AFRL-RH-WP-TR-2021-0028

ENHANCED AIRMAN ALIGNMENT: 2018-2021 PROJECT SUMMARY

Thomas R. Carretta
711 HPW/RHBC

Annette L. Rizer
Michael F. Brady
Infoscitex Corporation

April 2021
Final Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2021-0028 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//

THOMAS R. CARRETTA, PhD
Work Unit Manager
Performance Optimization Branch
Airman Biosciences Division

//signature//

R. ANDY MCKINLEY, DR-III, PhD
Core Research Area Lead
Performance Optimization Branch
Airman Biosciences Division

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
30-04-21	Final	December 2017 – February 2021

4. TITLE AND SUBTITLE

Enhanced Airman Alignment: 2018-2020 Project Summary

5a. CONTRACT NUMBER
FA8650-14-D-6500/FA8650-18-F-6825

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Thomas R. Carretta1, Annette L. Rizer2, and Michael F. Brady3

5d. PROJECT NUMBER
5329

5e. TASK NUMBER
09

5f. WORK UNIT NUMBER
H0SA (532909TC)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
2Infoscitex Inc., A DCS Corp
4027 Col Glenn Hwy, Suite 210
Dayton OH 45431

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Airman Biosciences Division
Performance Optimization Branch
Wright-Patterson AFB, OH 45433

Air Force Personnel Center
Strategic Research and Analysis Branch
550 C St West, Ste. 45
JBSA-Randolph, TX 78150-4747

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHBC

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2021-0028

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release.

13. SUPPLEMENTARY NOTES
AFRL-2021-1764, cleared 8 June 2021

14. ABSTRACT

In 2017, the US Congress set aside funding for a five-year Testing Modernization program to support advanced development of military enlisted personnel assessment methods and information technology (IT) infrastructure to enable the optimization of human capital. This report summarizes 19 US Air Force projects performed between December 2017 and February 2021. These project objectives fell into five broad areas. These were the development/validation of: 1) advanced cognitive assessment methods, 2) advanced non-cognitive assessment methods, 3) occupational performance criteria, 4) algorithms to optimize the use of human capital, and 5) improved IT infrastructure to enable remote special testing and efficient data processing and management.

15. SUBJECT TERMS
Personnel measurement, selection and classification; criterion development; remote special testing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Thomas R. Carretta
Unclassified	Unclassified	Unclassified	SAR	51	19b. TELEPHONE NUMBER *(Include Area Code)* N/A

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18

TABLE OF CONTENTS

LIST OF TABLES

PREFACE

The nineteen (19) projects summarized in this report are from the Enhanced Airman Alignment program from December 2017 through February 2021. Contributors included personnel from the prime contractor, Infoscitex (IST), and nine (9) subcontractors. The companies and their respective projects were:

1. BAM Technologies – Remote Special Testing Data Management System: Phase I - Planning and Proof-of-Concept Developmental/Operational Test & Evaluation
2. Campion Consulting – Examination of the Utility of Text Analysis Tools to Enhance Personnel Assessment
3. Drasgow Consulting Group (DCG) – Tailored Adaptive Personality Assessment System (TAPAS) Item Development and Examination of Alternate Item Formats; Dark Tetrad Item Development
4. Human Resources Research Organization (HumRRO) – Cyber Test Update; Evaluation of the Impact of Removing Armed Services Vocational Aptitude Battery (ASVAB) Technical Knowledge Tests; Air Force Compatibility Assessment (AFCA) Review; Department of Defense (DoD) Job Performance Criterion Development; Weighted Airman Promotion System (WAPS) Evaluation; Advanced Accessions
5. Infoscitex (IST) – Aptitude Requirements for Operators in Jobs with High Levels of Human-Automation Interaction
6. MDC & Associates (MDC) – Use of Technology-Enhanced Work Simulations (Serious Games) for Cyber Assessment, Phase I and II
7. National Academy of Sciences (NAS) – Strengthening U.S. Air Force Human Capital Management
8. Operational Technologies Corporation (OpTech) – Air Force Reading Abilities Test (AFRAT) Revision; Examination of the Utility of Months of Mission-Ready Service (MMRS) as a Personnel Selection and Classification Criterion
9. Osi Vision – Strategic Research and Assessment Program Data System Support: Phase I
10. PDRI - Navy Promotion Testing Evaluation; TAPAS, Navy Computerized Adaptive Personality Scales (NCAPS), and Self-Description Inventory (SDI) Comparison and Consolidation; Best Practices Guide for Assessment, Development, Validation, and Implementation

1.0 INTRODUCTION

In 2017, the US Congress set aside funding for a five-year Testing Modernization program to support advanced development of military enlisted personnel assessment methods and information technology (IT) infrastructure to enable the optimization of human capital. US Air Force project objectives fell into five broad areas. These were the development/validation of: 1) advanced cognitive assessment methods, 2) advanced non-cognitive assessment methods, 3) occupational performance criteria, 4) algorithms to optimize the use of human capital, and 5) improved IT infrastructure to enable remote special testing and efficient data processing/management. The technical approach involved: 1) identification of critical entry-level skills, abilities, and other characteristics (SAOCs), 2) identification of measurement gaps and development of new measures as needed, 3) data collection and psychometric analyses, and 4) development and evaluation of information technology to improve the efficiency of data collection and processing, enabling timely decision-making. Brief project summaries for the period January 2018 through February 2021 are provided below, grouped by broad project objective area.

2.0 DEVELOPMENT AND VALIDATION OF ADVANCED ASSESSMENT METHODS: COGNITIVE ABILITY

Several projects focused on the improved assessment of cognitive ability. These were:

1) an update to the content of the knowledge-based Cyber Test and development of a computer-adaptive test (CAT),
2) development of a game-based approach to assessment of cyber aptitude,
3) development of a new version of the Air Force Reading Abilities Test (AFRAT),
4) examination of aptitude requirements for personnel in jobs requiring high levels of human-automation interaction,
5) examination of the utility of text analysis tools to enhance personnel assessment, and
6) evaluation of the impact of removing one or more of the ASVAB technical knowledge tests on enlistment qualification, predictive validity, and classification efficiency.

2.1 Cyber Test Update

Since 2008, several studies have been conducted to develop and evaluate the Cyber Test (formerly known as the Information and Communications Technology Literacy (ICTL) Test, or. The Cyber Test is a knowledge-based measure used as a pre-enlistment assessment across the US Military Services, It has been shown to predict success in entry-level training in cyber-related military occupations (Trippe, Moriarty. Russell, Carretta, & Beatty, 2014). The goal of the current project was to transition the existing static Cyber Test forms to a CAT platform. Toward this end, 251 experimental items were pilot tested, calibrated, equated, screened, and added to the existing Cyber Test item pool. The resuilting items were assembled into three parallel forms (or item pools) from which the CAT algorithm will draw items. Additionally, the test blueprint was updated with the assistance of cyber subject matter experts across the Services, and 215 new items were developed and are ready to be pilot tested (Koch, Trippe, Beatty, & Shewach, 2019).

2.2 Use of Technology-Enhanced Work Simulations (Serious Games) for Cyber Assessment

2.2.1. Cyber Game Development Phase I. Numerous assessments have been developed or proposed for classification of US Air Force recruits into cyber careers. Most of these involve measurement of cognitive abilities, technical knowledge, and/or personality traits that have been the focus of extensive research across a broad range of career fields, and that employ common testing methods (e.g., paper-and-pencil or computer-based self-reports) and approaches to validation. One exception is the proposed use of serious games in aptitude assessment. The Phase I effort examined potential gains in predictive validity that could be achieved beyond traditional methods through the use of technology-enhanced work simulations (serious games) as part of an assessment for selection and classification into cyber career fields.

The first objective was to identify aptitudes and traits required for success in select Air Force enlisted and officer cyber careers using archival information. The second objective was to identify which cyber aptitudes and traits can be measured through existing DoD tests and where measurement gaps existed. The third objective was to provide a summary of relevant literature and recommendations for how a serious game approach could be used to measure cyber aptitudes and traits, including those where gaps in assessment currently exist. Four aptitudes and five traits were identified as the individual characteristics most related to on-the-job cyber performance for which there is currently inadequate assessment and which could be assessed via a serious game (Coovert, Martin, Howard, Kim, Dreibelbis, Arbogast, & Potter, 2019). Table 1 provides brief definitions of these aptitudes and traits.

2.2.2. Cyber Game Development Phase II. Cyber Game Development Phase II (Coovert, Weirnik, & Martin, 2020) extended previous work (Coovert et al., 2019) with the primary objective of developing a serious game to assess the suitability of military candidates for entry-level cyber career fields. For the purpose of this project, cyber career fields refer to the following Air Force Specialty Codes (AFSCs) - enlisted: 3D0X2 (Cyber Systems Operations), 1B4X1 (Cyber Warfare Operations), 1N4X1A (Digital Network Analyst), and 3D1X2 (Cyber Transport Systems); officer: 17DX/SX (Network Operations/Cyber Space Operations).

Table 1. Aptitudes and Traits Important for United State Air Force (USAF) Cyber Jobs Not Adequately Measured by Current USAF/DoD Tests

Aptitude/Trait	Type	Definition
Active Learning	Aptitude	The ability to understand the implications of new information for both current and future problem-solving and decision-making
Adaptability	Trait	The degree to which individuals are open to change (positive or negative) and to considerable variety in the workplace
Analytical Thinking	Trait	The degree to which individuals analyze information and use logic to address work-related issues and problems
Decision-Making	Aptitude	The ability to consider the relative costs and benefits of potential actions to choose the most appropriate alternative
Deductive Reasoning	Aptitude	The ability to apply general rules to specific problems to produce answers that make sense
Dependability	Trait	The degree to which individuals are responsible, reliable, and dependable, and fulfill obligations
Persistence	Trait	The degree to which individuals persist in the face of obstacles
Situational Awareness	Trait	The degree to which individuals pay attention to their surroundings and rarely get lost or surprised
Systems Thinking	Aptitude	The ability to understand how multiple parts of a system interact and influence each other

Using the results from the subject matter expert (SME) data collected in Phase I, a list of critical aptitudes and traits were refined to include six (analytical thinking, deductive reasoning, systems thinking, active learning, adaptability, and situational awareness) which become the focus and those to be assessed in a serious game. These aptitudes and traits are expected to have a strong potential for demonstrating significant improvement in predictive validity/classification efficiency when used with existing DoD tests.

With the six aptitudes and traits identified, work began on the development of a serious game to assess them in individuals. Several ideas were developed and, after input from a focus group, the development team settled on a serious game to manage a virus epidemic[1]. Game design began and, along with extensive pilot work, a spiral development approach was taken to flesh out the overarching theme, sub-themes, and vignettes. Measurement of analytical thinking and active

[1] This decision was made, and game development began approximately one year prior to the identification of COVID-19 in China and the resulting world-wide pandemic.

learning occur in a mini-game set in a selection context. A second mini-game, whereby the candidate repairs a damaged circuit board, measures deductive reasoning. The final three constructs (systems thinking, adaptability, and situational awareness) are measured throughout active game play where treatment centers are monitored, vehicles loaded and dispatched, and adjustments made to demands made by changing environmental conditions.

