental selection measures, including the Passenger Observation, Mental Math, Writing Knowledge, and Visual Recall Assessments, a Role-Play Exercise, and updated Structured Interview questions;
- Critical support documents and other resources for ensuring the proper and standardized administration and use of each of the selection measures, including administration guides, assessor guides, response booklets, and multimedia (e.g., images, videos);
- Training materials for the Role-Play Exercise, including protocols and materials for the Role-Play Exercise Assessors and Resource Person trainings;
- A behaviorally-based performance measure designed specifically for the BDO job and based on the current job analysis (AIR, 2010a);
- Training materials for the BDO Job Performance Measure;
- A scoring technical report (AIR, 2012), including scoring and standard setting recommendations; and
- This validation study technical report, which documents the procedures followed to validate the recommended selection measures, as well as key findings.

Key Findings

As described in Chapter V, the results of the validation study indicated that some of the experimental selection measures performed better than others. More specifically, the current study provided evidence that supports the use of the Passenger Observation Assessment, the Mental Math Assessment, and the Role-Play Exercise. The results of the study also suggested that the Watson-Glaser Critical Thinking Appraisal, the Writing Knowledge Assessment, and the Visual Recall Assessment were less promising for selection purposes. This section summarizes these key findings, beginning with the less promising measures. Next, the benefits afforded by the recommended selection battery (i.e., the Passenger Observation Assessment, the Mental Math Assessment, and the Structured Interview, and the Role-Play Exercise) are described.

Assessments Not Included in Recommended Selection Battery

Results from the validation study suggested that three of the measures in the experimental selection battery were less promising: the Watson-Glaser Critical Thinking Appraisal, the Writing Knowledge Assessment, and the Visual Recall Assessment. First, although the results of the job analysis supported using the Watson-Glaser Critical Thinking Appraisal, ratings from the feedback questionnaires suggested that BDOs did not perceive the connection between this measure and their job. Furthermore, some BDOs cautioned that this assessment may mislead job candidates about requirements of the job (e.g., the type educational background required) which could cause high-quality job candidates to withdraw from the hiring process. Finally, in general, the Watson-Glaser Critical Thinking Appraisal is more susceptible to adverse impact than the other assessments in the experimental selection battery and thus could increase the likelihood of a legal challenge to the BDO selection system (Hough et al., 2001).

The Writing Knowledge Assessment was also identified as less promising. Although this measure demonstrated adequate levels of reliability and criterion-related validity, findings from from the job analysis suggested that Written Communication, the underlying ability assessed by this measure, may not be as critical to BDO job performance as other abilities. This finding was supported by feedback received from BDOs. Specifically, ratings from the feedback questionnaire suggested that other measures in the experimental selection battery were more jobrelated. In addition, BDOs and SPOT Transportation Security Managers (STSMs) indicated that an existing writing measure (i.e., the writing assessment for the Transportation Security Inspector job series) may be more suitable for assessing a candidate's ability to identify and document relevant information in a SPOT Referral report. Furthermore, some BDOs suggested that the SPOT training program may be a more appropriate venue for assessing and developing job-specific writing knowledge and skills.

Finally, with regard to the Visual Recall Assessment, the results of the validation study suggested some limitations with the images for this measure. Specifically, many BDOs commented that the clarity of the images, particularly with regard to the background of the image, made it difficult to see specific passengers or objects. The results of the validation study also revealed a number of problematic items (e.g., items that lacked effective distractors or demonstrated low item-total correlations). In addition to these item-level analyses, the measure demonstrated low levels of internal consistency. Finally, the Visual Recall Assessment failed to diverge from the other measures included in the experimental selection battery (i.e., the Mental Math and Writing Knowledge Assessments). That is, the Visual Recall Assessment overlaps with the other measures in the experimental selection battery and is less likely to contribute unique information to the prediction of BDO job performance.

Given the findings for each of these measures, AIR did not include the Watson-Glaser Critical Thinking Appraisal, the Writing Knowledge Assessment, or the Visual Recall Assessment in the recommended selection battery.

Recommended Assessments

Findings from the validation study provide more support for the Passenger Observation Assessment, the Mental Math Assessment, and the Role-Play Exercise. These three measures demonstrated high levels of content and face validity as indicated by the information gathered during the job analysis (AIR, 2010a) and through feedback questionnaires. In addition, the minimal level of overlap among these assessments suggests that each can contribute unique information when predicting BDO job performance. Though differences existed across forms for each of the selection measures, at least one form of each selection measure was strongly correlated with ratings from either the BDO JPM or PASS. Thus, the Passenger Observation Assessment, the Mental Math Assessment, and the Role-Play Exercise are included in the recommended selection battery, along with the updated Structured Interview. (See the —Recommended Next Steps‖ section for AIR's recommendations for implementing this selection battery).

Proposed Selection System

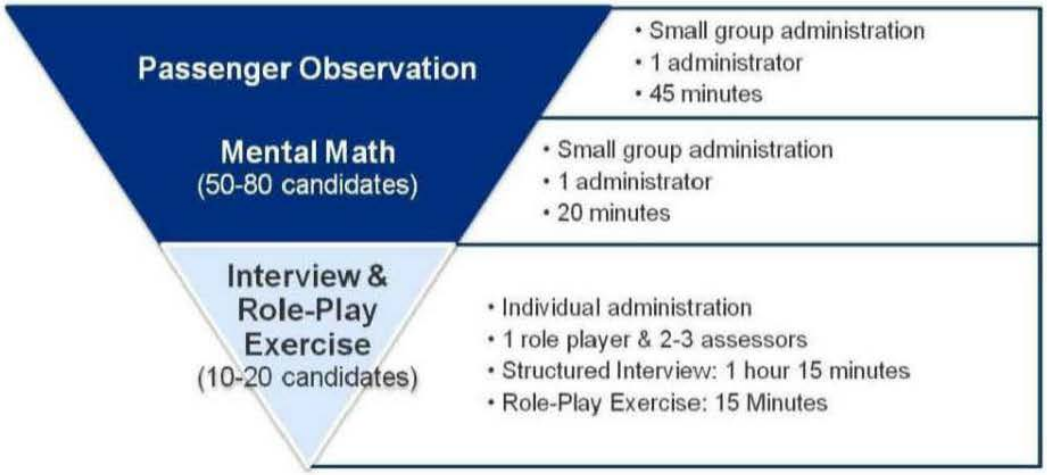
After identifying the most promising selection measures, AIR considered different approaches for combining the measures into a single selection system. One approach consists of using a compensatory model. In this model, the selection measures are weighted and combined into an overall composite. Selection decisions are based on a candidate's overall composite score. Thus, higher scores on one selection measure can compensate for lower scores on another measure. For compensatory measures, weighting schemes should be based on rational criteria, such as choosing to weight selection measures in order to emphasize certain abilities or skills that are critical to job performance (Guion, 1998).

An alternative approach involves a non-compensatory model. In these types of selection systems, job candidates must reach a certain score on each selection measure to be considered for employment. Thus, each selection measure within the battery is a —hurdle‖ that candidates must pass in order to be selected for the job. These types of models are appropriate in situations where each skill and ability measured within the system is uniquely critical for performance in a way that other strengths cannot compensate for weaknesses in them (Guion, 1998).

AIR recommends using a non-compensatory model to implement the recommended selection battery. In the first hurdle, AIR recommends administering the Passenger Observation and Mental Math Assessments to job candidates in small groups (i.e., 3-4 job candidates). Next, job candidates who score above the qualifying scores for these two measures would then complete

TSA's current Structured Interview and the newly developed Role-Play Exercise via a one-onone administration. This recommended system is depicted in Figure 6. The figure includes estimates of the number of job candidates that would complete each hurdle in the selection system using an example for illustrative purposes. This example shows how the number of job candidates could be reduced at each step. [30] The typical administration time and resource requirements for both sets of measures are depicted on the right-hand side of the figure.

Figure 6. Increased Efficiency of the Recommended Selection Battery



This system is recommended for two reasons. First, each of the underlying skills and abilities that are assessed by the recommended selection battery are critical for BDO job performance. As such, demonstrating skill in one area cannot compensate for lack of skill in another. For example, if a candidate demonstrates the ability to perform mental arithmetic quickly and accurately, but is unable to observe and distinguish behaviors or appearance factors, he or she would be unlikely to perform the BDO job successfully. Likewise, if a job candidate demonstrates skill in observing behaviors and appearance factors but is unable to elicit information during a conversation with a passenger, he or she would not likely be a good fit for the BDO job.

Second, by reevaluating job candidates after the first hurdle, TSA would be able to reduce the costs associated with administering the Structured Interview. While collecting data during the

[30] This example assumes that 50-80 job candidates are initially considered for a few job openings during one selection cycle. During the conduct of the validation study, some sites reported that they were interviewing 50-80 job candidates for 2-3 job openings.

validation study, some sites commented that they were interviewing up to 80 candidates for 2 – 5 job openings. Because each Structured Interview requires at least two Assessors and can take 45 to 75 minutes to administer, this can be a time-consuming and resource-intensive process. By including a preliminary hurdle that consists of measures that can be administered in small groups, TSA can reduce the number of candidates to be interviewed to fill a BDO job opening. Furthermore, because the Role-Play Exercise was designed to be administered immediately after the Structured Interview, the only additional resource requirement for that phase of the system involves procuring and training a role-player.

