



AFRL-RH-WP-TR-2023-0049

**OPTIMIZING QUALIFICATION TO U.S. SPACE FORCE
SPACEPOWER DISCIPLINES**

**Oren R. Shewach
Michael Ingerick
Reese Butterfuss
Human Resources Research Organization**

**Thomas R. Carretta
711 HPW/RHBC**

**Carl Persing
Infoscitex, a DCS company**

**August 2023
Interim Report**

DISTRIBUTION STATEMENT A. Approved for public release.

**AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
HUMAN EFFECTIVENESS 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-2023-0049 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH THE ASSIGNED DISTRIBUTION.

CARRETTA.THOMAS.R.1228984929
Digitally signed by
CARRETTA.THOMAS.R.1228984929
Date: 2023.09.28 16:30:28 -04'00'

THOMAS R. CARRETTA, PhD
Work Unit Manager
Performance Optimization Branch
Air and Space Biosciences Division

WILLIAMS.LOGAN.ANDREW.1273597634
Digitally signed by
WILLIAMS.LOGAN.ANDREW.1273597634
Date: 2023.09.29 10:57:49 -04'00'

LOGAN A. WILLIAMS, PhD
Human Performance Product Area Lead
Product Development
Performance Optimization Branch
Air and Space Biosciences Division

This report is published in the interest of scientific and technical information. 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	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to the Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YY) 14-08-23		2. REPORT TYPE Interim		3. DATES COVERED (From - To) 01-09-21 - 28-07-23	
4. TITLE AND SUBTITLE Optimizing Qualification to U.S. Space Force Spacepower Disciplines				5a. C30-08-21 ONTRACT NUMBER FA8650-21-F-4104	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Oren R. Shewach ^a , Michael Ingerick ^a , Reese Butterfuss ^a , Thomas R. Carretta ^b , and Carl Persing ^c				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER H12Q	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Human Resources Research Organization ^a Infoscitex, a DCS Company ^c 66 Canal Center Plaza, Suite 700 4027 Colonel Glenn Highway, Alexandria, VA 22314-1578 Suite 210 Dayton, OH 45431-1672				8. PERFORMING ORGANIZATION	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Materiel Command ^b Air Force Research Laboratory 711 th Human Performance Wing Airman Systems Directorate Air and Space Biosciences Division Performance Optimization Branch Wright-Patterson AFB, OH 45433				10. SPONSORING/MONITORING AGENCY ACRONYM(S) 711 HPW/RHBC	
				11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S) AFRL-RH-WP-TR-2023-0049	
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution Statement A: Approved for public release; distribution unlimited.					
13. SUPPLEMENTARY NOTES Report contains color. AFRL-2023-4700, cleared 22 September 2023					
14. ABSTRACT Predictive Success Models (PSMs) were developed and evaluated to qualify USSF enlisted and officer applicants for spacepower disciplines. A PSM development strategy was employed, combining scores on cognitive aptitude tests with personality assessments, that considered both predicted training success and gender racial-ethnic diversity of qualified applicants as objectives. Specifically, PSMs incorporated scores on two personality assessments, the Tailored Adaptive Personality Assessment System (TAPAS; Stark et al., 2014) for enlisted applicants and the Self Descriptive Inventory (SDI; Woolley et al., 2022) for officer applicants, in addition to scores on the Armed Services Vocational Aptitude Battery (ASVAB) for enlisted and Air Force Officer Qualifying Test (AFOQT) for officers. Enlisted PSMs showed improvements over existing ASVAB-based qualification composites, with results varying by career field. Officer PSMs showed more limited improvements over existing AFOQT-based qualification composites. Projected impact analyses highlight specific improvements in predicted training success, qualification rates, and adverse impact ratios for females and racial-ethnic minority applicants when using PSMs in job qualification.					
15. SUBJECT TERMS Personnel selection, personnel classification, validation, Armed Services Vocational Aptitude Battery, ASVAB, personality assessment, Tailored Adaptive Personality Assessment System, TAPAS, Air Force Officer Qualifying Test, AFOQT					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT: SAR	18. NUMBER OF PAGES 68	19a. NAME OF RESPONSIBLE PERSON (Monitor) Thomas R. Carretta 19b. TELEPHONE NUMBER (Include Area Code)
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			