Item response theory (IRT) and factor analysis methods were used to examine the construct validity of the indicators used to assess the six constructs. IRT analyses indicated good marginal reliabilities for all constructs (Analytical Thinking, Active Learning, Deductive Reasoning, Situational Awareness, Adaptability, and Systems Thinking). Situational Awareness, Adaptability, and Systems Thinking each had indicator variables developed from game play. The IRT and factor analyses indicated reasonable construct validity and discrimination for the items on those constructs.

In addition to face validity, results of the factor analysis and IRT analyses were used to select the final items from large initial pools. The sample of individuals involved in pilot work closely mirrored the population of those new or early in their Air Force career. The final items demonstrated good reliability and discriminability.

Throughout game development, the usability of each module was assessed, and adjustments were made when necessary. Two larger usability assessments were made as well, one following the development of the selection mini-game and a second of the full game. One must keep in mind this is a serious game, developed as an assessment tool. It is not a typical online game developed primarily for entertainment. Thus, the obtained ratings from players for training to play the game, the help feature, and overall aspects of the game are quite reasonable.

As originally envisioned, this project would also sample individuals in Basic Military Training (BMT) at Lackland AFB. Scores on the Virus Slayer serious game were to have been correlated with scores on the ASVAB and Air Force Tailored Adaptive Personality Assessment System (AF TAPAS). Testing with Air Force Basic Recruits and psychometric evaluation was delayed due to social distancing constraints resulting from the COVID-19 pandemic. Once the testing at BMT has been accomplished, a study is planned to administer the Virtual Slayer serious game to cyber training students at Keesler AFB to examine its predictive validity/incremental validity against training performance.

2.3 Air Force Reading Abilities Test (AFRAT) Update

The AFRAT; Mathews & Roach, 1983) was developed to assess reading grade level (RGL) of enlisted personnel. It is used to identify personnel with poor reading ability and assign them to remedial reading training, as an aid in counseling students, and for assessing the RGL of personnel in various occupational specialties. Although the AFRAT has proven useful for many years, its content has not been updated since its implementation in 1982. The objectives of this

effort were to develop two new forms of the AFRAT and develop two new sections to identify service members with potential reading disabilities (Pena, Martinez, Haight, & Shore, 2021). The existing AFRAT Reading Comprehension subtest was updated based on the types of written information enlisted personnel are expected to understand. A section was added to help identify enlistees with potential reading disorders so that they can be directed to appropriate resources for help, if needed. A sufficient number of items was developed to create two new parallel forms.

The new AFRAT forms were equated to the previous form (81A). Examination of item statistics showed that the new forms were parallel to each other and to the previous version. The new forms had an average item difficulty of .80 and average biserial correlation of .56. The previous version had an average difficulty of .84 and a biserial correlation of .70. Equated scores were converted to a RGL scale for the total scores.

Analyses of the new subtests designed to identify reading disabilities (Orthographic and Phonological Choice) indicated that low scores on them are related to lower RGL and to increased risk of having a reading disability. Low percentile scores identify Airmen most likely to have some form of reading disability. Data collection from a larger Air Force sample, along with the accumulated knowledge in identification of reading disabilities, will allow for the determination of a standard score that accurately identifies reading deficiencies. Pena et al. (2021) presented ways in which these tests, along with the RGL measure, can identify Airmen who should receive additional assessment or training.

2.4 Aptitude Requirements for Operators in Jobs with High Levels of Human-Automation Interaction

Over the past decade, there has been considerable interest in the US military in the development of systems with increased levels of automation and autonomy. Autonomous systems provide a considerable opportunity to enhance future Air Force operations by potentially reducing manning costs, increasing the range of operations, enhancing capabilities, providing new approaches to air power, reducing the time required for critical operations, and providing increased levels of operational reliability, persistence and resilience. Increased levels of autonomy can be brought to bear to enhance operations in both manned and unmanned aircraft and in operations in space, cyber, command and control, intelligence, surveillance and reconnaissance (ISR), readiness, and sustainment across the Air Force. The vision is for the development of highly autonomous/ automated systems that will work synergistically with human operators as part of an effective human-automation or human-machine team. The introduction of increased automation and autonomy provides several human-systems integration challenges.

2.4.1. Human-Automation Interaction Phase I. The first objective of this effort was to identify factors that affect operator aptitude requirements in jobs requiring high levels of human-

automation interaction (HAI) or human-machine teaming (HMT) (i.e., peer-to-peer). The second objective was to identify entry-level aptitude requirements (i.e., SAOCs) for these types of jobs.

A literature review (Brady, Shermadou, Gibson, & Carretta, 2020) identified several task characteristics (i.e., monitoring, multitasking, and effective human-machine interaction and operator states (i.e., boredom, complacency, perceived workload, situation awareness, stress, and trust) relevant to performance in HAI. The literature review also indicated that several cognitive abilities (i.e., attentional control, fluid intelligence, spatial ability, and working memory), personality traits (i.e., Big Five), and other characteristics (e.g., boredom proneness, coping style, perfect automation schema, and video gamer experience) have been studied in relation to HAI and HMT performance.

Results of meta-analysis indicated that cognitive ability had the strongest relationship with HAI performance - spatial ability (r = .52), working memory (r = .43), and attentional control (r = .33). Insufficient studies were found to examine the predictive validity of general mental ability, crystalized intelligence, or fluid intelligence. The strongest relationships for personality occurred for conscientiousness (r = .26), neuroticism (r = -.17), and agreeableness (r = .15). Analyses for other characteristics indicated video game experience (r = .30) and boredom proneness (r = .31) were related to HAI performance.

2.4.2. Human-Automation Interaction Phase II. This study examined the relations between identified critical entry-level SAOCs and performance on tasks requiring high levels of HAI (Brady, Shermadou, & Carretta, 2020). Criteria data were collected using the Adaptive Levels of Autonomy (ALOA) testbed, developed by OR Concepts Applied (ORCA; Johnson, Leen, & Goldberg, 2007; Johnson, Leen, Goldberg, & Chiu, 2005). ALOA realistically simulates complex tasks represented in UAV operations and incorporates different levels of automation that can be set by the researcher or adjusted by operators.

Correlations were computed to identify relationships among stable individual differences measured pre-task, performance metrics using the ALOA simulator, and operator states measured post-task. Results showed several statistically significant correlations between cognitive measures and objective performance. Personality was related to some distal performance outcomes, but was mostly relevant to how well participants utilized the automation's suggestions and detected errors.

Hierarchical regression analyses were performed to examine the incremental validity of individual differences variables beyond cognitive ability and personality. Results varied by criteria. For Image Analysis Accuracy, perceptual speed remained statistically significant as variables were added to the model. Conscientiousness and need for cognition each significantly improved multiple R-squared (R^2) beyond perceptual speed. Adding Perfect Automation Schema (PAS) did not significantly improve R^2.

The hierarchical regression analyses for Weapon Release Accuracy included an additional step not included in the image analysis task, since extraversion had a significant relationship with weapon release task accuracy. Extraversion was included in a separate step from conscientiousness in order to parse out their respective contributions. While conscientiousness did not significantly improve R^2, adding extraversion to the model did. PAS significantly improved R^2, as did need for cognition.

Because the hierarchical regression analyses were similar for accuracy on both the image analysis and weapon release tasks, he accuracy for the two tasks were combined into a composite accuracy variable in order to determine the contribution of these variables to overall accuracy. Adding conscientiousness to perceptual speed significantly improved R^2. Extraversion did not significantly improve R^2.Adding PAS significantly improved R^2, as did need for cognition.

For the Health Monitoring Reaction Time criterion, the beta coefficient for perceptual speed was significant in every step of the analyses. Perfect automation schema had incremental validity above perceptual speed and conscientiousness. However, when need for cognition was added (model 4), the beta coefficients for PAS were no longer significant.

The authors concluded by identifying opportunities to improve the measurement of promising constructs and to identify research gaps.

2.5 Examination of the Utility of Text Analysis Tools to Enhance Personnel Assessment

Text analysis methods have been applied in several areas including social networking and blogging, risk management, knowledge management, cybercrime prevention, customer care service, customer relations, text filtering, and tracking public opinion. Text analysis tools also have been applied to personnel assessment (e.g., Campion, Campion, Campion, & Reider, 2016). Campion et al. (2016) suggested that text analysis could be used in lieu of human raters in a personnel selection application. Several studies have examined the utility of text analysis for assessing cognitive ability and personality characteristics from written text. The objectives of this project were to 1) conduct a literature review of state-of-the-art text analysis software and analytical techniques and identify best practices, 2) identify where and how text analysis could augment existing US Air Force assessments, and 3) examine the utility of text analysis tools/methods using archival US Air Force data.

2.5.1. Literature Review of State-of-the Art Text Analysis Software and Analytical Techniques and Identify Best Practices.

The literature review identified 242 relevant articles regarding computer-assisted text analysis (CATA) (Campion & Campion, 2019). Three types of studies emerged from the literature review. The first group consisted of those that used qualitative CATA methods only (28.1%). The second group consisted of studies that used a combination of qualitative and quantitative CATA methods (62.4%). Group 3 consisted of review articles (9.5%) that introduced text mining, provided recommendations on how to

conduct CATA, presented specific techniques for conducting CATA, or summarized the use of CATA to measure constructs.

The literature review revealed that text analysis methods have been applied to virtually all types of textual data. Some examples are essays, transcripts of interviews or phone calls, news articles and press releases, survey responses, and tweets. The literature review also characterized studies in terms of whether a low or high human intervention and low or high level of computer automation was involved. Historically, mostly simpler and less sophisticated forms of CATA have been used, but there is a trend toward greater use of more sophisticated approaches.

Campion and Campion (2019) also discussed types of construct validation studies employed in text analysis studies (i.e., definition of content domain, internal structure, relationships with other measures, predictive validation, SME judgments). They also provided a review of types of software used in text analysis (e.g., Language Inquiry and Word Count [LIWC], Python, SPSS Modeler) and their intended uses and likely applicability to Air Force data sources.

2.5.2. Examination of the Utility of Text Analysis Methods using Archival US Air Force Data. Three studies were conducted (Campion & Campion, 2020). The first study illustrated the use of text analysis techniques to analyze the content of text responses to occupational surveys. The second study used text analysis to score responses to essay questions (i.e., structured interviews). The third study applied text analysis methods to score Officer Training School (OTS) applications and predict rated and non-rated OTS selection board scores.

2.5.2.1 Content analysis of archival occupational survey data. The Air Force Personnel Center had archival data from several occupational surveys of enlisted jobs (e.g., Security Forces, Remotely Piloted Aircraft (RPA) Sensor Operators). The surveys assessed factors affecting job satisfaction and identified both positive and negative aspects of the job. Human raters had previously reviewed responses to survey questions in order to develop content themes to create video-based realistic job previews (RJPs) to inform potential job applicants. The conclusion of the study was that text analysis methods did a reasonably good job of reproducing the categories identified by the human raters, and the differences in the categories were no less logical. In addition, a researcher using text analysis could conduct the analyses in a few hours that would take a week to do manually.