The system would also provide increased efficiency by providing job candidates with a realistic job preview (Premack & Wanous, 1985). Specifically, by placing the Passenger Observation and Mental Math Assessments early in the administration process, job candidates will have more information to determine whether they are qualified for or interested in the BDO. Some job candidates who are not suited for the BDO job may withdraw from the selection process at this stage and thereby reduce the number of candidates that need to be interviewed, thus reducing the cost and resources required to fill the job opening.

Recommended Next Steps

As described above, the results from the validation study suggests that three of the selection measures are likely to be useful components of the BDO hiring system. Although the validation process described in the previous chapters is the final phase of the current study, AIR is recommending several additional steps. These include:

- Gathering additional data regarding the recommended selection battery from job candidates using a predictive validation design;
- Considering best practices when implementing the final system;
- Implementing the updated Structured Interview; and
- Using the BDO JPM as part of ongoing or future research efforts.

Each of these recommendations is described in more detail below.

Evaluate the Recommended Selection Battery Using a Predictive Design

This section describes AIR's recommendation to collect additional data for each of the recommended selection measures using a predictive design. The section begins by describing the purpose of the recommended study and concludes by reviewing a proposed approach for conducting the study.

Purpose of Proposed Predictive Study

As described in Chapter V, the findings from the validation study provide initial support for the recommended selection battery (i.e., the Passenger Observation Assessment, the Mental Math Assessment, the Structured Interview, and the Role-Play Exercise). Due to the small sizes and the observed differences across forms, however, AIR is recommending that TSA collect additional data to further evaluate these assessments. Specifically, AIR is recommending the conduct of a predictive, criterion-related validation study using the recommended selection battery (McPhail & Stelly, 2010). A predictive design would involve administering the selection battery to BDO job candidates but only making hiring decisions using the updated Structured Interview. After the newly hired BDOs had accumulated enough job experience, TSA could collect performance data for these employees using either training assessments, PASS, and/or the BDO JPM. These performance ratings could then be correlated with the original scores on the recommended selection battery.

Collecting additional data using a predictive design would address two limitations in the current study. First, it would help address the range restriction observed for the current sample of BDOs (McPhail & Stelly, 2010). Because these BDOs were experienced (average tenure = 2.93 years, *SD* = 1.17), had been previously screened using the existing Structured Interview, and have undergone extensive training, AIR anticipates that their scores on some selection measures are likely to differ from scores from a sample of job candidates (Van Iddekinge & Ployhart, 2008). For example, BDOs engage passengers in conversation on a daily basis. Thus, their performance on the Role-Play Exercise is likely higher than what would be observed for a sample of job candidates. Also, the job performance ratings provided for this sample of BDOs are likely higher and more consistent than what would be observed among a broader sample of job candidates or a sample of recent BDO hires.

The observed level of range restriction likely attenuated the relationships between scores on the selection measures and the job performance ratings. AIR attempted to account for this attenuation by using established methods for correcting for range restriction (Hunter et al., 2006). Nevertheless, administering the selection measures to a sample of job candidates would allow TSA to directly evaluate the performance of these measures among a sample of job candidates. Specifically, TSA could correlate scores on the selection measures from recent hires with their subsequent job performance ratings. Furthermore, the data from administering the selection measures to actual job candidates would provide a better source for estimating the true relationship between the selection measures and job performance. These direct calculations would likely provide a more accurate estimate of the degree of criterion-related validity for the recommended battery.

Second, collecting additional data using a predictive design would help increase the sample size for each of the selection measures. In the current study, sample sizes for each form were relatively small (i.e., approximately 100 incumbents per form). These sample sizes make it difficult to determine if the form differences observed in the current study should be attributed to differences in test content or the BDOs that completed each form. In other words, it is possible that the observed form differences may be attributed to differences in items on each selection measure or unusual patterns of test scores among a small number of BDOs. Ideally, in the predictive validity study, predictor data would be collected from at least 250 or more job candidates per form. This would allow for a more complete evaluation of the observed form differences.

In addition to addressing some of the limitations with the current study, a predictive design would provide TSA with an opportunity to identify and address operational issues while implementing the recommended selection battery nationwide program. These issues may include calibrating equipment across airports, securing space for group administrations, and identifying and training personnel to serve as Administrators or Assessors. Some of these issues arose during the validation study. For example, six airports were unable to secure the videobased equipment needed for the Passenger Observation and Mental Math Assessments. Also, one site had difficulty securing the space for the separate administration sessions. By collecting only experimental data while initially implementing the recommended system, TSA could identify and resolve administration issues before the recommended selection battery is used to inform hiring decisions.

Method of Proposed Predictive Validation Study

AIR recommends continued data collection for a predictive, criterion-related validity study. The order of administration (depicted in Figure 7) corresponds with the recommended approach for enhancing the efficiency of the existing selection battery.

Figure 7. Order of Administration for Predictive, Criterion-related Validity Study

First, AIR recommends job candidates complete the Watson-Glaser Critical Thinking Appraisal. Although this measure was not included in the recommended selection battery, it could serve as a comparison measure for the predictive validation study (Russell & Peterson, 1997). Specifically, TSA could evaluate whether the recommended selection measures (i.e., the Passenger Observation Assessment, the Mental Math Assessment, and the Role-Play Exercise) correlate with the Watson-Glaser Critical Thinking Appraisal measure in patterns that are similar to what was observed in the current study and what could be expected based on the literature. TSA could use these data to further examine the form differences observed in the current study.

Immediately following the Watson-Glaser Critical Thinking Appraisal, job candidates would complete the Passenger Observation Assessment and the Mental Math Assessment. These assessments, like the Watson-Glaser Critical Thinking Appraisal, can be administered in a group format. Thus, multiple job candidates could participate in a single testing session.

Next, job candidates would participate in the Structured Interview and the Role-Play Exercise.[31] Similar to the validation study, AIR is recommending that job candidates complete these

183

assessments individually. Finally, hiring decisions would be based only on a job candidate's performance on the Structured Interview.

AIR also recommends continuing to use the feedback questionnaire during the predictive validation study. The data collected using this questionnaire would allow TSA to determine if job candidates, like BDOs, perceive the measures as job-relevant. Also, some of the instructions may be rated as clear by BDOs, but could be confusing to job candidates because they lack contextual knowledge. Perceptions of clarity and job-relatedness can impact important outcomes such as the likelihood of a candidate accepting an offer or referring other job candidates to apply to an organization (Hausknecht et al., 2004).

These test and feedback data could be used in a sequential manner. Initially, job candidates' scores on the recommended selection battery could be examined using descriptive, reliability, and construct validity analyses. These results could first be used to more accurately estimate the degree of range restriction in the sample of job incumbents in the current study. Then, the improved estimate of range restriction could be used to recalculate the corrected criterion-related validity coefficients from the current study. In addition, if an adequate sample size is obtained, these data could also be used to conduct preliminary adverse impact analyses (e.g., subgroup differences). Finally, these data would allow TSA to conduct preliminary analyses to examine if the form differences observed in the current study can be attributed to differences in test content.

After hiring decisions are made based on the Structured Interview, TSA could collect job performance data. These data may include test scores from their initial SPOT training, ratings from TSA's PASS, and STSM ratings using the BDO JPM. Each of these job performance measures is likely to contribute unique information when evaluating the recommended selection battery. For example, training scores would likely indicate a candidate's propensity to learn new information while BDO JPM ratings would correspond to their ability to perform job-related duties (Schmidt & Hunter, 1998).

[31] Although the Conversation Skills rating scale did not predict job performance in the current study, it did appear to diverge from the Elicitation of Information rating scale (r = .45-.46). Given that these scales appear to assess unique aspects of a candidate's performance, AIR recommends including it in the predictive validation study as an

211

As seen in the current study, it is critical that STSMs have an adequate opportunity to observe newly hired BDOs' performance. Specifically, it is important that raters have an opportunity to observe the new hires perform a range of job duties (e.g., performing SPOT, writing reports, conducting referrals). These opportunities can impact the quality of managers' ratings. Thus, the new hires should have at least three to six months of job experience as a BDO before these ratings are collected. After these ratings are collected, scores from the selection measures could then be correlated with performance data to determine the predictive validity of the selection measures. Because scores on the selection measures would be based on job candidates' performance, these coefficients would provide a direct estimate of the operational validity for each of recommended selection measures.

experimental measure. It is possible that the low levels of criterion-related validity may be attributed to small sample sizes.

Consider Best Practices if Implementing the Recommended Selection Battery

If the results of the predictive, criterion-related validity study provide evidence supporting the use of the recommended selection battery, AIR recommends implementing this system. This section describes AIR's recommendations for the implementation process. These include:

- Provide information to job candidates about the BDO job prior to or early on in the selection process;
- Orient STSMs to the new selection measures;
- Consider approaches for minimizing cheating or dishonest test behavior;
- Standardize data management and record keeping processes; and
- Communicate the results of both the concurrent and predictive validation studies to all the participants.