TABLE OF CONTENTS

ACKNOWLEDGEMENT	iv
1.0 EXECUTIVE SUMMARY	1
1.1 Background and Overview of Technical Approach.....	3
2.0 PREPARING ANALYSIS SAMPLES AND PREDICTOR-CRITERION DATA ...	5
2.1 Identifying Criterion Measures Aligned with Job Requirements	5
2.2 Preparation of Analysis Samples and Construction of Criterion Composites ..	9
2.2.1 Preparation of Analysis Samples.....	11
2.2.2 Construction of Criterion Composites.....	12
2.3 Analyzing Relationships Between Predictors and Criteria.....	13
2.3.1 Validity and Subgroup Difference Estimation Approach	13
2.3.2 Enlisted Validity and Subgroup Difference Results	16
2.3.3 Officer Validity and Subgroup Difference Results	18
3.0 DEVELOPING PREDICTIVE SUCCESS MODELS (PSMs) BY CAREER FIELD21	
3.1 PSM Modeling Methodology	21
3.2 Enlisted Predictive Success Model (PSM) Results.....	26
3.3 Officer Predictive Success Model (PSM) Results	33
4.0 EVALUATING PROJECTED IMPACT OF CANDIDATE PSMs.....	37
4.1 Enlisted Projected Impact Approach	38
4.2 Enlisted Projected Impact Results	40
4.3 Officer Projected Impact Approach	48
4.4 Officer Projected Impact Results	48
5.0 DISCUSSION	53
6.0 CONCLUSION	56
7.0 REFERENCES.....	57
LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS	60

LIST OF TABLES

Table 1. Enlisted and officer targeted AFSCs/SFSCs and sample sizes	10
Table 2. Enlisted and officer predictor subtests/facets.....	14
Table 4. Candidate Standards for 1N4X1B/5I4X1B.....	46
Table 5. Candidate Standards for 3D0X1/5C0X1	47
Table 6. Officer Projected Impact – Career Field Matching for Space Operations	51
Table 7. Officer Projected Impact – Fixed QR for Space Operations	52

LIST OF FIGURES

Figure 1. Most common work activities for enlisted AFSCs	7
Figure 2. Most common work activities for officer AFSCs	8
Figure 3. ASVAB/TAPAS subtest FPPS validity estimates across career fields.....	17
Figure 4. ASVAB/TAPAS subtest subgroup difference estimates in USAF enlisted applicant population	18
Figure 5. AFOQT/SDI subtest FSG validity estimates, across career fields.....	19
Figure 6. AFOQT/SDI subtest subgroup difference estimates in USAF officer applicant population	20
Figure 7. Space Systems Operations Pareto frontiers for candidate PSMs and operational composites.....	27
Figure 8. Space Systems Operations candidate PSM #3 composition.....	28
Figure 9. Intelligence – All Source Analysts Pareto frontiers for candidate PSMs and operational composites	29
Figure 10. Intelligence – All Source Analysts candidate PSM #5 composition.....	29
Figure 11. Intelligence – Cryptologic Analysts Pareto frontiers for candidate PSMs and operational composites	30
Figure 12. Intelligence – Cryptologic Analysts candidate PSM #4 composition.....	31
Figure 13. Cyber Operations Pareto frontiers for candidate PSMs and operational composites.....	32
Figure 14. Cyber Operations candidate PSM #5 composition.....	32
Figure 15. Space Operations Pareto frontiers for candidate PSMs and operational composites.....	33
Figure 16. Space Operations candidate PSM #6 composition.....	34
Figure 17. Intelligence Operations Pareto frontiers for candidate PSMs and operational composites.....	35
Figure 18. Intelligence Operations candidate PSM #4 composition.....	35
Figure 19. Cyber Operations Pareto frontiers for candidate PSMs and operational composites.....	36
Figure 20. Cyber Operations candidate PSM #4 composition.....	37

Figure 21. Mean Predicted Performance Across Enlisted AFSCs on Operational and Candidate Standards.....	42
Figure 22. Mean Qualifying Rates Across Enlisted AFSCs on Operational and Candidate Standards.....	42
Figure 23. Mean Qualifying Rates by Subgroup Across Enlisted AFSCs on Operational and Candidate Standards.....	43
Figure 24. Mean Adverse Impact Ratios by Subgroup Across Enlisted AFSCs on Operational and Candidate Standards.....	44