2.5.2.2 Scoring responses to essay questions (structured interviews). Using text analysis to score structured interview responses was based on four considerations: (1) interview responses are narrative, (b) interviews are time-consuming to administer, so text analysis may save time, (3) an automated structured interview might complement the current Air Force use of aptitude tests, and (4) interview questions might measure constructs that could complement the aptitude tests. Four hundred sixty-three (N = 463) Air Force Basic Recruits responded to six interview questions by typing their responses. The responses were scored by two Air Force

Personnel Center (AFPC) interns and by the text analysis software. The interns received training to improve their level of reliability. Results indicated that the text analysis software extracted a meaningful number of categories and the categories showed small positive relations with aptitude test scores (ASVAB and TAPAS). Results also indicated that text analysis can be used to score interview responses, the scores show good validity in predicting interviewer (human scorers) ratings, the scores show construct validity with other assessments, and use of text analysis would have practical utility for making selection decisions.

2.5.2.3 Scoring Officer Training School (OTS) applications. OTS applications are a potentially rich source of textual data for personnel assessment. Much of the information is narrative, such as past employment, educational degrees, statements, and letters of recommendation. Manual review of the applications to determine applicant suitability and make personnel selection decisions is labor intensive and time-consuming. Text analysis methods were applied to determine if they could improve the efficiency of application review and predict OTS rated and non-rated board scores for prior service and non-prior service applicants. Data were available for about 497 rated (flying training) and 1,331 non-rated (non-flying training) prior service and non-prior service applicants. Results indicated that numeric scores (e.g., AFOQT composites, GPA, number of jobs, and essay scores) were consistent predictors of OTS Board scores (*R* ranged from .48 to .76) and that text variables provided incremental validity (between 3.6% and 8% increase in R^2). Results indicated that text analysis techniques have the potential to improve the efficiency of scoring text material (i.e., reduce the time required to score text material by human raters) and to improve measurement of job-related constructs, increase predictive validity, and reduce adverse impact.

2.6 Evaluation of Impact of Removing Armed Services Vocational Aptitude Battery (ASVAB) Technical Knowledge Tests

The primary objective of this effort was to analyze archival US Air Force data to determine the impact of removing one or more ASVAB subtests on prediction of training performance, subgroup qualification rates (i.e., adverse impact), and classification efficiency (Johnson & Zeidner, 1995; Zeidner & Johnson, 1991, 1994). The ASVAB subtests are Arithmetic Reasoning (AR), Assembling Objects (AO), Auto and Shop Information (AS), Electronics Information (EI), General Science (GS), Math Knowledge (MK), Mechanical Comprehension (MC), Paragraph Comprehension (PC), and Word Knowledge (WK). Verbal Expression (VE) is a weighted composite of PC and WK. Analyses focused on the ASVAB technical knowledge subtests (General Science, Mechanical Comprehension, Electronics Information, and Auto/Shop). These subtests contribute to the USAF Mechanical and Electronics composites as shown below:

Mechanical (M) = AR + MC + AS + 2VE

Administrative (A) = MK + VE

General (G) = AR + VE

Electronics (E) = GS + AR + WK + EI

Armed Forces Qualification Test (AFQT) = AR + MK + 2VE

A secondary objective was to examine the utility of content changes to the USAF Administrative (A} and General (G) composites. Neither of these include the ASVAB technical knowledge tests. For the A composite, analyses focused on whether the addition of Assembling Objects (AO), AR, and/or GS improve its utility. For the G composite, analyses focused on whether the addition of AO, MK, and/or GS would improve its utility. The largest loss in predictive validity would occur for the Mechanical composite (.447 - .472 = -.025).

Table 2 summarizes results of the predictive validity analyses. Generally, a model resembling the AFQT containing both verbal and both math subtests showed the highest validity for the alternate composites for all four occupational groups.

Table 2. Aggregated Predictive Validity Results for the Current MAGE Composites and Alternatives

	Training Grades	
	Validity	**Delta**
Current Mechanical Composite	.472	
AR + PC + WK	.429	-.043
AR + PC + WK + MK	.447	-.025
AR + PC + WK + AO	.408	-.064
AR + PC + WK + AO + MK	.432	-.040
Current Electronics Composite	.472	
AR + MK	.401	-.071
AR+ MK + WK	.452	-.020
AR + MK + PC	.447	-.025
AR + MK + AO	.376	-.095
AR + MK + PC + WK	.467	-.005
AR + MK + AO + WK	.429	-.043
AR + MK + AO + PC	.427	-.044
AR + MK + AO + PC + WK	.452	-.020
Current Administrative Composite	.398	
MK + PC + WK + GS	.390	-.008
MK + PC + WK + AR	.415	.017
MK + PC + WK + AO	.375	-.023
MK + PC + WK + AR + GS	.409	.011
MK + PC + WK + AO + GS	.381	-.017
MK + PC + WK + AO + AR	.399	.001
MK + PC + WK + AO + AR + GS	.400	.002
Current General Composite	.429	
AR + PC + WK + MK	.452	.023
AR + PC + WK + AO	.407	-.022
AR + PC + WK + GS	.441	.012
AR + PC + WK + AO + MK	.434	.005
AR + PC + WK + GS + MK	.459	.030
AR + PC + WK + AO + GS	.427	-.002
AR + PC + WK + AO + GS + MK	.448	.019

Table 3 summarized the results of the analyses that examined qualification rates for the majority group versus the minority group. All of the alternative composites showed less adverse impact for the Mechanical and Electronics analyses. In contrast, addition of subtests to the Administrative and General composites tended to slightly increase adverse impact.

Results of the predictive validity (Table 2) and qualification rate (Table 3) analyses must be considered together to inform Air Force policy. Although predictive validity is lower for the alternative Mechanical and Electronics composites, they greatly reduce adverse impact and seem to provide viable alternatives to the operational composites.

Table 3. Qualifying Rate/Adverse Impact Ratio for Current and Alternate Composites

	Qualifying Rate - Adverse Impact Ratio		
	Male/Female	**NHW/NHB**	**NHW/HW**
Current Mechanical Composite	.606	.501	.793
AR + PC + WK	.836	.695	.864
AR + PC + WK + MK	.868	.730	.895
AR + PC + WK + AO	.877	.745	.942
AR + PC + WK + AO + MK	.901	.777	.961
Current Electronics Composite	.715	.669	.867
AR + MK	.893	.836	.973
AR+ MK + WK	.880	.776	.917
AR + MK + PC	.891	.804	.947
AR + MK + AO	.914	.840	.997
AR + MK + PC + WK	.883	.770	.907
AR + MK + AO + WK	.904	.803	.961
AR + MK + AO + PC	.909	.820	.979
AR + MK + AO + PC + WK	.904	.797	.952
Current Administrative Composite	.972	.902	.95
MK + PC + WK + GS	.923	.850	.921
MK + PC + WK + AR	.938	.873	.951
MK + PC + WK + AO	.966	.903	.978
MK + PC + WK + AR + GS	.905	.835	.928
MK + PC + WK + AO + GS	.940	.878	.961
MK + PC + WK + AO + AR	..948	.888	.973
MK + PC + WK + AO + AR + GS	.927	.866	.962
Current General Composite	.891	.820	.929
AR + PC + WK + MK	.927	.854	.942
AR + PC + WK + AO	.932	.865	.964
AR + PC + WK + GS	.880	.804	.906
AR + PC + WK + AO + MK	.940	.874	.972
AR + PC + WK + GS + MK	.895	.820	.920
AR + PC + WK + AO + GS	.910	.843	.948
AR + PC + WK + AO + GS + MK	.917	.850	.955

3.0 DEVELOPMENT AND VALIDATION OF ADVANCED ASSESSMENT METHODS: NON-COGNITIVE ASSESSMENTS

Four projects focused on improving the assessment of non-cognitive characteristics. These involved expansion of the constructs being measured, examination of alternate item formats, and comparison of assessment methods used across the Services. These were:

1) TAPAS item development and examination of alternate item formats
2) TAPAS, NCAPS, and SDI comparison and consolidation
3) Dark Tetrad item development
4) Air Force Compatibility Assessment (AFCA) review

3.1 TAPAS Item Development and Examination of Alternate Item Formats

Three versions of the Air Force Tailored Adaptive Personality Assessment System (TAPAS) were developed (Chernyshenko, Drasgow, Stark, & Nye, 2019). All assess the same 15 facets underlying the Big Five personality dimensions. The first was a 90-item static version that used the traditional single statement format with a 4-point Likert response scale. The second and third forms each had 120 forced-choice item pairs. The second version used a two-alternative forced choice response format where the pairs of statements were unidimensional (unidimensional pairwise preference, UDPP). The third version also used the two-alternative forced choice format, but most of the paired statements were multidimensional (multidimensional pairwise preference, MDPP). Software was developed to score all three response formats.

Six hundred eighty-nine (689) US Air Force Basic Recruits completed the three TAPAS forms to evaluate their psychometric characteristics. The forms were presented in two orders to control for order effects: (1) Likert, multidimensional forced-choice, and unidimensional forced-choice; and (2) unidimensional forced-choice, Likert, and multidimensional forced choice. The Basic Recruits were first asked to answer honestly for research purposes, and then to complete the forms a second time, where they were asked to do their best to "convince the Air Force that you would make a good Airman" (i.e., fake good). The Basic Recruits also completed five self-report scales that served as criterion measures (Situational Decision-Making, Communications, Decision-Making and Management, Leading Others, and Displaying Professionalism). They used a 5-point Likert format.

The single statement Likert form demonstrated higher reliability than either the UDPP or MDPP form, but was more susceptible to faking. Cross-format correlations revealed that the single statement Likert and UDPP scales tended to correlate with each other more highly than the MDPP scales did with either the Likert or UDPP scales. All 15 scales were used to predict each of the criteria. Examination of the validities indicated that scale scores computed from responses in the Faking condition did not predict the criterion variables. None of the adjusted R^2 values

exceeded .17, and many were less than .10. In contrast, scale scores from the Honest condition had adjusted R^2 values generally between .2 and .4, indicating fairly good validity. The Likert format scales generally had the highest correlations with the criterion variables. Their average adjusted R^2 was .38, which was noticeably higher than the adjusted R^2 values of the unidimensional forced-choice scales (.27) and multidimensional forced-choice scales (.22). One possible explanation for this pattern of results is that the criterion variables were assessed using Likert response scales, so the Likert AF TAPAS may have shared some mono-method response consistency error variance.

3.2 TAPAS, NCAPS, and SDI Comparison/Consolidation

A comprehensive review of three personality assessments used by the US military services for personnel selection and/or job placement (Kantrowitz, Kingry. Madaj, & Nye, 2019) was conducted. These were the TAPAS, NCAPS, and SDI. The objective of the project was to improve testing and test evaluation within the military, and specifically the consolidation of personality assessments developed by the US military over the past decade.