Provide Information to Job Candidates Prior to the Selection Process

During the data collection process, subject matter expert (SMEs) indicated that TSA has been trying to ensure more information about the job is provided to internal job candidates. These efforts have improved the overall communication process. In order to build on these initial improvements, AIR recommends providing additional information to job candidates prior to the selection process. For example, job candidates would likely benefit from a comprehensive job description. While collecting data for the job analysis and validation study, SMEs noted that the following requirements are not currently emphasized in the current BDO job description: amount

of (1) standing/walking, (2) sustained attention, and (3) interaction with the public. For example, some SMEs indicated that the BDO job often requires incumbents to stand on their feet and interact with passengers for an entire shift. Others noted the need to perform repetitive mental activities while maintaining a high level of accuracy.

These descriptions could enhance the efficiency of the hiring system by reducing the number of unqualified or poor fit candidates who apply for the BDO job. Specifically, SMEs suggested that providing a more detailed job description would provide candidates with the opportunity to withdrawal if they were not interested in the position or did not believe they could perform the requisite duties.

In addition to a comprehensive job description, AIR recommends providing job candidates with read-ahead materials that describe the purpose and procedures involved in the selection process. These materials would likely have a similar effect as the job description. Specifically, job candidates who were not interested in completing the necessary selection measures, or did not believe they were qualified, may choose not to apply.

Finally, AIR recommends providing additional information to internal TSO candidates about the consequences of withdrawing from the SPOT program. For example, some SMEs noted that some candidates were unaware that their seniority would be lost if they transferred into the BDO program but later chose to return to their previous position. SMEs commented that this policy encourages some incumbents, who may not demonstrate adequate levels of performance, to remain within the SPOT program. Also, SMEs suggested that providing information about this policy earlier in the hiring process may have prevented less qualified job candidates from applying. In general, providing clear information about the BDO job, the selection process, and the career path may help increase efficiencies in the hiring process.

Orient STSMs to the New Selection Measures

In addition to providing information to job candidates, AIR recommends informing STSMs about the new system. During the validation study, SMEs stressed the importance of ensuring that adequate guidance and communication be provided to STSMs regarding the recommended selection system. For example, information about the purpose, the development process, and the administration procedures should be provided to SPOT managers at all airports. This information may be shared during open forums that are held on-site or remotely. During these forums, STSMs can learn about the new system and ask questions. STSMs may also be encouraged to share this information with their airport leadership. This type of communication will increase the likelihood that the system is well received and implemented appropriately.

In addition to providing information about the recommended selection measures, some SMEs suggested providing guidance on how to integrate BDOs that are hired with the new system into the existing SPOT program population. These SMEs commented that BDOs who are selected using the recommended selection battery may perform at a higher level. Some SMEs recommended that TSA provide guidance on how to merge and integrate groups of incumbents hired under different systems. This guidance may consist of training on managing organizational change.

Consider Approaches for Minimizing Cheating or Dishonest Test Behavior

During the group feedback sessions, BDOs provided suggestions for how to reduce the likelihood of candidates cheating during the selection process. Some BDOs suggested that it may be helpful to provide barriers between job candidates during the administration of the Work Sample Test (i.e., the Passenger Observation Assessment and the Mental Math Assessment).
These barriers would prevent job candidates from recording another candidate's responses. Also, these barriers would not allow candidates to use another test-taker's behaviors as cues during the administration process. For example, job candidates could not see when another person records his or her response during the Passenger Observation Assessment and determine which passengers to observe.

Standardize Data Management and Record Keeping Processes

Data management and record keeping is critical for ensuring the effectiveness of a selection system. Ongoing data collection efforts allow organizations to continually evaluate the reliability and validity of their selection measures (Messick, 1989). Furthermore, data management and record keeping can be especially important for organizations faced with a legal challenge.

During the current study, AIR made several requests for archival data (e.g., PASS ratings, training scores, selection ratios). Although TSA attempted to fulfill each of these requests, some challenges arose with accessing certain types of data. For example, training scores for incumbent BDOs were not readily available. Likewise, TSA was unable to provide estimates of the selection ratios for some of their SPOT programs.

Many of these challenges may be attributed to a paper-based data management system. Specifically, some sites and offices rely on paper-based record keeping. Although this approach requires less training and electronic resources, it may limit the efficiency with which data can be shared throughout the organization. Implementing a standardized, electronic process for managing and retaining data nationwide will allow TSA to maximize the utility of all the components of its human resource management system. For example, TSA could examine empirically-based estimates of selection ratios, career progression, and training performance. This information could then inform future pilot tests and human capital initiatives.

Accomplishing this goal will likely require coordination across multiple TSA offices. However, this information would benefit a variety of stakeholders (e.g., Office of Security Operations [OSO], Office of Human Capital [OHC]).

Communicate the Results of the Validation Studies to Participants

Finally, at the conclusion of the predictive, criterion-related validation study, AIR recommends communicating the results of both studies to all SMEs who participated. Providing summary reports about the findings and outcome of the research demonstrates that the organization has taken action on the basis of the information and input provided. Furthermore, sharing the findings of these studies increases the likelihood that SMEs will be willing to participate in future research efforts.

Implement Updated Structured Interview

AIR recommends integrating the updated Structured Interview items into the current BDO selection system. These new behaviorally-based items increase the number of forms available and improve the test security of the Structured Interview. In addition, AIR recommends including the situational items designed to assess Honesty-Integrity. These items showed adequate levels of variability during the external pilot test. Furthermore, these items were reviewed by SMEs and found to assess a competency that is critical for BDO job performance. In addition to these new items, AIR has proposed modifications to competency probes, definitions, rating scales, and administration processes. Each recommendation was designed to increase the standardization of the administration of the Structured Interviews across airports. These changes also were designed to increase the ease with which Assessors can make accurate and reliable evaluations of job candidates using the Structured Interview Administration Guide (see Chapter IV).

Consider use of BDO Job Performance Measure for Other Research Studies

Along with the updated Structured Interview, AIR recommends including the BDO Job Performance Measure (JPM) in future research studies. This measure was developed using the results of the job analysis (AIR, 2010a) and revised based on the feedback of SMEs. In addition to measuring job-relevant aspects of performance, the BDO Job Performance Measure demonstrated adequate levels of variability. Specifically, STSMs used the entire rating scale while providing their ratings. Furthermore, STSMs demonstrated adequate levels of inter-rater agreement. Given these encouraging findings, it is likely that the BDO Job Performance Measure could provide useful information in future research studies. For example, the BDO Job Performance Measure could be used for current or future pilot projects or job analyses. In addition, once TSA has finished evaluating the recommended selection system, the BDO Job

Performance Measure could be used in conjunction with the current PASS system to inform administrative decisions (e.g., promotions, salary adjustment). The BDO Job Performance Measure may be especially helpful as an administrative measure because it appears to assess different aspects of job performance than PASS.

CHAPTER VIII: SETTING QUALIFYING SCORES FOR THE RECOMMENDED SELECTION BATTERY

When implementing the recommend selection measures, TSA will need to conduct a formal process for establishing qualifying scores. A qualifying score for a selection measure is the minimum score an individual must achieve on the test in order to be considered qualified for the job (Cizek & Bunch, 2007). Qualifying scores are influenced by multiple factors such as test validity, success rate, numbers of position openings and applicants (i.e., the selection ratio), and other organizational factors (Guion, 1998; SIOP, 2003).

As described in Chapter V, the results of the validation study indicated that some of the experimental selection measures performed better than others. More specifically, the current study provided evidence that supports the use of the Passenger Observation Assessment, the Mental Math Assessment, and the Role-Play Exercise (see Chapter VII). However, due to observed differences across forms, AIR is recommending that TSA collect additional data to further evaluate these assessments. Specifically, AIR is recommending the conduct of a predictive, criterion-related validation study using the recommended selection battery (McPhail & Stelly, 2010) (see Chapter VII).

Because AIR has recommended that the proposed selection battery undergo additional evaluation using a predictive, criterion-related validity design, the qualifying scores described in this chapter of the validation report should be considered preliminary. Specifically, these qualifying scores are meant to illustrate a process that could be used once additional data have been collected. Nevertheless, these preliminary qualifying scores may be useful for informing future standard setting efforts.

This chapter begins by summarizing the different approaches that AIR considered for the current study. Next, the process used to identify preliminary qualifying scores for the individual assessments is reviewed. Then, an example for how to combine these separate preliminary qualifying scores for a single selection battery is presented. The chapter concludes with recommendations for monitoring the effectiveness of qualifying scores. Additional information for each step of the process for establishing qualifying scores is presented in a separate scoring technical report (AIR, 2012).

Processes Considered for Establishing Preliminary Qualifying Scores

A number of approaches have been proposed for setting qualifying scores (Cizek & Bunch, 2007). In general, the approaches for establishing qualifying scores can be classified as either empirical or rational methods (Johnson & Oswald, 2010). This section describes each approach and summarizes some of the advantages and disadvantages of each.