ACKNOWLEDGEMENT

The Human Resources Research Organization (HumRRO) wishes to thank the following individuals for their technical support and valuable feedback on the evaluation documented in this report:

- Dr. Katie Gunther, Mr. Angelo Haygood, Dr. Emily Anderson, and Mr. Daniel Streeter (United States Space Force)
- Ms. Annette Rizer, Dr. Julia Walsh, and Dr. Montana Woolley (Infoscitex)
- Dr. Bobbie Dirr, Dr. John Trent, Dr. James Johnson, Mr. Matt Palomo, Dr. Ashanti Edwards, Dr. Sophie Romy, and Mr. Ken Schwartz (United States Air Force)
- Dr. Wilfried De Corte (University of Ghent)
- Dr. Laura Ford, Dr. Maura Burke, and Dr. Dan Putka (Human Resources Research Organization)

1.0 EXECUTIVE SUMMARY

The United States Space Force (USSF) is developing an innovative, 21st century talent management strategy (U.S. Space Force, 2021). The USSF, with its United States Air Force (USAF) partners, is reviewing current talent management policies and experimenting with new ones to attract, recruit, and access the talent needed for spacepower disciplines while mitigating potential barriers to service. Accessioning high-quality and diverse applicants, enlisted and officer, to serve as Guardians in the USSF is essential to this talent management strategy.

To enable this strategy, the primary objective of the current effort was to develop and evaluate new Predictive Success Models (PSMs) to qualify USSF applicants for spacepower disciplines. The PSMs were intended to optimize predicting training success while opening-up qualification to larger, more diverse applicant pools with the right talents to succeed relative to current qualification standards.

Developing and evaluating the PSMs consisted of the following activities:

1. Identifying, compiling, and preparing relevant job- and person-side data on enlisted and officer personnel for analysis. Relevant job-side data includes information on current, operational job qualification standards and important work activities performed in the seven spacepower disciplines (career fields) targeted (four for enlisted and three for officer). Relevant person-side data (scores on cognitive aptitude tests and personality assessments, demographics, technical training completion, and final school grades [FSG]) were prepared for 194,859 enlisted applicants or accessions and 46,887 officer applicants or accessions.
2. Developing and validating PSMs for each targeted spacepower discipline (career field) using the data prepared in (1). The PSMs were developed using an innovative multi-objective optimization method, Pareto optimization, that combines scores on the cognitive aptitude tests and personality assessments to optimize technical training success and the diversity of qualifying enlisted and officer applicants in combinations specific to each career field. The PSMs were subsequently evaluated relative to the current qualification standards for the targeted career fields.
3. Evaluating the projected impact of using the best PSMs from (2) to qualify or match applicants to the targeted spacepower disciplines or the specialties (jobs) comprising each discipline, benchmarked against the qualification standards in current operational use. For enlisted personnel, using PSM scores to qualify applicants for jobs to complement or replace the current operational qualification standards was evaluated. For officers, two potential uses were evaluated: (a) using PSM scores to qualify applicants to spacepower disciplines and (b) using PSM scores to match applicants to the spacepower discipline that they were expected to perform best in training.

Enlisted PSM scores (a combination of scores on selected ASVAB cognitive subtests and TAPAS facets) showed distinct improvements over current job qualification standards using

scores on the current ASVAB-based composites (Mechanical, Administrative, General, and Electronics [MAGE] composites). For some enlisted jobs, using the PSM scores yielded improvements in expected training success and opened-up qualifications to a larger, more diverse pool of applicants with the right talents to succeed. For the remaining enlisted jobs, using PSM scores for job qualification resulted in improvements in qualification rates overall and for female and minority applicants, with minimal to no loss in the expected training success of qualifying applicants. Officer PSMs (a combination of scores on selected AFOQT cognitive subtests and Self-Description Inventory [SDI] facets) showed limited improvements from using AFOQT cognitive-based scores to qualify applicants for career fields. However, improvements were observed when PSM scores matched applicants to the career fields where they were expected to perform their best in training.

Based on the findings, we make the following recommendations:

Enlisted

1. Implement one or more of the best PSM candidates identified in the projected impact evaluation for the targeted spacepower disciplines operationally, in place of or as an additional compensatory standard to the current operational MAGE standard(s).
2. Conduct a follow-up test and evaluation of the implemented PSM candidates once sufficient data accumulates within USSF.
3. Use the follow-up test and evaluation results to refine the implemented PSM candidates as needed.