The project had four parts. Part 1 was an archival review of previously conducted research studies on the TAPAS, NCAPS, and/or SDI. Part 2 consisted of mapping analyses of the TAPAS with the NCAPS and SDI to identify scale overlap/uniqueness. Part 3 was an examination of the extent of shared variance between the TAPAS, NCAPS, and SDI and/or with other variables. Part 4 summarized a meta-analysis of validity results for each scale. In general, many of the assessments provided evidence of conceptual and empirical overlap between the tests. Some concerns exist regarding the operational performance of the assessments and suggestions were provided for how the military may improve the extent to which a combined approach to personality assessment under the TAPAS framework may improve job selection and/or placement decisions.

Results provide some indication that the US military can achieve the objective of moving toward a consolidated approach based on the TAPAS, NCAPS, and SDI to assess personality for selection and/or job placement. However, there are some challenges with fully achieving this objective. One of the most salient and central findings resulted from the investigation into overlap of the constructs measured across the three tests. The qualitative and quantitative reviews and analyses indicated a generally high level of convergence that the findings regarding construct overlap are robust and actionable.

The findings regarding the criterion-related validity of the three assessments were less robust and less interpretable. The archival research demonstrated typically zero or small effect sizes for facet-level scores with various criterion domains. The expectations for the magnitude of criterion-related validity coefficients were modest in recognition of the fact that single scales from any of these assessments are typically not evaluated or used on their own, not all scales are

predicted to relate to all criteria, and small effects, when combined, can have more substantial relationships with other variables. Even with low expectations for robust validity coefficients, most validities were characterized as zero and a few TAPAS scales consistently demonstrated at least a small effect size ($r > 0.10$). The NCAPS demonstrated somewhat more robust validity coefficients, although the number of studies was very limited. The criterion-related validity for the SDI was not typically robust. The meta-analyses further reinforced the conclusion that many of the constructs studied across various assessments tend to result in zero to small validities, and that there is much variability in the performance of the assessments, with most results near zero. Based on these results, it is not clear from the research reviewed, that TAPAS is likely to add robust prediction to job placement/selection scenarios or add incremental prediction to other assessment tools.

It should be noted that the criterion-related validity portion of this project was limited by the nature of the criteria data used/obtained in the quantitative analyses that were conducted.

3.3 Dark Tetrad Item Development

The AF TAPAS is a DoD-owned personality assessment measure rooted in the Big Five theory of personality. It contains 15 facets designed to assess personality factors related to performance in military specialties. The purpose of this effort was to expand personality assessment beyond the Big Five to include measures of “dark” traits.

The Dark Triad in psychology focuses on the personality traits of narcissism, Machiavellianism, and psychopathy (Pauihaus & Williams, 2002). Use of the term "dark" implies that people possessing these traits have malevolent qualities. People scoring high on these traits are more likely to commit crimes, cause social distress, and create severe problems for an organization (e.g., counterproductive work behavior), especially if they are in leadership positions.

Although the three Dark Triad traits are conceptually distinct, empirical evidence shows them to overlap. They are associated with a callous-manipulative interpersonal style (Jones & Paulhaus, 2010). Narcissism is characterized by egotism, grandiosity, pride, and a lack of empathy (Kohut, 1977), Machiavellianism is typified by manipulation and exploitation of others, a cynical disregard for morality, and a focus on deception and self-interest (Jacobwitz & Egan, 2006). Psychopathy is characterized by continuing antisocial behavior callousness, impulsivity, remorselessness, and selfishness (Skeem, Polaschek, Patrick, & Lilienfeld, 2011).

Some have suggested expanding the Dark Triad to include a fourth facet, sadism (Mededovic & Petrovic, 2015). Sadism represents a combination of different behavioral, cognitive, and interpersonal characteristics related to pleasure in connection with inflicting emotional or physical pain on others (Reidy, Zeichner, & Seibert, 2011) and to control, punish, and humiliate

others (Myers, Burket, & Husted, 2006). Reidy et al. (2011) found that sadism, separately from psychopathy, predicted unprovoked aggression in the laboratory context.

Drasgow, Chernyshenko, Stark, Nye, Zhang, Sun, and Li (2020) developed two forms to assess the Dark Tetrad traits. Both assess psychopathy, narcissism, Machiavellianism, and sadism, as well as four Bright side facets (achievement, even tempered, selflessness, and virtue) of the Big Five personality dimensions. The versions include a traditional single statement format with a 5-point Likert response scale and a two-alternative forced-choice format with multi-dimensional paired statements. Software for scoring the response formats was also developed. The statements were administered to a sample of USAF Basic Recruits and MTurk workers to assess their psychometric characteristics. After the statements were calibrated, the forms were assembled and administered to samples of MTurk workers and Prolific workers. Reasonably good cross-form correlations were found, as well as substantial correlations with alternative measures of the same construct. The single statement versions of the Dark Tetrad traits showed reasonable resistance to faking good.

Data were collected using both "Honest" and "Fake Good" response instructions. In addition to the TAPAS measures, criteria measures of the Big Five personality domains, Dark Tetrad, counterproductive work behavior, organizational citizenship behavior, and subjective well-being were collected. Convergent validity cross-format correlations were found to be reasonably good for the Bright Side scales collected in the Honest condition, but lower for the Dark Tetrad scales. They ranged from a mean correlation of .54 for the Bright Side scales to a mean correlation of .35 for the Dark Tetrad scales. For data collected in the Faking Good condition, cross-method correlations were much lower, ranging from a mean of .35 to .45 for the Bright Side scales and from .14 to .40 for the Dark Tetrad scales. After correcting for unreliability in the criteria, these correlations were even higher.

3.4 Air Force Compatibility Assessment Review

The purpose of this effort was to conduct an independent evaluation of the AFCA. The AFCA was developed by Air Force Personnel Center, Strategic Research and Analysis Branch (AFPC/DSYX) psychologists to identify enlisted applicants who are at high risk to exhibit counterproductive work behaviors. An independent team with expertise on integrity testing and test validation reviewed AFCA technical reports and analyzed raw data files from the AFCA project (Sackett, Shewach, Anderson, & Tomeh, 2019). Sackett et al. also provided a summary of the literature on each of the constructs assessed by the AFCA and summarized research on critical or promising integrity constructs not included in the AFCA. Independent peer review is a standard best practice in social science research because it helps to ensure rigor in application of scientific methods and enhance confidence in study results. An independent review is especially crucial for the AFCA, given the high stakes nature of the test, and sensitivity of issues surrounding the interpretation of results. Sackett et al. concluded that the results previously

reported by AFPC were supported by the reanalysis in regard to AFCA reliability, validity, comparability of passing rates across subgroups, and utility. The authors made several recommendations about improved measurement of current AFCA constructs (e.g., aggression, disinhibition, Machiavellianism) and expansion of content (e.g., Dark Triad, agentic goal orientation, and moral licensing).

4.0 DEVELOPMENT AND VALIDATION OF OCCUPATIONAL PERFORMANCE CRITERIA

Four projects focused on the development and validation of occupational performance criteria. These were:

1) DoD Job Performance Criterion Development
2) Weighted Airman Promotion System (WAPS) Evaluation
3) Navy Promotion Testing Evaluation
4) Months of Mission-Ready Service (MM-RS) Evaluation

4.1 DoD Job Performance Criterion Development

The overall objective of this effort was to document the job performance criteria unique to each Service's accession and classification testing efforts and to develop a unified set of job performance criteria that all Services could use. A standardized core set of job performance criteria would enable joint-Service validation studies and evaluation of the generalizability of predictor-criterion relationships.

4.1.1. Development of a Performance Taxonomy for Entry-Level, Skilled and Technical Occupations

The first step in this project was to develop a job performance taxonomy to describe the entire domain of early career enlisted job performance, including performance in training and during the first term of enlistment across Services (Allen, Ford, Russell, Carretta, & Kirkendall, 2019; Allen, Russell, Ford, Kirkendall, & Carretta, 2020). A literature review was conducted to identify a comprehensive list of performance dimensions for entry-level skilled and technical jobs. Campbell's (2012) model was used as a starting point because it has been well researched and documented, and is relevant for entry-level jobs. This process resulted in 33 dimensions. Seventeen expert raters organized these dimensions into broader categories. The resulting model had 4 broad dimensions at the highest level – Technical Performance, Organizational Citizenship and Peer Leadership, Psychological Well-Being, and Physical Performance, which had 11 sub-dimensions. In addition to job performance, Allen et al. (2019, 2020) identified two other important criterion domains – Attitudes and Organizational Outcomes. The job performance taxonomy subsequently was used to categorize hundreds of criterion instruments and identify (a) gaps and redundancies in extant criterion measures and (b) core, joint-Service criterion measures.

4.1.2. Identification of Measurement Gaps

A taxonomic structure was developed to organize current criterion instruments and identify measurement gaps (Allen et al., 2020). The resulting taxonomy had three broad domains – Job Performance, Attitudes, and Organizational Outcomes. Allen et al. (2020) identified 74 current criterion instruments and mapped them against the taxonomy. There were 13 job performance rating scales, 13 performance tests, 20 attitudinal surveys, and 28 variables from administrative data (e.g., attrition and performance records). Allen et al. (2020) recommended the Services 1)

maximize the use of administrative data, 2) improve measurement of attitudinal, outcome, and performance constructs, and 3) improve measurement of job performance to include measures of organizational citizenship and peer leadership and measures of technical performance with job performance ratings.

4.1.3. Development of Joint-Service Criterion Instruments for Enlisted Jobs

The purpose of this project was to develop a unified set of criterion measures that can be used by all US military Services. In the first phase of the project, Allen et al. (2020) developed taxonomies of job performance, attitudes, and organizational outcomes for first-term enlisted personnel; constructed a database of criterion measures used by the Services; and linked criterion measures to the performance domain constructs. The second phase of the project focused on the development of unified criterion measures, including performance appraisal rating forms, self-assessment tools, a situational judgment test, and a methodology for measuring attrition consistently across Services (Ford, Hu, Graves, Huber, Russell, Wilmot, &Ellis, 2020). The new unified set of criterion procedures and measures can be used in joint-service projects to evaluate the criterion-related validity of the ASVAB and other predictor measures.

4.1.3.1 Procedures to Align Outcome Variables across Services. After reviewing methods the Services use to measure and track outcome variables, Ford et al. (2020) developed procedures to standardize (or align) attrition, training outcomes, physical fitness scores, and counts of disciplinary incidents across the Services. Based on their review of the currently available databases, the new procedures for constructing attrition and training outcome variables should be feasible for immediate implementation. Similarly, the information needed to construct physical fitness scores appears to be available in military personnel databases, despite the relatively infrequent use of physical fitness scores in prior research. Of the outcome variables that were reviewed, disciplinary incidents were the most sparsely documented. The possibility of accessing the Defense Incident Based Reporting System (DIBRS) to obtain disciplinary incident data should be investigated before any associated recommendations can be implemented.

4.1.3.2 Performance Rating Scales to Address Job Performance Constructs. Cross-Service job performance rating scales (PRS) were developed to address the first -term enlisted job performance constructs identified in the first phase of this project (Allen et al., 2020). Two forms of PRS were developed: (a) one to be administered at the end-of-training (EOT) and (b) the other to be administered during the 1st term of Service. Both forms are suitable for use by supervisors or peers. The PRS were designed to measure all the constructs of the job performance taxonomy including: (a) Technical Proficiency, (b) Organizational Citizenship and Peer Leadership, (c) Psychosocial Well-Being, and (d) Physical Performance.