Rational Methods

Rational methods rely on estimates and judgments provided by SMEs (Cizek & Bunch, 2007; Kehoe & Olson, 2005). Typically, SMEs consider each item, or a group of items, on a selection measure and estimate the proportion of minimally competent job applicants who would answer that item correctly (Angoff, 1971; Cizek & Bunch, 2007; Truxillo, Donahue, & Sulzer, 1996). Typically, these estimates are first provided individually. After the first round of estimates, panelists discuss their ratings and review additional data (e.g., average test scores). After receiving this additional information and discussing their ratings, panelists provide a second set of individual ratings. Researchers then use the estimates for each item to develop an overall qualifying score. [32] For example, if the average proportion of minimally competent job applicants to answer each item correctly across a measure was 75%, and the measure consisted of 20 items, the initial qualifying score would be set at 15.

[32] Note that there are many variations to this general approach; a general overview has been presented for the purposes of this report.

Although these approaches leverage SMEs' understanding of the job and the assessment, they typically require a substantial amount of training (Cizek & Bunch, 2007). Panelists must develop a clear concept of the term —minimally competent job applicantl and understand the knowledge, skills, and abilities required for the job. Also, panelists may be asked to complete each assessment in the test battery so they understand the items they are evaluating. In addition to a comprehensive training process, some researchers have indicated that panelists may struggle with these approaches when items are very easy or very difficult (Mueller, Norris, & Oppler, 2007).

Empirical Methods

Empirical methods rely on observed test scores for a group of test-takers (Kehoe & Olson, 2005; Mueller et al., 2007). For example, the borderline and contrasting groups methods consider the distributions of scores for successful and unsuccessful incumbents. Initially the qualifying score is set at the point where the two distributions intersect. Regression-based methods, on the other hand, usually employ linear regression equations to identify a test score that corresponds with an optimal level of job performance (Kehoe & Olson, 2005). Finally, expectancy charts provide visual depictions of the relationship between the selection measures and the job performance

ratings (Mueller et al., 2007). Specifically, these charts display the level of job performance associated with each observed score for a selection measure.

The effectiveness of each of these approaches varies (Kehoe & Olson, 2005; Mueller et al., 2007). For example, the borderline groups method is usually intuitively appealing for experts. However, it can be challenging to implement if there is a large amount of overlap between the distribution of test scores for successful and unsuccessful incumbents. Regression-based approaches tend to yield different results and often require large sample sizes in order to provide precise qualifying scores. Finally, expectancy charts are usually both intuitive and often well received by stakeholders. Furthermore, this approach has received support from the courts during legal challenges (Kehoe & Olson, 2005; Mueller et al., 2007). Nevertheless, expectancy charts may be difficult to interpret if there is not a clear relationship between the selection measures and job performance.

Current Approach

Although some approaches for setting qualifying scores have been studied more extensively than others, researchers recommend considering both the types of data that are available and the amount of access one has to subject matter experts (Cascio, Alexander, & Barrett, 1988). After considering the different approaches within each category, AIR used an empirical method for three reasons. First, the number of experts and amount of time required to implement a rational method would have been prohibitively costly. For instance, AIR estimated that it would take at least 6-10 subject matter experts three to five 8-hour workdays to complete a judgment based standard setting process for the entire experimental selection system. This panel would need to consist of SMEs with a diverse understanding of the BDO job and the hiring process (e.g., trainers, SPOT Transportation Security Managers [STSMs], human resource analysts). Furthermore, ideally the panel would need to represent a geographic and demographic sample of relevant SMEs. Convening such a group of SMEs for an entire workweek was not feasible given the current study's timeline and existing resources.

Second, it is likely that SMEs would have had difficulty using the content of some of the measures to estimate qualifying scores. For instance, estimating qualifying scores on the Passenger Observation Assessment using individual items may have been challenging because each item consists of a separate behavior or appearance factor that is displayed at certain points in the video while other behaviors are displayed on screen. Thus, the difficulty of an item is dependent on the context in which it occurs. SMEs who review individual items, or subsets of items, may provide inaccurate estimates unless they considered the assessment holistically.

Third, during the validation study, AIR collected data in order to examine the relationship between selection measure scores and job performance ratings. These data are typically the cornerstone for many empirical methods used to establish qualifying scores. Methods that incorporate empirical information are usually judged to be superior to methods that rely solely

on experts' judgment (Mueller et al., 2007). Thus, AIR chose to leverage the existing validation study data to inform preliminary judgments about qualifying scores.

Based on these considerations, and the benefits of using empirical methods, AIR developed expectancy charts for setting preliminary qualifying scores for the current study. AIR anticipated that multiple stakeholders may be involved in examining the preliminary qualifying scores after the conclusion of this study. Thus, it was critical to adopt an empirical approach that was both technically sound and accessible to future stakeholders. Expectancy charts, as indicated earlier, are often characterized as some of the most accessible methods for establishing qualifying scores. Furthermore, this approach has been supported by both researchers and the courts (Kehoe & Olson, 2005; Mueller et al., 2007).

Review of Preliminary Qualifying Scores

This section provides an example of how AIR developed the preliminary qualifying scores for the proposed selection battery using expectancy charts. In the scoring technical report (AIR, 2012), we illustrate how this method could be used to identify low, moderate, and high preliminary qualifying scores for the Passenger Observation Assessment, the Mental Math Assessment, and the Role-Play Exercise. Table 77 summarizes this process.

Table 77. Steps for Developing Expectancy Charts for Individual Selection Measures

Step	Description
Identify Relevant Job Performance Criteria	AIR included means, standard deviations, and frequencies of composite scores for Performance Accountability and Standards System (PASS) and the BDO Job Performance Measure (JPM).
Record Frequency of Observed Scores for the Selection Measure within the Expectancy Chart	AIR recorded the frequency of each observed score for a particular selection measure. For example, AIR recorded the number of BDOs that received a particular score on the Passenger Observation Assessment (e.g., the number of BDOs who received a score of 18).
Record Mean Values and Frequencies of Job Performance Criteria for Each Observed Score on the Selection Measure	AIR recorded the mean value and frequency for the composite scores on the PASS and BDO JPM for each observed score for a particular selection measure.
Examine the Expectancy Chart to Identify Low, Moderate, and High Preliminary Qualifying Scores	AIR examined the expectancy chart to determine at which point on the distribution of selection scores job performance ratings began to vary.

This process was repeated for each form of the recommended selection measures. AIR then examined the separate expectancy charts to identify low, moderate, and high preliminary qualifying scores.

Using the low, moderate, and high preliminary qualifying scores identified for each selection measure, AIR developed expectancy charts for the recommended selection battery. These charts were developed to demonstrate the impact of setting qualifying scores for the entire recommended selection battery. The expectancy charts only pertain to BDOs who completed each of the selection measures included in the recommended battery. Specifically, BDOs who completed either form of the Passenger Observation Assessment, and the Role-Play Exercise, as well as Mental Math – Form A were included (n = 111). For the current chapter, the expectancy chart for the low preliminary qualifying scores is presented below. Examples of moderate and high preliminary qualifying scores are presented in a separate scoring technical report (AIR, 2012).

It is important to note that these charts were developed using the test scores from incumbent BDOs. Thus, these data have been used to illustrate an approach that could be carried out for data collected from job applicants if a predictive study is conducted.

Example Preliminary Qualifying Scores

An example of an expectancy chart for the recommended selection battery with low preliminary qualifying scores is presented in Table 78. This system consists of the three low qualifying scores that were identified using the expectancy charts for individual selection measures. Specifically, for this system, the qualifying score for the Passenger Observation Assessment was 8, for Mental Math – Form A it was 12, and for the Role-Play Exercise it was 4.

Overall, 33% of the BDOs in the current sample (n = 37) would be rejected using the current set of qualifying scores. In other words, this system would yield a selection ratio of 66%. Nevertheless, this selection battery still improves the efficiency of TSA's hiring system. Specifically, by examining the selection ratio after Mental Math – Form A, one can see that TSA would be able to interview 27% fewer candidates. For medium and large-sized airports, this could significantly reduce the costs of their hiring process.

Table 78. Example Expectancy Chart for the Experimental Selection System

(b)(3):49 U.S.C. § 114(r)

[33] PASS Level categories were established by TSA leadership based on the distribution of total scores on the PASS for all incumbents during the performance period. (b)(3):49 U.S.C. § 114(r)

(b)(3):49 U.S.C. § 114(r)

combining different selection measures can impact the anticipated performance of job candidates.

Additional examples using more stringent qualifying scores are presented in the scoring technical report (AIR, 2012).

Ongoing Evaluation of Preliminary Qualifying Scores

Qualifying scores represent decision points that may change depending on organizational circumstances (Guion, 1998). Thus, it is helpful to conceptualize the scores presented in this chapter as preliminary qualifying scores. Furthermore, it is likely that they will need to be revised as more information on the selection measures and other organizational factors becomes available. This section describes some organizational factors that should be considered before finalizing preliminary qualifying scores.

Applicants' Performance on Selection Measures

The current qualifying scores were based on test scores for BDO incumbents. Thus, these scores should be interpreted cautiously for two reasons. First, the sample of test-takers in the current study had already been screened prior to being included in the study. The previous screening measures most likely removed many of the low-performing BDOs. Specifically, prior to the current validation study, BDOs had successfully completed a structured interview and SPOT training. Furthermore, on average, BDOs who participated in the current study had 2.93 years of experience in the SPOT program (SD = 1.17) and received slightly elevated performance ratings. Thus, it is unlikely that, as a group, BDOs' scores on the selection measures and job performance ratings would show the same levels of variability as scores from a group of job applicants.