4.1.3.3 Self-Report Assessments. Three versions of self-report assessments were developed to extend coverage of the criterion domain.

- End-of-training (EOT) assessment – to measure dimensions of attitudinal constructs, performance constructs and organizational outcome constructs (physical and technical performance, reprimands, and awards);
- In-unit (IU) assessment – to measure dimensions of attitudinal, performance, and outcome constructs as with the EOT assessment;
- Exit survey – to provide a more fine-tuned measure of attrition, focusing on detailed reasons for attrition, withdrawal cognitions, counterproductive work behavior (CWB), and work satisfaction.

All three assessment types were designed for use across Services, and each include the minimal number of items required to measure the targeted constructs. In application, researchers will be able to supplement the core set of items with Service-specific items in support of their unique research needs.

4.1.3.4 Situational Judgment Test. A cross-service item pool for a situational judgment test (SJT) was developed to assess constructs in the job performance taxonomy that lacked coverage from other measures. The constructs targeted were (a) technical performance (decision making, problem solving, and innovation) and (b) organizational citizenship and peer leadership (planning and structuring work, conscientious initiative, support for peers, and organizational support).

The SJT was designed to assess test-takers' judgments about various problems they might encounter as E-3s and E-4s. Each SJT item consists of two parts: (a) a description of a hypothetical scenario and (b) several possible responses to that scenario. Examinees are asked to evaluate the effectiveness of different responses by comparing them to one another (e.g., indicating the best and worst responses) or by separately rating the effectiveness of each response.

4.1.3.5 DoD-Wide Criterion Measurement Research Plan. Based on input from military personnel researchers, we developed a plan for evaluating the psychometric characteristics of the new set of cross-service measures. The research plan describes a strategy for conducting the data collections that will be required for thorough evaluation of the criterion measures.

4.2 Weighted Airman Promotion System (WAPS) Evaluation

The WAPS determines promotions to the ranks of E-5 to E-9 within the USAF. The WAPS comprises a formula for weighting various components characterizing a person's readiness for promotion. Two standardized tests serve as WAPS components: (a) a Specialty Knowledge Test (SKT) – a measure of technical knowledge pertaining to the Air Force specialty (AFS) to which

the individual belongs, and (b) the Promotion Fitness Exam (PFE) – a measure of general Air Force knowledge covering topics such as history, customs, resource management, dress and appearance, and security. SKTs are specific to each AFS. The PFE is given to all service members of a given rank, regardless of AFS. Objective 1 was to analyze archival data to evaluate overall predictive (i.e., criterion-related) validity of the WAPS tests. Objective 2 was to analyze archival data based on examinees' first-time and repeated item exposure, Objective 3 was to summarize the literature on item exposure and tools/best practices to prevent and detect test compromise, and a summary of recommendations. Objective 4 was to evaluate a SJT prototype and detailed recommendations for future WAPS SJT development.

4.2.1. Examination of Overall Predictive Validity of the WAPS Tests

Bradley, Dahlke, McCloy, Reader, and Hu (2019) examined the predictive validity of the SKT and PFE scores for Enlisted Performance Report (EPR) ratings. Although there were some instances of non-negligible outcome prediction, there was little evidence overall regarding the predictive efficacy of the SKT or PFE. Similarly, weak relations were observed for the ASVAB, which has substantial evidence of its criterion-related validity for job performance. Bradley et al. speculated that the lack of supportive validity findings may be attributable more to properties of the outcome measures (criteria) than to deficiencies associated with the SKT or PFE. Specifically, the EPR ratings (a) were highly restricted in terms of score variability and (b) included non-technical, non-duty relevant considerations that appear relatively distinct from aspects of performance likely to be predicted by SKT or PFE scores. Although the EPR ratings might provide value to the Air Force from an operational standpoint, they might not be well-suited for validation research compared to measures that better differentiate Airman performance and are more relevant to the focal predictor constructs (e.g., research-only performance rating scales, hands-on work sample performance measures). Bradley et al. concluded that the evidence regarding the predictive validity of the SKT and PFE was inconclusive.

4.2.2. Examination of Predictive Validity of Examinees' First Time and Repeated Item Exposure

Airmen may complete the SKT or PFE on multiple occasions. Bradley et al. (2019) found little evidence that exposure affects properties of the items or examinee test scores. Further, they found little evidence that item exposure systematically affected demographic subgroup item-level performance or relations between Airman experience and item-level performance. In addition, there was little evidence that item-level psychometric properties (i.e., difficulty, discrimination, correlations between item-level scores and external variables) differed between first-time and repeat examinees, or that item exposure moderated these differences. Results from differential item functioning (DIF) analyses provided no systematic evidence that items functioned differently between first-time and repeat examinees, regardless of whether the items had been exposed. Given the general lack of supportive findings concerning the predictive validity of the PFE and SKT, Bradley et al. did not conduct analyses to investigate whether item exposure attenuates the criterion-related validity of the test scores.

4.2.3. Literature Review on Item Exposure and Tools/Best Practices to Prevent and Detect Test Compromise

To help address Air Force concerns regarding the effects of item exposure on WAPS test performance, Waugh, Walion, Burgoyne, and McCloy (2019) examined the testing literature concerning item exposure, test compromise, forensics to detect item/test compromise, and test security in general.

Ideally, recommendations should be based on a thorough test security audit. Some specific recommendations for WAPS are highlighted below. Obviously, some recommendations might not be feasible at this time. Some other recommendations might already be implemented or be planned for the near future.

1. *Use a short testing window.* A 1 or 2-week window is common for professional certification programs. The shorter the window, the less chance that test content will be communicated to examinees in that window. Longer windows might be necessary, but the length should be no longer than needed.

2. *Administer the exam electronically*. It is much more difficult to control access to paper exams. Electronic administration also allows for collection of item response time data which has various uses—including detecting test fraud and determining optimal testing time limits. Finally, CBT allows on-the-fly randomization of the item order.

3. *Create multiple test forms for each test administration*. The benefit of a candidate obtaining the test content on one test form is greatly reduced if he/she is administered a different test form. This might not be feasible for the SKT because each AFS has its own test. If paper forms are used, answer-copying or pre-knowledge of the answer key could be reduced if two or more versions are created that contain the same items but are in different orders.

4. *Ban cell phones from the testing room*.

5. *Formally train the test proctors*.

6. *Train item writers and other test contributors in test security*. Require them to sign a non-disclosure agreement.

7. *Carefully control test materials during test development*. This includes strict controls over paper materials and encrypting of electronic materials.

8. *Consider using a web-based item-banking application for test development*. User permissions can be set up so that item authors, reviewers, and other contributors see only the item content they need to see.

9. *Communicate to examinees that sharing item content is forbidden*. Require them to sign a non-disclosure agreement. Clearly communicate the consequences for test fraud.

10. *Consider converting the PFE to a computerized adaptive test (CAT)*.

11. *Compute some forensic statistics*. The appropriate statistics depend considerably on whether the test is CBT or paper, whether item response latency data is obtained, the testing volume, resources, and the statistical expertise of the staff. It also depends on the forensic goals and the type of test fraud that is most likely. It might be best to hire a consultant such as Caveon to perform the forensic services or, at least, to help determine what types of forensics would be appropriate.

12. *Analyze item parameter drift (e.g., changes in item difficulty over time) to help identify items that might be compromised*.

4.2.4. Literature Review on Best Practices in Situational Judgment Test (SJT) Development

Sullivan, Whetzel, and McCloy (2019) conducted a literature review summarizing best practices in SJT development. As with any selection method (e.g., job knowledge tests, assessment centers, interviews), SJT quality is influenced by decisions made regarding its design, development, and scoring. It is clear from both psychometric properties and test-taker response behavior that not all SJT designs are the same, and not all designs may be appropriate for the intended use and assessment goals. The approach to SJT development and scoring ultimately depends on several factors, including the assessment goals and end-user preferences, which is a testament to the extremely versatile, informative nature of SJT-based assessment.

Based on their review of the literature, Sullivan et al. (2019) suggested the following guidelines and best practices that will not be appropriate for every SJT, but provide a good starting point for developers seeking to employ an SJT in their selection system.

- *Response Instructions*: Use should-do questions unless seeking to assess personality traits or other behavioral tendencies, where would-do questions are better suited.

- *Response Format*: Consider the ***rate*** format. This provides data (e.g., effectiveness ratings) for all response options rather than just, say, two (as in the ***most/least*** format). It also thereby permits the largest range of potential scoring options.

- *Scoring*: Rational scoring is the most feasible approach.

- *SJT Reliability*: Test-retest or alternate forms (if the situation permits it) reliability estimates are preferred to internal consistency estimates.

- *Method of Presentation*: Video-based SJTs have several advantages in terms of higher face and criterion-related validity.
- *Faking*: Prevarication is more of a problem with SJTs requesting would-do responses, but currently the measures do seem less vulnerable overall than traditional personality measures.
- *Coaching*: There is some evidence that responding to SJTs can be coached, although some researchers believe some scoring methods are likely less coachable (e.g., within-person standardization). More research on this topic is needed.

4.3 Navy Promotion Testing Evaluation

Kubisiak, Kaplan, and Zorzie (2020) conducted a series of analyses to assist the US Navy (USN) in studying item exposure policies and other item and exam development policies as they relate to Navy-Wide Advancement Examinations (NWAE). Performance on the NWAE is one of several factors used in the Navy Enlisted Advancement System (NEAS) to determine advancement to the ranks of E-4 through E-7. The goal was to provide feedback on current practices and recommendations regarding ways to ensure that item and exam development processes are fair, valid, and credible.

The work consisted of three primary components. First, a focused literature review was conducted on topics related to item exposure. Additionally, analyses on archival data provided by the USN was performed to assess how repeated item exposure across multiple administrations of testing affected subsequent item analyses and Sailor item performance. Kubisiac et al. (2020) also provided a summary of best practices and recommendations on exam development policies concerning item exposure, random and randomized equivalent exams, development policies concerning item exposure, random and randomized equivalent exams, parallel items, SJT, and other measures.

The authors noted that many tools are available to detect possible negative effects of item exposure, but the key is to use multiple techniques over time. This is the only way to ensure consistent and reliable information that accurately identifies, and can be used to address, negative effects of item exposure. Many methods of controlling for item exposure were catalogued, such as modifications to tests and test banks, the addition of alternate forms, and changes to the testing environment. However, each of these methods have different limitations and benefits, so the selection of which to use must be carefully considered in the broader context of the overall testing process. Additionally, alternative methods of testing are available, such as SJTs, which can enhance and add incremental information to the testing process in a reliable, valid manner, while mitigating the impact of previous exposure to test takers.

Analyses of archival NWAE data indicated a consistent finding that the proportion of items that became easier was relatively stable across different lengths of time between administrations, whereas the proportion of items that became harder increased. Kubisiac et al. (2000) also found that the percentage of items that became easier was consistently, significantly higher than the percentage of items that became harder. Further, the percentage of items that became easier decreased as the length of time between administrations increased. Finally, it was observed that among candidates who exhibit performance changes between repeat and non-repeat items, a significantly greater number performed better on repeat items. Several interpretations of these results were discussed. Though they provide some unique insight into the current testing program, in and of themselves, they do not provide clear evidence for specific revisions such as increasing the length of time between administrations.

Overall, the results underscore the need to be careful about consistently analyzing item performance and ensuring that trends in the data do not change in unexplainable ways over time. If shifts are identified, tools are available to investigate what may be causing them, and several solutions can be implemented. However, the application of the appropriate tool is not always straightforward, and may have unanticipated secondary effects. Therefore, any actions taken must be considered in the context of the overall testing program to ensure fair, valid, and accurate testing.

4.4 Months of Mission-Ready Service (MM-RS) Evaluation

MM-RS is a measure of job performance that combines the periods that Airmen spend at different skill levels during a fixed time period (i.e., four years). The potential benefit of measuring and studying MM-RS is that it could serve as a long-term criterion that could improve the Air Force selection and classification process. The intent of the study was to assess the usefulness of using data routinely collected at time of enlistment to predict MM-RS (Gonzalez, Haight, & Martinez, 2021).

The core of MM-RS relies on the idea that a person can provide capability, or mission-readiness, to an AFS they are qualified to perform. This qualification is obtained after successful completion of Technical Training and award of a 3-skill level in an AFS. While trainees are learning, they are not performing, and thus they are not providing any mission-readiness for that specialty. To standardize measurement, MM-RS is calculated for the first 48 months after an individual enters active duty, regardless of their term of enlistment. For this study, the first 48 months of an Airmen's career was divided into four sections: BMT, Technical Training, 3-skill level, and 5-skill level. Credit was awarded for time spent providing readiness. MM-RS is *not* a direct measure of productivity, rather it is the *potential* to perform at a particular level of proficiency. Time spent during BMT and Technical Training were awarded no points. After graduation from Technical Training, Airmen start contributing to the readiness of their AFS. However, they are still gaining experience and receiving on-the-job training (OJT). This is reflected by weighting 3-skill level time half as much as 5-skill level time. Adding the weighted

3- and 5-skill level time in months produces MM-RS as calculated in previous studies. The only data needed to calculate MM-RS were: entered active duty date, 3- and 5-skill level dates of award, and the date of separation (DOS) or the date 48 months after entering active duty, whichever came first.

Gonzalez et al. (2021) replicated and extended a study by Halper, Goodman, and Alley (2010) with a larger and more current sample. An additional objective was to determine whether personality measures provide incremental prediction of MM-RS beyond that provided by the use of the variables established by Halper et al.

The initial goal of a direct replication of Halper et al. (2010) was modified by improvements in the calculation of MM-RS and the unavailability of predictors used in the original study. The data available for the study required extensive cleaning, and not all original variables could be satisfactorily salvaged. The calculation of MM-RS was also adapted to represent a standardized measure of mission-readiness such that the measures were more comparable across AFSs than had been possible in earlier studies.

Descriptive and inferential statistics were used to determine the extent to which MM-RS could be predicted by the available variables. Hierarchical regression analyses were used to determine the incremental validity of personality. The regressions were corrected for range restriction. Results indicated that MM-RS was significantly predicted by a combination of ASVAB scores, educational level, age, enlistment type, and enlistment term (R = .30). Correction for range restriction had little effect on the validity results. Although the addition of personality scores was statistically significant, the effect size was small.

Recommendations include using MM-RS as an added consideration for classifying Airmen into positions in which they will be most successful or in which the projected need of the Air Force is greatest. Further recommendations include adjusting the definition of MM-RS so that it focuses on setbacks in career progression instead of qualified time. Such a measure could avoid delays outside of the control of the airmen. Such an adjustment should reduce the error variance in the measure and improve its accuracy. Finally, alternative variables may better predict MM-RS, such as work interests (tasks, work environment) as measured by the Air Force Work Interest Navigator (AF-WIN). Consideration of these predictors has the potential to further enhance the selection and classification of Air Force personnel

5.0 DEVEOPLMENT AND VALIDATION OF ALGORITHMS TO OPTIMIZE THE USE OF HUMAN CAPITAL (PERSON-JOB FIT)

5.1 Advanced Accessions

The primary objective of this initiative was to develop a tool to optimize organizational and individual outcomes when matching enlisted personnel (applicants or trainees) to entry-level jobs (Ingerick, 2019). A core premise of the person-job matching tool is that both organizational and individual outcomes for the Air Force are optimized by the effective management of job-relevant talent pools. Talent pools are effectively managed when the aggregate supply of talent within the applicant or accessions population are optimally matched to jobs (or job clusters) to minimize over- and under-qualification, within and across jobs. Accordingly, the tool recommends "best" job matches, based on the projected payoff(s) from combining the selected inputs (person, job, organization) when matching people among qualifying jobs to minimize the over or under supply of talent. The report describes an initial prototype of the person-job matching tool. The project was planned as a two-year development effort. The objective of the second year was to refine the tool and build-in selected dynamic modeling feature(s). However, the project was ended after completion of Year 1 because similar projects were being conducted by other USAF organizations.

6.0 DEVELOPMENT AND VALIDATION OF IMPROVED IT INFRASTRUCTURE TO ENABLE REMOTE SPECIAL TESTING AND EFFICIENT DATA PROCESSING AND MANAGEMENT

6.1 Remote Special Testing/Data Management System: Phase I

The Air Force Recruiting Service (AFRS) uses entrance testing to screen applicants for enlistment qualification and suitability for specific career fields. These tests are administered at a limited number of locations, including Military Entrance Processing Stations (MEPS) and Military Entrance Test Sites (METS). The geographic separation between these testing locations and recruiter offices, as well as a limited capability for data sharing, can result in lengthy delays between an applicant expressing interest in an Air Force career and a recruiter having the necessary information to determine their eligibility. AFRS would like to leverage an existing testing platform, the Test of Basic Aviation Skills (TBAS) platform, to expedite the testing process for certain aptitude tests. The TBAS was developed for AFPC/DSYX to support the Pilot Candidate Selection Method (PCSM) program. TBAS systems and a scaled-down version of these TBAS systems (without joysticks or rudder pedals), which are already approved to administer the entrance tests required by AFRS, would be deployed to select recruiter offices within the continental United States. Tests given on these testing stations would be proctored remotely by a 3rd party vendor. This would alleviate the need for recruiters in proximity to these systems to send applicants to a MEPS or METS facility. Bennett, Forsythe, Kellaher, and Barborak (2019) researched the existing TBAS system infrastructure and conducted interviews with recruiters and test control officers in order to develop an implementation plan to achieve the testing capabilities required by AFRS. The plan outlines a method and timeline to develop, acquire, and deploy additional TBAS systems, as well as modernize the PCSM infrastructure to allow rapid data sharing between TBAS systems and recruiters.

6.2 Strategic Research and Assessment Program Data System Support: Phase I

The mission of the AFPC/DSYX is to improve person-job match in support of human capital management for optimal Air Force mission capability. To accomplish this, it must obtain all possible sources of testing, assessment, applicant, personnel, training, and performance data available regarding Air Force applicants and current Air Force members. These data are used to support an array of studies and analyses regarding the development, revision, and validation of assessment screening tools and processes. There are many sources of data, including Air Force recruiting, Basic Military Training, Air Education and Training Command (AETC) technical training, job performance, the emerging AETC training ecosystem, etc. The DoD and Air Force have begun initiatives to explore the potential use of social media and other forms of "Big Data" to contribute to these processes.

The purpose of this Phase I effort (Weissmuller & Cazaras, 2020) was to:

- Document the existing Strategic Policy Analysis Resource and Knowledge Base (SPARK) system,
- Apply specific expertise in system analysis, data management, IT processes, etc. to review current data sources and data processing procedures, investigate and identify other untapped or emerging sources of data, and
- Develop a vision and a roadmap with milestones of how to establish an effective, flexible approach for establishing the needed data infrastructure, methodologies for pulling and applying the data, searching and pulling meaningful patterns from the data, etc. to support the emerging multi-domain selection and classification assessment processes.

The Phase II objective is to research, evaluate, and recommend the way forward to replace existing SPARK processes with new programming and hosting capabilities. These new processes will facilitate data importation and integration to support AFPC/DSYX analytic research programs and efforts to brief and implement recommendations for operational deployments and on-going monitoring programs.

7.0 BEST PRACTICES GUIDE FOR ASSESSMENT DEVELOPMENT, VALIDATION, AND IMPLEMENTATION

As AFPC/DSYX continues to develop assessments and refine selection and classification models for a large number of Air Force career fields, consistency in the procedures used and how work is presented will be increasingly important. The objective of this project was to develop summaries of best practices to allow AFPC/DSYX researchers with advanced degrees in fields involving test development and statistical modeling to conduct studies that meet professional standards for quality and completeness. This series of reports consolidates the experience, wisdom, and tools the Air Force has accumulated in its selection and classification work, and blends them with best practice recommendations from industry. The resulting reports will help ensure that the Air Force remains at the forefront in developing and using rigorous and innovative selection and classification methods.

In addition to covering content at a conceptual level, to maximize efficiency, ease of understanding, and use, it was considered essential that the best practices guides incorporate the following types of ancillary materials wherever possible and relevant:

- Concrete examples (e.g., examples tailored to, or that could be easily generalized to, an Air Force context)
- Step-by-step how-to's (e.g., running and interpreting all relevant statistical analyses)
- Sample materials (e.g., examples of effective graphic displays)
- Sample code (e.g., commonly-used SAS code)

AFPC/DSYX personnel wrote two reports regarding internal practices. The four reports developed by the contractor concerned the following topics:

1. Test Development and Validation
2. Selection and Classification Model Development
3. Reporting/Briefing Results
4. Ethical and Legal Considerations

7.1 Test Development and Validation

This report (LeBreton, 2021) covers test development and validation, providing recommendations and best practices in the Air Force. The recommendations are based on over a century of scientific research and practice, both within the USAF and in the scientific literature more generally. This report addresses five major topics. The first is validity and the validation process. The second covers steps/stages in the test development and validation process. The third discusses using classical test theory (CTT) to evaluate items and build tests. The fourth goes over using IRT to evaluate items and build tests. The last section discusses item bias and test bias. The report also includes appendices containing annotated R code for conducting CTT and IRT item analyses described in the text proper.

7.2 Selection and Classification Model Development

This report (Ployhart, 2021) provides recommendations and best practices in selection and classification for the Air Force Personnel Center/Strategic Research and Assessment Branch. The recommendations are based on over a century of scientific research and practice, both within the USAF and in the scientific literature more generally.

Selecting the right talent and classifying each person into the specialty and occupation that best fit their talents is vital for effective individual and organizational performance. Selection and classification are the first steps in the management of talent. Consequently, every downstream activity (training, development, succession planning) benefits from more rigorous selection and classification. Performance, learning, development, retention, and satisfaction are all improved by effective selection and classification.