Another factor to consider regarding the use of applicant data for setting qualifying scores is motivation and its effect on test scores. For example, despite efforts to properly motivate study participants (i.e., job incumbents), there is no direct way to determine whether they put forth the effort to perform their best. Conversely, nearly all job candidates will be motivated to perform their best to obtain a position for which they are applying. To the extent that not all study participants put forth their best effort, it may be the case that higher job performers received lower test scores than which they are capable, which can ultimately affect preliminary qualifying scores.

While developing the preliminary qualifying scores, AIR considered how BDOs' job experience and training could affect their performance on the selection measures. On some selection measures BDOs' performance may have been higher than what may typically be observed for

applicants. For instance, most BDOs received high ratings for the Role-Play Exercise. Because the Role-Play Exercise resembles job duties that BDOs perform on a daily basis, it is likely that their scores would be higher than those observed for most applicants. AIR has recommended that TSA further examine the psychometric properties of these selection measures using a predictive validation design. Thus, the data and approach presented in this report is designed to provide a framework for future users to consider when re-examining the qualifying scores using applicant data. Given these considerations, it is recommended that TSA continue to monitor applicants' scores on the selection measures and before finalizing the preliminary qualifying scores.

Potential Impact on Workforce Diversity

TSA will also want to consider the degree to which qualifying scores impact the diversity of their workforce (Kehoe & Olson, 2005). For example, organizations often emphasize a diverse workforce in order to achieve particular business goals. Because the Passenger Observation Assessment, the Mental Math Assessment, and the Role-Play Exercise are all high-fidelity assessments that simulate tasks on the BDO job, they are less likely to lead to adverse impact than typical measures of problem solving or general cognitive ability (Roth et al., 2008). Nevertheless, measures that assess specific cognitive abilities, like the Passenger Observation and Mental Math Assessments, may result in differential hiring rates between members of particular subgroups (e.g., ethnic, gender) (Hough, et al., 2001). Although AIR considered the likely impact of the preliminary qualifying scores on hiring rates for different subgroups, the sample sizes for particular subgroups in the current study (e.g., Blacks versus Whites versus Hispanics) were too small to provide stable estimates of adverse impact or differential validity. Thus, it is recommended that TSA examine the impact of the preliminary qualifying scores on the hiring rates of different subgroups and protected class members as additional data from job applicants becomes available.

CHAPTER IX: FUTURE CONSIDERATIONS

Throughout the validation study, the AIR collected feedback from BDOs and their managers to inform the development and implementation of the recommended selection battery. During these discussions, subject matter experts (SMEs) also provided recommendations for training, performance management, and the BDO job in general (e.g., career progression and duties of the BDO job). These recommendations are documented for TSA's consideration and include:

- Converting the Work Sample Test measures to an electronic and un-proctored administration format;
- Examining the utility of a personality measure (to be used for selection purposes);
- Examining the advantages and disadvantages of using an external applicant pool for the SPOT program;
- Evaluating the validity of the recommended selection battery for selecting candidates into the SPOT Assessor Program;
- Using job analysis data to conduct a training alignment study;
- Developing recurring training/performance measures modeled after the experimental selection battery;
- Examining the BDO career path;
- Examining the human capital processes used for contractor BDOs;
- Investigating the reciprocal relationship between a BDO's behavior and a passenger's behavior during the screening process; and
- Examining the affect of fatigue on a BDO's vigilance.

In this chapter, each of these areas for future consideration is described in more detail, organized by topic (i.e., selection, training and performance management, and the BDO job in general).

Selection

This section presents areas for future consideration that relate to selecting BDOs.

Conversion of the Work Sample Test to an Electronic and Un-Proctored Format

During group feedback sessions conducted during the pilot test and the validation study, several SMEs recommended administering the Work Sample Test using an electronic and un-proctored format. SMEs noted that the automated assessments would eliminate the need for a

Administrator and hand entry of job candidates' test responses. These changes would increase efficiency and reduce the costs associated with these measures.

In addition to the improved efficiency, these modifications may help address challenges with the Passenger Observation Assessment experienced by participants. Specifically, though the majority of participants performed very well on the assessment, several SMEs provided feedback that they had difficulty (1) recording the observed behaviors and appearance factors in their response booklets and (2) identifying and following the passengers throughout the screening line. Creating an electronic version of this assessment would allow job candidates to provide their responses directly on the screen and may address challenges related to providing responses. For example, capabilities that allow job candidates to highlight behaviors and appearance factors of interest as they watch the video could be added, which would minimize the amount of time job candidates divert their attention from the screen to record responses or read passenger descriptions. In addition, a mechanism could be added to highlight passengers of interest throughout the duration of the video (rather than only for the first few seconds the passenger is on the screen), which would eliminate challenges associated with identifying and following passengers. SMEs commented that these changes would likely improve perceptions of fairness of the assessment. Note that, if an electronic and un-proctored version of the Work Sample Test is developed, methods for preventing cheating and addressing job candidates' questions would need to be considered. One way to combat the potential issues of un-proctored administration is to create an electronic version of the assessment that is still administered in a proctored environment.

Examination of the Utility of a Personality Selection Test

Throughout the validation study data collection, several SMEs inquired about the utility of incorporating a measure of personality into the BDO hiring system. Personality measures are typically self-report questionnaires that provide an evaluation of the characteristics that make a job candidate unique (Gatewood, Feild, & Barrick, 2008). These characteristics (e.g., what a person habitually wants, does, says, feels) often interact with an individual's environment to influence his or her behavior. Several knowledge, skills, abilities, and other characteristics (KSAOs) that were identified and prioritized for selection during the job analysis process specifically related to conscientiousness (i.e., responsible, organized, dependable, willing to achieve, and persevering [Gatewood et al., 2008]). These KSAOs included detail orientation;

skill in working with minimal or no supervision; and ability to work in a reliable, responsible, and dependable manner. Given this finding, personality measures that specifically measure whether a person is conscientious may be considered job-relevant.

During the test specification phase of the validation study, AIR opted not to include a measure of personality for two primary reasons. First, discussions with key stakeholder groups indicated that there may be some resistance to implementing a personality measure for the selection of BDOs. Furthermore, AIR determined that several of the KSAOs related to conscientiousness were currently measured in the current BDO selection system (via the Structured Interview). In addition to these primary reasons, other disadvantages regarding the use of personality measures AIR considered included potential negative job candidate reactions (Anderson & Witvliet, 2008), susceptibility to faking (Morgeson et al., 2007; Ones, Dilchert, Viswesvaran, & Judge, 2007; Rothstein & Goffin, 2006) and cost.

Despite these disadvantages, there are several reasons TSA may want to consider the utility of a personality measure for BDO selection in the future. First and foremost, as stated above, several of the KSAOs rated as highly important and linked to the job were related to the personality construct of conscientiousness. Second, personality measures typically result in only small subgroup differences (Foldes, Duehr, & Ones, 2008; Outtz, 2002) and, thus, result in fewer instances of differential prediction (i.e., differential hiring rates between members of particular subgroups [e.g., ethnic, gender]). In addition, personality measures are typically easy to administer and do not require extensive training. With regard to cost, proprietary personality measures require minimal financial investment once initial purchase fees are paid. Finally, and most notably, a personality measure of conscientiousness may diverge from the measures included in the recommended selection battery that relate to cognitive ability (i.e., the Passenger Observation Assessment and the Mental Math Assessment) and thus could provide some additional predictive capabilities beyond the measures currently included. As such, future consideration may be given to the inclusion of a personality measure in the BDO hiring system and/or in the predictive, criterion-related validation study that AIR recommends as part of recommended next steps described in Chapter VIII. Refer to the BDO Test Specifications report (AIR, 2010b) for additional descriptive, psychometric, and operational information about personality measures as well as specific personality measures to consider.

Consideration of Feedback Regarding Use of External Job Candidates

Presently, to become a BDO, current TSA employees must be internally promoted from their position as a TSO. Should the recruitment and selection process for BDOs be open to external job candidates, TSA may wish to consider the input and feedback SMEs provided about this issue. This information is briefly discussed below.

Overall, SMEs agreed that use of external job candidates would increase the pool of eligible job candidates. That is, a wider variety of job candidates may be interested in pursuing the more specialized BDO position if they are not required to first serve as a TSO, potentially increasing the diversity of backgrounds among job candidates. Nevertheless, SMEs provided a few recommendations for effectively implementing this change. First, SMEs suggested that it will be important to consider how external job candidates can be recruited and trained so that they are as competent as internal BDO job candidates. For example, some SMEs suggested that external job candidates should be required to serve in a conditional role as a TSO for a given period of time 200

or, at very least, that they be required to complete TSO training. This is because fundamental TSA knowledge and skills learned while serving as a TSO are also essential for performing as a BDO (e.g., knowledge of the differences in roles of checkpoint personnel, knowledge of the behavior of checkpoint personnel that can affect a passenger's demeanor/behavior).