The report begins with a brief summary of existing Air Force practices and challenges, introducing key definitions and basic concepts. This leads into a discussion of model development including job analysis as a foundation and how to select or generate predictor and criterion measures. From here, the report describes techniques for establishing evidence of predictive relationships, including methods for handling artifacts and conditions affecting statistical estimates of those relationships, and methodologies for combining predictor scores.

The report then turns to a discussion of the different types of selection and classification systems, breaking them down into five broad approaches, identifying key characteristics and utility of each. This is followed by additional practical considerations, including subgroup differences and adverse impact. Next, the report reviews strategies for generalizing from experimental to operational use of the selection and classification models, and concludes with the identification of future trends to monitor going forward.

7.3 Reporting Briefing Results

Proper deployment of selection techniques requires a high degree of technical sophistication to properly use and interpret results from complex psychometric and statistical analyses. Because of this, the selection expert communicating this information must execute a careful balancing act, maintaining the precision and transparency demanded by professional standards, while simultaneously conveying the essence of findings to a non-technical audience.

This report (Morris, 2020) provides a review of relevant professional standards that exist to guide reporting of research findings, both at a broad level and specific to criterion-related validation. It then discusses general considerations for reporting criterion-related validity results, and specific guidance for information that should be provided at a minimum. Finally, it covers strategies and techniques for effectively communicating validity evidence in a way that retains the required technical information, while making it accessible to audiences.

7.4 Ethical and Legal Considerations

The purpose of this effort (Gutman, 2020) was to discuss best practices for understanding and evaluating critical legal issues related to adverse impact. The report is divided into six sections:

Section 1 provided an overview of types of discrimination, including distinctions between facial discrimination, disparate treatment, pattern or practice of discrimination, and adverse impact. It was noted that because both adverse impact and pattern or practice of discrimination use statistical disparities, these two forms of discrimination are often confused with each other.

Section 2 summarized major Supreme Court rulings on adverse impact. It was noted that the early traditions in *Griggs* and *Albemarle* were temporarily altered in *Wards Cove* before being recovered, for the most part, in CRA-91.

Section 3 focused on key guidelines in the *Uniform Guidelines* that were overturned in the court. This included overturning guidance prohibiting use of content validity to validate tests of mental ability and the *Connecticut* ruling relating to multiple hurdles. Thus, adverse impact must be defended if there is any adverse impact in any of the steps.

Section 4 discussed how to create a test battery based on critical knowledge, skills, and abilities (KSAs) derived from a job analysis. The discussion included previously validated in-house tests, off-the-shelf tests, and newly-developed tests. A criterion validity study was illustrated.

Section 5 first noted that off-the-shelf tests should be criterion validated because they are copyrighted and their items cannot be removed or added to connect to critical KSAs of a job analysis. It was noted that well-conducted content validity studies (from scratch) can be used for cognitive tests. A method for content validating a newly-developed test can be used and connected to critical KSAs from a job analysis.

Section 6: The final section focused on methods for reducing adverse impact. It was concluded that race norming is strictly illegal under CRA-91; that a limited form of race-conscious banding may be used as a tie breaker after prior steps have been satisfied; that using an alternative test or selection procedure that reduces or eliminates adverse impact is legal; that altering items on a test after it has been administrated has been supported in one case, but this procedure was used because of a consent decree; and lastly, a good faith effort of losing an adverse impact claim is not sufficient - certainty of losing such a claim is the criterion.

It is important to recognize that the Federal antidiscrimination laws do not apply directly to military selection issues. That said, they are important insofar as the regulations that govern some of these laws, most notably adverse impact, play an important role in determining if a selection test or other selection procedures are valid. Valid tests are needed in order to successfully predict positive outcomes in personnel selection.

8.0 OPTIMIZATION OF HUMAN CAPITAL

The objectives of this project were to:

- Explore the historical, cultural, and organizational contexts related to human capital management within the Air Force system of operations
- Review the personnel research/personnel assessment/selection and classification processes and practices that are based on Air Force policy related to military personnel selection and classification testing
- Review the current and emerging military personnel selection and classification research programs within the Air Force personnel community and operational commands
- Identify state-of-the-art approaches in selection and testing contexts (methods, tools, analyses, and decisions)
- Review emerging approaches to selection (e.g., gamification, social media, big data) and how these can be vetted and leveraged, while still ensuring a strong scientific foundation and rigorous validation of methods
- Examine prediction and selection systems that operate in the broader context of recruiting, accession, and attrition processes
- Identify strengths, weaknesses, needs, complementarities, and duplications of those current military personnel selection and classification research programs, as informed by established professional principles and best scientific evidence in this domain
- Recommend a roadmap of goals and timeline for executing enhancements that include needs for design and implementation.

To assist the Air Force in preparing now for the challenges its Airmen will face in the future, the National Academies of Sciences, Engineering, and Medicine (NASEM; 2020) report provided a Flight Plan to strengthen the USAF human capital management system through three priorities, each with the committee's overarching recommendation and specific implementable action item

8.1 Study Approach

In 2019, to better understand where and how to implement changes, the Air Force requested that the NASEM conduct a study to examine how to strengthen the USAF human capital management system in support of optimal mission capability. Their report represents the final consensus of the inter-disciplinary expert committee appointed by NASEM to conduct that study, under the auspices of the NASEM Board on Human-Systems Integration. Members of the committee served as volunteers and represent relevant academic and other research areas (e.g., industrial and organizational psychology, economics, human-systems integration, computer sciences, and cybersecurity) and human capital management practitioners, including experts in Department of Defense and USAF human capital. In conducting its study, the committee

received input from numerous stakeholders located across the United States who provided unique perspectives representing multiple communities inside and outside the Air Force. Additionally, the committee considered relevant information provided by invited expert speakers from academia, government, and private industry, as well as numerous previously published products including published research and professional guidelines and standards and Air Force doctrine, strategic documents, and studies.

The USAF human capital management (HCM) system is not easily defined or mapped. It affects virtually every part of the Air Force because workforce policies, procedures, and processes impact all offices and organizations that include Airmen, and responsibilities and relationships change regularly. To ensure the readiness of Airmen to fulfill the mission of the Air Force, strategic approaches are developed and issued through guidance and actions of the Office of the Deputy Chief of Staff for Manpower, Personnel and Selection (AF/A1) and the Office of the Assistant Secretary of the Air 102 Force for Manpower and Reserve Affairs (SAF/MR). The committee's Flight Plan is designed to assist those offices.

In conducting its study, the committee focused on understanding the opportunities and challenges associated with related interests and needs across the USAF human capital management system as a whole. To begin to understand the dynamics of the system, the committee developed a model of the ecosystem of the USAF human capital management system that emphasizes influential internal (e.g., standards for enlistment) and external (e.g., local unemployment rates) variables and relationships rather than specific Air Force offices or organizations. The resulting ecosystem model makes very clear the cascading or rippling effects of policy decisions felt throughout the organization, sometimes with 111 far-reaching and unanticipated effects.

8.2 The Path Ahead

The recommendations and action items of this report offer the Air Force a strategic approach, across a connected human capital management system, to develop 21st century human capital capabilities essential for the success of 21st century Airmen. Although some elements of the needed system now exist, this report points to a number of critical gaps. Should these gaps be left unfilled, the Air Force will fall short of the excellence in human capital management essential to meeting its responsibilities under the National Defense Strategy. Executing the agenda laid out by the Flight Plan will be demanding, but the rewards will be commensurate with the investment made into the USAF human capital management system. Four recommendations were provided:

1. The USAF should deliberately manage Airman through a connected human capital management system, using data-driven decisions based on data systematically collected and analyzed.

2. The USAF should ensure Force effectiveness through evidence-based practices across a connected human capital management system to optimally match Airman to career fields, training, and job assignments.

3. The USAF should invest in research that ensures that decisions about Airmen from accession to separation reflect professional best practices, evolve with changing technology and mission demands, and are integrated across the human capital management system.

REFERENCES

Allen, M., Ford, L., Russell, T., Carretta, T. R., & Kirkendall, C. (2019). *Development of a performance taxonomy for entry-level skilled and technical occupations.* Presentation at the 34th annual conference of the Society for Industrial and Organizational Psychology, National Harbor, MD.

Allen, M., Russell, T., Ford, L., Kirkendall, C., & Carretta, T. R. (2020). *Identification of criterion constructs and measures for joint-service enlisted jobs,* AFRL-RH-WP-TR-2020-0019. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Warfighter Interface Division, Collaborative Interfaces and Teaming Branch.

Beatty, A. S., Burke, M. I., Koch, A. J., & Trippe, D. M. (2020). *Next-generation Armed Services, Vocational Aptitude Battery (ASVAB),* AFRL-RH-WP-TR-2020-0099. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Bennett, G., Forsythe, B., Kelleher, S., & Barborak, G. (2019). *Air force remote special testing and data management system implementation plan,* AFRL-RH-WP-TR-2019-0062. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Warfighter Interface Division, Collaborative Interfaces and Teaming Branch.

Bradley, K. M., Dahlke, J. A., McCloy, R. A., Reader, M. C., & Hu. M. (2019). *Evaluation of criterion-related validity and potential item exposure effects for the Weighted Airman Promotion System (WAPS),* AFRzl-RH-WP-TR-2019-0067. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Warfighter Interface Division, Collaborative Interfaces and Teaming Branch.

Brady, M. F., Shermadou, Z. S., & Carretta, T. R. (2020). An investigation of aptitude requirements for airmen working with autonomous systems, AFRL-RH-WP-TR-2020-0114. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Brady, M. F., Shermadou, Z. S., Gibson, A., & Carretta, T. R. (2020). *An investigation of aptitude requirements for human operators in human-automation interaction,* AFRL-RH-WP-TR-2020-0021. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Warfighter Interface Division, Collaborative Interfaces and Teaming Branch.

Campbell, J. P. (2012). Behavior, performance, and effectiveness in the 21st century. In S. Kozlowski (Ed.), *The Oxford Handbook of Organizational Psychology*: Volume 1. NY: Oxford University Press.

Campion, M. A., & Campion, E. D. (2019). *Literature review of computer-assisted text analysis research software analytical techniques, and best practices,* AFRL-RH-WP-TR-2019-0100. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Warfighter Interface Division, Collaborative Interfaces and Teaming Branch.

Campion, M. A., & Campion, E. D. (2020). *Final report on text analysis tools to enhance personnel assessment,* AFRL-RH-WP-TR-2020-0079. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Campion, M. C., Campion, M, A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101,* 958-975.

Chernyshenko, O. S., Drasgow, F., Stark, S. & Nye, C. D. (2019). *The development of three static personality research forms and on-line scoring tools for the U. S. Air Force,* AFRL-RH-WP-TR-2019-0090. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Collaborative Interfaces and Teaming Branch.

Coovert, M. D., Martin, J., Howard, D. J., Kim, F., Dreibelbis, R., Arbogast, M. S., & Potter, S. (2019). *Use of technology-enhanced work simulations for cyber assessment,* AFRL-RH-WP-TR-2019-0043. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Collaborative Interfaces and Teaming Branch.