Alternatively, other SMEs suggested that external job candidates be given a conditional offer, during which they are required to pass a probationary period (e.g., of 6 months to a year) before they officially become a BDO. If such a conditional hiring process is implemented, SMEs suggested that it may be important to consider issues related to the classification of the conditional position (i.e., job title) and pay grade (i.e., whether external job candidates awarded a conditional offer receive the same pay as an internal job candidate who is not required to pass a probationary period). Alternatively, TSA may consider extending the initial training program for external job candidates. This training could focus on the type of information that BDOs acquire while serving as TSOs. This approach may reduce the time period of the probationary period, but could require additional training costs. For either approach, TSA may wish to consider the potential legal implications of establishing a separate career path for external job candidates.

In addition to the SME input described above, if external job candidates are included in the BDO hiring pool, AIR also suggests that the recommended selection battery be reviewed to determine whether it will require modifications. Though all assessments were developed in a manner that would permit their use for selection of internal or external job candidates, additions could likely be made to further assist external job candidates. For example, for the administration of the Role-Play Exercise during the pilot test, AIR had included an image of a BDO in uniform to help job candidates visualize the scene. Following the pilot test, AIR decided not to include the image in the administration process because BDO incumbents did not need the visual cue and found the image distracting and/or confusing. Though internal TSO job candidates would also not likely require the visual cue provided by the image, external job candidates may benefit from its provision.

Selection of BDOs for the SPOT Assessor Program

As previously mentioned, the current validation study was conducted on the basis of a job analysis conducted in FY2010. Since that time, some changes in the BDO job have arisen. For example, the SPOT Assessor Program was recently pilot tested at Logan Airport in Boston, Massachusetts and the Detroit Metropolitan Wayne County Airport in Detroit, Michigan. Information AIR has gathered about this program reveals some changes to the job as compared to how it was previously conducted (e.g., BDO is stationed at the Ticket Document Checker [TDC] rather than —walking the line‖).

At this stage, it is unclear whether changes in the design of the job have resulted in new or different work behaviors (e.g., types of interactions with passengers). Future research could be conducted to examine whether the validity evidence gathered in the current validation study can be generalized to the newly developed variations of the BDO job, such as the SPOT Assessor Program. Specifically, a test transportability technique—a process of using job analytic

information to inform hiring decisions across similar jobs—could be used to determine the generalizability of the findings (Hoffman & McPhail, 1998). For example, if research suggests that the same major work behaviors apply to the BDO job as it was performed in FY2010 and variations that have since been developed, then the selection measures identified as valid measures for the BDO job in the current effort could be generalized to those variations. Test transportability also requires that the variations of the BDO job rely on similar pools of job candidates and the fairness of the selection measures are well documented (Gibson & Caplinger, 2007). Conducting this research and establishing the relationship between the BDO job and its variations would further increase the usefulness of the current validity research.

Training and Performance Measurement

This section presents areas for future consideration that relate to BDO training—both initial and recurrent—and performance measurement.

Implementation of a Training Alignment Study for Initial BDO Training

Throughout the job analysis (AIR, 2010a) and validation study, BDO incumbents commented that the initial training is invaluable to their success as a BDO. Nevertheless, many BDOs expressed a desire for the enhancement of the initial BDO training. Given this feedback, data collected during the BDO job analysis (AIR, 2010a) could be used to conduct a training alignment study to review the initial BDO training curriculum. Training alignment studies (also known as curriculum assessments, training assessments, and training gap analyses) enable an organization to determine the match between training needs and existing training programs. Such assessments allow for the identification of training shortfalls or —gapsl. Once gaps are identified, new training content could then be developed or existing content could be revised to ensure alignment with the current requirements of the job. In addition, evaluations of the training program can then be performed to determine whether or not trainees have acquired the critical KSAOs needed to perform the BDO job as a result of participation in the training program.

Should a BDO training alignment study be conducted, important questions to be answered might include:

- Is training focused on the right level of skills and knowledge? Are there any training gaps?;
- Should the initial BDO training be enhanced to emphasize casual conversation skills? If so, how?;
- How well do airports implement on-the-job training? What factors improve or hinder implementation?;
- Is refresher training needed and, if so, in what topics? What mechanisms most efficiently and effectively deliver refresher training and maximize the transfer of what is learned in training to the job?;
- Are there airport security personnel training models from other countries that can be examined to identify best practices?;
- Are there unfulfilled training needs related to BDO supervisory roles? What types of training are needed?;
- How can the Online Learning Center be leveraged to make tools, work aids, and other information available to the workforce?;

- Would a formal mentoring program be useful in this environment? How should this program be structured?; and
- How well is the National Standardization Team program working? Are there program improvements that could be implemented?

Ultimately, the effectiveness of SPOT hinges on how well the BDOs perform their job. A training alignment study could help examine the comprehensiveness of BDO training and whether it leads to the appropriate on-the-job behaviors that result in effective job performance.

Development of Recurrent BDO Training

In terms of recurrent training and professional development, incumbents expressed the need for additional opportunities to maintain and improve their skills, (b)(3):49 U.S.C. § 114(r) (b)(3):49 U.S.C. § 114(r).

(b)(3):49 U.S.C. § 114(r)

In addition, measures that did not perform as well in terms of selection (and were thus not included in the recommended selection battery) could be used as training tools instead. For example, the Visual Recall Assessment could be used in recurrent training as a method for maintaining and/or sharpening BDOs' visual recall skills. As mentioned in Chapter V, if TSA uses the Visual Recall Assessment in future validation studies or as recurrent training, it may be helpful to include items that were removed from the final version of the assessment for the validation study analyses. Specifically, TSA could collect additional data on each item—including those identified as problematic—and determine which items function appropriately (e.g., in terms of item difficulty, distractor analyses) with a larger sample of test-takers. Likewise, the Writing Knowledge Assessment could be used to sharpen or maintain BDOs' writing skills.

Development of Additional BDO Performance Measures

In addition, SMEs indicated that similar assessments would be extremely helpful for both training and performance measurement purposes. Regarding performance measurement, managers conveyed the need for objective performance appraisal measures. Some suggested modifying the Work Sample Test and the Role-Play Exercise for use as on-going assessments of BDO performance. In addition, one manager suggested the development of a —livel Passenger Observation Assessment in which an individual (i.e., a Resource Person) could travel through the passenger screening line displaying particular SPOT behaviors, and BDOs could then be evaluated on how many behaviors they successfully identify and/or whether they accurately identify the individual as a referral.

In general, more objective performance appraisal data could provide important information regarding the SPOT workforce's capabilities and inform organizational training and strategic planning.

BDO Job in General

This section presents areas for future consideration that relate to the BDO job in general.

Examination of the BDO Career Path

During data collection, several questions and issues were posed with regard to the BDO career path. Some came from incumbents who were unclear of their own career path. For example, questions were posed about career progression, opportunities for specialization, and differences between SPOT pay bands. Future consideration could be given to examining the BDO career path and disseminating information to the BDO workforce regarding these areas of uncertainty.

Other suggestions came from managers interested in examining the career path of a BDO before they enter the BDO workforce. Specifically, managers were interested in learning about the positions that are typically held by individuals before entering the BDO job and whether and which KSAOs gained in previous positions best prepare them for the job. Given that this type of information could be helpful for gaining further understanding of KSAOs required of the BDO job, future consideration could be given to conducting a survey of BDO incumbents to inquire

about previous positions held, KSAOs gained as a result, and perceptions of whether KSAOs gained are beneficial (or harmful) to performing the BDO job. Such information would also be useful for informing the possible development of other types of predictor measures, such as biodata measures.

Consideration of Human Capital Processes Used for Contractor BDOs

Under current legislation, airports can apply to use private contractors for screening activities through TSA's Screening Partnership Program (TSA, 2012). Sixteen airports currently rely on private companies to provide their screening services, including SPOT operations. Despite their contractor status, the BDOs perform their duties in accordance with TSA requirements and oversight.

On the other hand, the human capital operations used to recruit, select, and train BDOs are left to the discretion of the private companies. Thus, prior to beginning work, contractor BDOs may go through different processes than do BDOs hired directly by TSA. A comparison between the recruitment, selection, and training processes used by private companies and TSA could prove informative. Such a comparison could identify best practices that may enhance the human capital operations at either TSA or privately-operated airports. For example, TSA may be able to offer ways to improve the efficiency of the hiring process at privately-operated airports. On the other hand, private contractors may be able to offer effective strategies for recruiting candidates from large pools of applicants.

Examination of Reciprocal Determinism and its Effect on Casual Conversation

As a result of AIR's experience with developing and administering the Role-Play Exercise, AIR identified a potential area for future consideration with regard to casual conversation. Specifically, research by Albert Bandura known as reciprocal determinism suggests that a person's behavior both influences and is influenced by his/her environment (Bandura, 1997). Given this phenomenon, future research may be conducted to examine the relationship between a BDO's behavior and a passenger's behavior (e.g., does a particular BDO behavior result in specific behavioral indicators being displayed by the passenger?). If such a relationship exists, it may be the case that a BDO's behavior has an effect on the final outcome resulting from a casual conversation. If research in this area is conducted, specific consideration could be given to examining or evaluating ways in which this relationship may vary across different subgroups (e.g., male/female BDOs interacting with male/female passengers or minority/non-minority BDOs interacting with minority/non-minority BDOs).

Examination of BDO Fatigue and its Effect on Vigilance

Data gathered during both the job analysis (AIR 2010a) and the validation study clearly indicates that the BDO job is cognitively complex. In particular, mental fatigue is one factor that may negatively affect the vigilance (i.e., the ability to maintain a high level of attention for extended periods of time) required to perform successfully on the job. There are conflicting views about the existence and impact of fatigue on BDO performance. In fact, a recent Government Accountability Office (GAO) report (GAO, 2010) outlined concerns about the vigilance issue. Here, AIR proposes a series of research activities that could help determine whether, how, and to what extent fatigue affects BDO job performance with regard to vigilance. Research ideas include:

- Explore the literature related to other security positions and/or other high-risk jobs to determine what type(s) of fatigue should be expected and/or investigated;
- Conduct follow-up interviews or a survey of BDOs to learn more about BDO fatigue (e.g., how does physical fatigue vary by the set-up of work area? How does mental fatigue vary by context, such as time of day, shift, or passenger flow?);
- Evaluate how long a BDO can perform his or her duties before significant performance decrements appear;
- Identify which factors (e.g., being seated versus standing, methods of rotating assignments during a shift) may moderate the effect of fatigue on performance; and
- Determine how what is learned about BDO fatigue could be integrated into recruitment and selection, staffing, and training processes.

In the current political climate, vigilance is often a key to providing national security including detecting terrorism or aggressive intentions from individuals, groups, or nations. BDOs serve in one of many TSA jobs that provide a layer of security within the nation's airports and other venues. A vigilance study will provide information needed to determine the best way to structure the BDO job and the work environment to maximize job performance. Ensuring BDOs are well-positioned for successful job performance will have an important indirect effect on the safety of the traveling public.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Institutes for Research. (2009). *Job analysis update: Training decision paper*. Washington, DC: Author.

American Institutes for Research. (2010a). *Updating the personnel selection system for behavior detection officers*: *Job/task analysis report* (Vol. 1). Washington, DC: Author.

American Institutes for Research. (2010b). *Updating the personnel selection system for behavior detection officers*: *Task 2 deliverable - Test specifications*. Washington, DC: Author.

American Institutes for Research. (2012). *Updating the personnel selection system for behavior detection officers*: *Scoring procedures technical report*. Washington, DC: Author.

Anastasi, A. (1988). *Psychological testing*. New York, NY: MacMillan Publishing Company.

Anderson, N., & Witvliet, C. (2008). Fairness reactions to personnel selection methods: An international comparison between the Netherlands, the United States, France, Spain, Portugal, and Singapore. *International Journal of Selection and Assessment*, *16*, 1-13.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (ed.), *Educational measurement* (2nd ed.) (pp. 508-600). Washington, DC: American Council on Education.

Arvey, R., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, *43*, 695-716.

Aviation and Transportation Security Act. (2001). Pub L. No. 107-71, § 101, 115 Stat. 597.

Ayers, J. D. (1953). A simple technique for estimating optimum time limits. *Journal of Canadian Psychology*, *7*, 81-85.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: Freeman.

Bass, B. M., & Barrett, G. V. (1972). *Man, work and organizations: An introduction to industrial organizational psychology*. Oxford, England: Allyn & Bacon, Inc.

Bates, M. E., & Lemay, E. P. (2004). The d2 test of attention: Construct validity and existing scoring techniques. *Journal of the International Neuropsychological Society*, *10*, 392400.

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology*, *54*, 387-419.

Becker, T. E., & Klimonski, R. J. (1989). A field study of the relationship between the organizational feedback environment and performance. *Personnel Psychology, 42*, 343358.

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work.* Boston: Kent Publishing Company.

Bernardin, H. J., & Buckley, M. R. (1981), Strategies in rater training. *Academy of Management Review,* 6, 205-212.

Bobko, P., Roth, P. L., & Buster, M. A. (2005). Work sample selection tests and expected reduction in adverse impact: A cautionary note. *International Journal of Selection and Assessment, 13*(1), 1-10.

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48,* 587-605.

Borman, W. C., Bryant, R. H., & Dorio, J. (2010). The measurement of task performance as criteria in selection research. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection*. New York: Routelage Academic.

Borman, W.C., & Motowidlo, S.J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10*, 99-109.

Brannick, M. T., & Levine, E. L. (2002). *Job analysis: Methods, research, and applications for human resource management in the new millennium*. Thousand Oaks, CA: Sage Publications.

Sensitive Security Information (SSI)

Cardy, R. L., & Dobbins, G. H. (1994). *Performance appraisal: Alternative perspectives*. SouthWestern Publishing Co.: Cincinnati, Ohio.

Carretta, T. R., & Ree, M. J. (2000). General and specific cognitive and psychomotor abilities in personnel selection: The prediction of training and job performance. *International Journal of Selection and Assessment*, *8*, 227-236.

Cascio, W. F. (1998). *Applied psychology in human resource management* (5th Ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Cascio, W. F., Alexander, R. A., & Barrett, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional issues and guidelines. *Personnel Psychology*, *41*, 1-24.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.

Costigan, T. E., Makonnen, Z. E., Taylor, T. S., Sawyer, K., Myers, T., & Toplitz, M. (2011). *SPOT referral report validation study final report: Volumes 1-4.* Washington, DC: American Institutes for Research.

Cullen, M. J, & Sackett, P. R. (2004). Integrity testing in the workplace. In Thomas, J. (Eds.), *Comprehensive handbook of psychological assessment, Vol. 4: Industrial and organizational assessment* (p. 149-165). Hoboken, NJ: John Wiley & Sons Inc.

Department of Labor. (2000). *Testing and assessment: An employer's guide to good practices*. Washington, DC: Author.

Dipboye, R. L., Smith, C. S., & Howell, W. C. (1994). *Understanding industrial and organizational psychology: An integrated approach.* Orlando, FL: Harcourt Brace College Publishers.

Drasgow, F., Nye, C. D., & Tay, L. (2010). Indicators of quality assessment. In Scott, J. C., & Reynold, D. H. (Eds.), *Handbook of workplace assessment*, (pp. 671-710). San Francisco, CA: Jossey-Bass.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). *Uniform guidelines on employee selection procedures*. Federal Register, 43.

Gatewood, R. D., & Feild, H. S. (2001). *Human resource selection*. Fort Worth, TX: Harcourt College Publishers.

Gatewood, R. D., Feild, H. S., & Barrick, M. (2008). *Human resource selection*. Mason, OH: South-Western.

Giambra, L. M., & Quilter, R. E. (1989). Sex differences in sustained attention across the adult life span. *Journal of Applied Psychology*, *74*, 91-95.

Gibson, W. M., & Caplinger, J. A. (2007). Transportation of validation results. In McPhail, S. M. (Ed.) *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 29-81). San Francisco, CA: Jossey-Bass.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review*, *18*, 694-734.

Gottfredson, L. (2002). Where and why *g* matters: Not a mystery. *Human Performance*, *15*, 2546.

Government Accountability Office. (May, 2010). *Aviation security—efforts to validate TSA's passenger screening detection program underway, but opportunities exist to strengthen validation and address operational challenges* (GAO-10-763). Washington, DC: Government Printing Office.

Guion, R. M. (1965). *Personnel testing*. New York, NY: McGraw-Hill.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum Associates.

Foldes, H. J., Duehr, E. E., & Ones, D. (2008). Group differences in personality: Meta-analysis comparing five U.S. racial groups. *Personnel Psychology*, *61*, 579-616.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Lawrence Erlbaum Associates: Mahwah, New Jersey.

Harvey, R. J. (1991). Job analysis. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol. 2* (2nd ed., pp. 71-163). Palo Alto, CA: Consulting Psychologists Press.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*, *57*, 639-683.

Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology*,*87*, 243-254.

Hendrickson, C., Matheson, N., Amodeo, A. M., Norris, D., & Sparano, E. (2008). *Practice analysis for open source professionals*. Washington, DC: American Institutes for Research.

Hoffman, C. C., & McPhail, S. M. (1998). Exploring options for supporting test use in situations precluding local validation. *Personnel Psychology*, *51*, 987-1003.

Hollowitz, J. (1999). Investigations of a structured interview for pre-employment integrity screening (Doctoral dissertation, University of Nebraska, 1998). *Dissertation Abstracts International*, *60*, 398B.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*, 152-194.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and metaanalytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*, 897-913.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restrictions for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*, 594-612.

Jeanneret & Associates, & AIR. (2000, December). *United States Marshals Service: Development of a selection process for deputy U.S. Marshals in the GS-082 classification*. Houston, TX: Jeanneret & Associates.

Johnson, J. W., & Oswald, F. L. (2010). Test administration and the use of test scores. In Farr, J., & Tippins, N. T. (Eds.), *Handbook of employee selection* (pp. 151-170). New York, NY: Rutledge.

Jones, J. W., & Terris, W. (1991). Selection alternative to the preemployment polygraph. In J. Jones (Ed.), *Preemployment honesty testing: Current research and future directions* (3952). New York, Quorum Books.

Kehoe, J. F., & Olson, A. (2005). Cut scores and employment discrimination litigation. In Landy, F. J. (Ed.), *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives* (pp. 410-449). San Francisco, CA: Jossey-Bass.