Coovert, M. D., Wierenik, B. N., & Martin, J. (2020). *Use of technology-enhanced simulations for cyber aptitude assessment: Phase II prototype development,* AFRL-RH-WP-TR-2020-0055. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Collaborative Interfaces and Teaming Branch.

Drasgow, F., Chernyshenko, O. S., Stark, S., Nye, C. D., Zhang, B., Sun, T., & Li, L. (2020). *Development of dark tetrad research forms for the U. S. Air Force,* AFRL-RH-WP-TR-

2020-0106. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Ford, L. A., Hu, M. C., Graves, C. R., Huber, C. R., Russell, T. L., Wilmot, M. P., & Ellis, B. 2020). *Development of joint-service criterion instruments for enlisted jobs,* AFRL-RH-WP-TR-2020-0036. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Supervisory Collaborative Interfaces and Teaming Branch.

Gonzalez, M., Haight, N., & Martinez, L. (2021). *Months of mission-ready service: Refinement of a criterion measure for a modernized selection and classification system,* AFRL-RH-WP-TR-2021-0013. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Guttman, A. (2020). *Air Force Personnel Center best practices guide: Legal and ethical issues in personnel testing,* AFRL-RH-WP-TR-2020-0103. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Halper, M. L., Goodman, T. M., & Alley, W. E. (2010). *Months of mission-ready service.* San Antonio, TX: Operational Technologies.

Ingerick, M. (2019). *Advanced accessions system: Documentation on year 1 prototype person-job matching tool,* AFRL-RH-WP-TR-2019-0041. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Supervisory Collaborative Interfaces and Teaming Branch.

Jacobwitz, S., & Egan, V. (2006). The 'dark triad' and normal personality traits. *Personality and Individual Differences, 40,* 331-339.

Johnson, C. D., & Zeidner, J. (1995). *Differential assignment theory sourcebook*, ARI Research Note 95-43. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Johnson, R., Leen, M., & Goldberg, D. (2007). *Testing adaptive levels of automation (ALOA) for UAV supervisory control,* AFRL-HE-WP-TR-2007-0068. Wright-Patterson AFB, OH: Air Force Research Laboratory, Warfighter Interface Division, Systems Control Interfaces.

Johnson, R., Leen, M., Goldberg, D., & Chiu, M. (2005). *Adaptive levels of automation (ALOA) for supervisory control,* AFRL-HE-WP-TR-2005-0062. Wright-Patterson AFB, OH: Air Force Research Laboratory, Warfighter Interface Division, Systems Control Interfaces.

Jones, D. N., & Paulhaus, D. L. (2010). Differentiating the dark triad within the interpersonal circumplex. In L. M. Horowitz & S. N. Stark (Eds.), *Handbook of interpersonal theory and research* (pp. 249-267). NY: Guilford.

Kantrowitz, T., Kingry. D., Madaj, C., & Nye, C. (2019). *Navy Computerized Adaptive Personality Scales (NCAPS) and Self-Description Inventory (SDI) wind down and merger with Tailored Adaptive Personality Assessment System (TAPAS),* AFRL-RH-WP-TR-2019-0095. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Collaborative Interfaces and Teaming Branch.

Koch, A. J., Trippe, D. M., Beatty, A. S., & Shewach, O. R. (2018). *Cyber test development,* AFRL-RH-WP-TR-2018-0120. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Collaborative Interfaces and Teaming Branch.

Kohut, H. (1977). *The restoration of the self.* NY: International University Press.

Kubisiak, C., M., Kaplan, & Zorzie, M. (2020). *Navy promotion testing research: Item exposure analyses,* AFRL-RH-WP-TR-2020- 0028. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Collaborative Interfaces and Teaming Branch.

LeBreton, J. (2021). *Air Force Personnel Center best practices guide: Test development and validation,* AFRL-WP-TR-2021-0002. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Mathews, J. J., & Roach, B. W. (1983). Reading abilities tests: *Development and norming for Air Force use,* AFHRL-TR-82-26. Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Mededovic, J., & Petrovic, B. (2015). The dark triad: Structural properties and location in the personality space. *Journal of Individual Differences, 36*, 228-236.

Morris, S. B. (2020). *Air Force Personnel Center best practices guide: Briefing validation results,* AFRL-RH-WP-TR-2020-0086. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Performance Optimization Branch.

Myers, W. C., Burket, R. C., & Husted, D. S. (2006). Sadistic personality disorder and comorbid mental illness in adolescent psychiatric inpatients. *Journal of the American Academy of Psychiatry and the Law Online, 34,* 61–71.

National Academies of Sciences, Engineering, & Medicine (2020). *Strengthening U. S. Air Force human capital management: A flight plan for 2020-2030.* Washington, DC: National Academies Press.

Pauihaus, D. L., & Williams, K. M. (2002). The dark triad in personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality, 36,* 556-563.

Pena, D., Martinez, L., Haight, N., & Shore, C. W. (2021). *Air Force Reading Abilities Test: Revisions and additions,* AFRL-RH-WP-TR-2021-0012. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Performance Optimization Branch.

Ployhart, R. E. (2021). *Air Force Personnel Center best practices guide: Selection and classification model development,* AFRL-RH-WP-TR-2021-0003. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Performance Optimization Branch.

Reidy, D. E., Zeichner, A., & Seibert, L. A. (2011). Unprovoked aggression: Effects of psychopathic traits and sadism. *Journal of Personality, 79,* 75–100.

Sackett, P. R., Shewach, O. R., Anderson, E. L., & Tomeh, D. H. (2019). *Air Force Compatibility Assessment review.* Presentation to the Manpower Accession Policy Working Group.

Skeem, J. L., Polaschek, D. L., Patrick, P. J., Lilienfeld, S. O. (2011). Psychopathic personality: Bridging the gap between scientific evidence and public policy. *Psychological Science in the Public Interest, 12,* 95-162.

Sullivan, T. S., Burgoyne, T. C., McCloy, R. A., & Whetzel, D. L. (2019). *Review of situational judgment test (SJT) prototype development process and materials,* AFRL-RH-WP-TR-2019-0058, Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human

Performance Wing, Airman Systems Directorate, Collaborative Interfaces and Teaming Branch.

Sullivan, T. S., Whetzel, D. L., & McCloy, R. A. (2019). *Literature summary: Best practices in situational judgment test (SJT) development,* AFRL-RH-WP-TR-2019-0044. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Collaborative Interfaces and Teaming Branch.

Trippe, D. M., Moriarty. K. O., Russell, T. L., Carretta, T. R., & Beatty, A. S. (2014). Development of a cyber/information technology knowledge test for military enlisted technical training qualification. *Military Psychology, 26,* 182-198.

Waugh, G. S., Walion, N. J., Burgoyne, T. C., & McCloy, R. A. (2019). *Summary of literature review of item exposure, test security, and forensics for the Weighted Airman Promotion System (WAPS),* AFRL-RH-WP-TR-2019-0042. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Supervisory Control and Cognition Branch.

Weissmuller. J. J., & Cazares, K. A. (2020). *Strategic Policy Analysis Resource and Knowledge (SPARK) version 1.0 system documentation,* AFRL-RH-WP-TR-2020-0075. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Systems Directorate, Performance Optimization Branch.

Zeidner, J., & Johnson, C. D. (1991). *The economic benefits of predicting job performance, Vol. 3: Estimating the gains of alternative policies*. NY: Praeger,

Zeidner, J., & Johnson, C. D. (1994). Is personnel classification a concept whose time has passed? In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 377-410). Hillsdale, NJ: Erlbaum.

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMNS

R^2	Multiple R-squared
17DX/SX	Network Operations/Cyber Space Operations
1B4X1	Cyber Warfare Operations
1N4X1A	Digital Network Analyst
3D0X2	Cyber Systems Operations
3D1X2	Cyber Transport Systems
A	Air Force ASVAB Administrative Composite
AETC	Air Education and Training Command
AF/A1	Office of the Deputy Chief of Staff for Manpower, Personnel and Selection
AFCA	Air Force Compatibility Assessment
AFQT	ASVAB Armed Forces Qualification Test Composite
AFPC/DSYX	Air Force Personnel Center, Strategic Research and Analysis Branch
AFRAT	Air Force Reading Abilities Test
AFRS	Air Force Recruiting Service
AFS	Air Force Specialty
AFSC	Air Force Specialty Code
AF-WIN	Air Force Work Interest Navigator
ALOA	Adaptive Levels of Autonomy
AO	ASVAB Assembling Objects Subtest
AR	ASVAB Arithmetic Reasoning Subtest
AS	ASVAB Auto & Shop Subtest
ASVAB	Armed Services Vocational Aptitude Battery
BMT	Basic Military Training
CAT	Computer-Adaptive Test
CATA	Computer-Assisted Text Analysis
CBT	Computer-Based Testing
CTT	Classical Test Theory
CWB	Counterproductive Work Behavior
DIBRS	Defense Incident Based Reporting System
DIF	Differential Item Functioning
DoD	Department of Defense
DOS	Date of Separation
E	Air Force ASVAB Electronics Composite
EI	ASVAB Electronics Information Subtest
EOT	End of Training
EPR	Enlisted Performance Report
G	Air Force ASVAB General Composite
GS	ASVAB General Science Subtest
HAI	human-automation interaction
HCM	Human Capital Management
HMT	human-machine teaming
ICTL	Information and Communications Technology Test
IRT	Item Response Theory
ISR	Intelligence, surveillance, and reconnaissance
IT	Information Technology

IU	In-Unit
KSA	knowledge, skills, and abilities
LIWC	Language Inquiry and Word Count
M	Air Force ASVAB Mechanical Composite
MC	ASVAB Mechanical Comprehension Subtest
MDPP	Multidimensional Pairwise Preference
MEPS	Military Entrance Processing Stations
METS	Military Entrance Test Sites
MK	ASVAB Math Knowledge Subtest
MM-RS	Months of Mission-Ready Service
NASEM	National Academies of Sciences Engineering, and Medicine
MTurk	Mechanical Turk
NCAPS	Navy Computer-Adaptive Personality System
NEAS	Navy Enlisted Advancement System
NWAE	Navy-Wide Advancement Examinations
OJT	On-the-Job Training
OTS	Officer Training School
PAS	Perfect Automation Schema
PCSM	Pilot Candidate Selection Method
PC	ASVAB Paragraph Comprehension Subtest
PFE	Promotion Fitness Exam
PRS	Performance Rating Scale
R	Multiple Correlation
RGL	Reading Grade Level
RJP	Realistic Job Preview
RPA	Remotely Piloted Aircraft
SAOC	skills, abilities, and other characteristics
SAF/MR	Office of the Assistant Secretary of the Air Force for Manpower and Reserve Affairs
SDI	Self-Description Inventory
SJT	Situational Judgment Test
SKT	Specialty Knowledge Test
SME	Subject Matter Expert
SPARK	Strategic Policy Analysis Resource and Knowledge
TAPAS	Tailored Adaptive Personality Assessment System
TBAS	Test of Basic Aviation Skills
UDPP	Unidimensional Pairwise Preference
USAF	United States Air Force
USN	US Navy
VE	ASVAB Verbal Expression Composite
WAPS	Weighted Airman Promotion System
WK	ASVAB Word Knowledge Subtest