Klehe, U., Konig, C. J., Richter, G. M., Kleinmann, M., & Melchers, K. G. (2008). Transparency in structured interviews: Consequences for construct and criterion-related validity. *Human Performance*, *21*, 107-137.

Knippen, J. T., & Green, T. B. (1994). How the manager can use active listening. *Public Personnel Management*, *23*, 357-359.

Kolen, M. J., & Brennan, R. L. (2004). *Statistics for social science and public policy: Test equating, scaling, and linking* (2nd edition). New York, NY: Springer.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Landon, T. E., & Arvey, R. D. (2007). Practical construct validation for personnel selection. In McPhail, S. M. (Ed.) *Alternative validation strategies: Developing new and leveraging existing validity evidence*. San Francisco, CA: Jossey-Bass.

Landy, F. J. (1989). *Psychology of work behavior* (4th edition). Pacific Grove, CA: Brooks-Cole.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72-107.

Lievens, F. (1999). Development of a simulated assessment center. *European Journal of Psychological Assessment*, *15*, 117-126.

Lievens, F., & Anseel, F. (2007). Creating alternate in-basket forms through cloning: Some preliminary trends. *International Journal of Selection and Assessment*, *15*, 428-433.

Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment*, *10*, 245-257.

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high fidelity simulations for predicting job performance in advanced high-stakes selection. *Journal of Applied Psychology*, *96*, 927-940.

Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology*, *47*, 715-738.

Mansilla, V. B., Duraisingh, E. D., Wolfe, C. R., & Haynes, C. (2009). Targeted assessment rubric: An empirically grounded rubric for interdisciplinary writing. *The Journal of Higher Education*, *80*, 334-353.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, *79*, 599-616.

McPhail, S. M., & Stelly, D. J. (2010). Validation strategies. In Scott, J. C., & Reynold, D. H. (Eds.), *Handbook of workplace assessment*, (pp. 671-710). San Francisco, CA: JosseyBass.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (pp. 13-103). New York, NY, Macmillian.

Moregeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*, 683-679.

Mueller, L., Norris, D., & Oppler, S. (2007). Implementation based on alternative validation procedures: Ranking, cut scoring, banding, and compensatory models. In McPhail, S. M. (Ed.) *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 349-405). San Francisco, CA: Jossey-Bass.

Mullaney, H. A., Makonnen, Z. E., & Costigan, T. E. (2009). *Casual conversation: Related literature findings and training recommendations.* Submitted to the Department of Homeland Security, Washington, DC: American Institutes for Research.

Neuman, G. A., & Wright, J. (1999). Team effectiveness: Beyond skills and cognitive ability. *Journal of Applied Psychology*, *84*, 376-389.

Nunally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGrawHill.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. (2007). In support of personality assessment in organizational settings. *Personality Psychology*, *60*, 995-1027.

Outtz, J. (2002). The role of cognitive ability tests in employment selection. *Human Performance*, *15*, 161-171.

Outtz, J. L. (2010). Addressing the flaws in our assessment decisions. In Scott, J. C., & Reynold, D. H. (Eds.), *Handbook of workplace assessment*, (pp. 711-727). San Francisco, CA: Jossey-Bass.

Pearson, Inc. (2006). *Watson-Glaser Critical Thinking Appraisal Short Form manual*. San Antonio, TX: Author.

Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: Contemporary practice and theory (3rd ed.).* Mahwah, NJ: Lawrence Erlbaum Associates.

Premack, S. L., & Wanous, J. P. (1985). A meta-analysis of realistic job preview experiments. *Journal of Applied Psychology*, 70, 706-719.

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

Pursell, E. D., Campion, M. A., & Gaylord, S. R. (1980). Structured interviewing: Avoiding selection problems. *Personnel Journal*, *59*, 907-912.

Putka, D. J., & Sackett, P.R. (2010).Reliability and validity. In Farr, J., & Tippins, N. T. (Eds.), *Handbook of employee selection* (pp. 9-50). New York, NY: Rutledge.

Rawson, R. E., Quinlan, K. M., Cooper, B. J., Fewtrell, C., & Matlow, J. R. (2005). Writingskills development in the health professions. *Teaching and Learning in Medicine*, *17*, 233-238.

Roth, P. L., Bevier, C. A., & Bobko, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*, 297330.

Roth, P., Bobko, P., McFarland, L., & Buster, M. (2008). Work sample tests in personnel selection: A meta-analysis of black-white differences in overall and exercise scores. *Personnel Psychology*, *61*, 637-662.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review*, *16*, 155-180.

Russell, T.L., Norris, D.G., & Goodwin, G.F. (2000, June). *Literature review: Alternative methods of measuring CPA knowledge and skill*. Washington, DC: American Institutes for Research.

Sensitive Security Information (SSI)

Russel, T. L., & Peterson, N. G. (1997). The test plan. In Whetzel, D. L., & Wheaton, G. R. (Eds.), *Applied measurement methods in industrial psychology*. Palo Alto, CA: DaviesBlack.

Ryan, J. J., & Paolo, A. M. (2001). Exploratory factor analysis of WAIS-III in a mixed patient sample. *Archives of Clinical Neuropsychology*, *16*, 151-156.

Sackett, P. R., & Ostgaard, D. J. (1994). Job-specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *Journal of Applied Psychology*, *79*, 680-684.

Sackett, P. R., & Wanek, J. E. (1996). New developments in the use of measures of honesty, integrity, conscientiousness, dependability, trustworthiness, and reliability for personnel selection. *Personnel Psychology*, *49*, 787-829.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*, 112-118.

Sanchez, J. I., & Levine, E. L. (2009). What is (or should be) the difference between competency modeling and traditional job analysis? *Human Resource Management Review*, *19*, 53-63.

Sands, W. A., Alf, E. F., & Abrahams, N. M. (1978). Correction of validity coefficients for direct restriction in range occasioned by univariate selection. *Journal of Applied Psychology*, *63*, 747-750.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262-274.

Schmitt, N. (2007). The value of personnel selection: Reflections on some remarkable claims. *Academy of Management Perspectives*, *21*, 19-23.

Schmitt, N., & Mills, A. E. (2001). Traditional tests and job simulations: Minority and majority performance and test validities. *Journal of Applied Psychology, 86*(3), 451-458.

Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., & Ramsay, L. J. (2004). The impact of justice and self-serving bias explanations of the perceived fairness of different types of selection tests. *International Journal of Selection and Assessment*, *12*, 160-171.

Schwab, D. P., Heneman, H. G., & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology, 28,* 549-562.

Siddique, C. M. (2004). Job analysis: A strategic human resource management practice. *International Journal of Human Resource Management*, *15*, 219-244.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149-155.

Stelly, D. J., & Goldstein, H. W. (2007). Application of content validation methods to broader constructs. In McPhail, S. M. (Ed.) *Alternative validation strategies: Developing new and leveraging existing validity evidence*. San Francisco, CA: Jossey-Bass.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH.

Thompson, D. E., & Thompson, T. A. (1982). Court standards for job analysis in test validation. *Personnel Psychology, 35*(4), 865-874.

Thornton, G. C., & Mueller-Hanson, R. A. (2004). *Developing organizational simulations: A guide for practitioners and students*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Tippins, N. T. (2010). Adverse impact in employee selection procedures from the perspective of an organizational consultant. In Outtz, J. L. (Ed.) *Adverse impact: Implications for organizational staffing and high stakes selection*. San Francisco, CA: Jossey-Bass.

Transportation Security Administration. (2006). *Administration guide for Behavior Detection Officer (BDO) and Transportation Security Manager-BDO (TSM-BDO) selection* (Version 6). Washington, DC: Author.

Transportation Security Administration. (2010). *Performance accountability & standards system: User's guide and metrics*. Washington, DC: Author.

Transportation Security Administration. (2012). Screening partnership program: Merging private-sector expertise and public-sector know how. Retrieved from http://www.tsa.gov/what_we_do/optout/index.shtm.

Truxillo, D. M., Donahue, L. M., & Kuang, D. (2004). Work samples, performance tests, and competency testing. In Thomas, J. C., & Hersen, M. (Ed.), *Comprehensive handbook of psychological assessment industrial and organizational assessment*, (345-370). Hoboken, NJ: Wiley, Inc.

Truxillo, D. M., Donahue, L. M., & Sulzer, J. L. (1996). Setting cutoff scores for personnel selection tests: Issues, illustrations, and recommendations. *Human Performance*, *9*, 275295.

Truxillo, D. M., Steiner, D. D., & Gilliland, S. W. (2004). The importance of organizational justice in personnel selection: Defining when selection fairness really matters. *International Journal of Selection and Assessment*, *12*, 39-53.

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection measures: A critical review and recommendations for practice. *Personnel Psychology*, *61*, 871-925.

Wainer, H., & Lewis, C. (1989). *Toward a psychometrics for testlets*. Princeton, NJ: Educational Testing Services.

Watson-Glaser Critical Thinking Appraisal. (1998). In *The thirteenth mental measurements yearbook*. Available from http://buros.unl.edu/buros/.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Wonderlic, Inc. (2002). *Wonderlic Personnel Test and Scholastic Level Exam User's Manual*. Libertyville, IL: Author.

Wonderlic Personnel Test. (2001). In *The fourteenth mental measurements yearbook*. Available from http://buros.unl.edu/buros/.