Officers

1. Revisit and replicate the current analyses to identify the best PSM candidates for use in matching officers to spacepower disciplines after the proposed AFOQT refinements have been completed.
2. Implement one or more of the best PSM candidates as a tool to match officers to spacepower disciplines operationally.
3. Conduct a follow-up test and evaluation of the implemented PSM candidates once sufficient data accumulates.
4. Use the follow-up test and evaluation results to refine the implemented PSM c.0

INTRODUCTION

The USSF is the nation's newest military service responsible for cultivating, developing, and advancing spacepower to ensure national prosperity and security (U.S. Space Force, 2020). USSF's goal is to be a lean, mission-focused digital Service, a goal that requires people with unique technical and non-technical talents. The USSF recognizes that attracting, recruiting, and accessioning people with those unique talents will be challenging, especially with today's competitive market for science, technology, engineering, and mathematics (STEM) talent—a talent management challenge shared by its sister service, the USAF. Accordingly, the USSF is developing an innovative, 21st-century talent management strategy (U.S. Space Force, 2021), working with its USAF stakeholders to review current talent management policies and

experimenting with new processes to attract, recruit, and access the talent needed for spacepower disciplines while minimizing potential barriers to service.

Highlighting the importance of this objective, the Secretary of Defense (2020) made diversity and inclusion recommendations for the U.S. military in a memorandum to senior Pentagon leadership. One of the Secretary's recommendations was to "analyze, identify, and remove as applicable, barriers that adversely impact diversity while retaining rigorous screening processes necessary to access a high-quality force" (Secretary of Defense, 2020). The Human Resources Research Organization (HumRRO) conducted a program of research to develop and evaluate new Predictive Success Models (PSMs) to qualify USSF applicants for spacepower disciplines. The new PSMs were intended to optimize predicted training¹ success in combination with opening qualifications to a larger, more diverse talent pool, consistent with the Secretary's recommendation. Accordingly, we employed a multi-objective PSM development strategy that combines scores on cognitive aptitude and personality assessments to optimize training success and diversity in combination with and relative to existing qualification standards.

1.1 Background and Overview of Technical Approach

The USAF and USSF use scores on cognitive aptitude tests to screen for service qualification (i.e., access to USAF or USSF) and qualifications for specific jobs. For enlisted personnel, USAF and USSF primarily use scores on the Armed Services Vocational Aptitude Battery (ASVAB) for service and job qualification. The ASVAB is a cognitive aptitude battery consisting of nine subtests measuring domains such as verbal, math, science, and mechanical aptitudes. All U.S. military branches use scores on the Armed Forces Qualification Test (AFQT), a weighted composite of the ASVAB verbal and math subtests, to qualify applicants to serve. Enlisted applicants must meet a minimum AFQT score to qualify for access to the USAF or USSF. Regarding a job qualification, four ASVAB-based operational composites, Mechanical, Administrative, General, and Electronics (MAGE) composites, are used to qualify enlisted Airmen and Guardians for specific jobs. Different MAGE composites and standards (cut scores) are used for job qualification, depending on the required tasks and difficulty levels of specific jobs/training programs. For some Air Force Specialty Codes/Space Force Specialty Codes (AFSCs/SFSCs), there are additional composites used to qualify applicants for jobs that are customized to predict success for a given job. These composites are called PSMs. PSMs are customized by tailoring subtest weighting schemes to optimal levels to predict training or job success within a given job. PSMs generally use scores on the ASVAB subtests. Some PSMs also incorporate scores on a personality assessment called the Tailored Adaptive Personality Assessment System (TAPAS).

For officers, the USAF and USSF primarily use scores on the Air Force Officer Qualifying Test (AFOQT). Like the ASVAB, the AFOQT is a cognitive aptitude test battery comprised of 10 subtests measuring multiple aptitudes and domain-specific knowledge. The USAF and USSF use scores on the AFOQT to screen officer candidates for service who commission through the Reserve Officers' Training Corps (ROTC) or Officer Candidate School (OCS). The current service qualification standard is AFOQT-Verbal composite (AFOQT-V) ≥ 15 and AFOQT-

¹ The PSMs optimize *training* success rather than *on-the-job* success, given that only technical training data was available for model development and validation.

Quantitative composite (AFOQT-Q) ≥ 10 . Other than the rated (Aviation) career fields, there are not currently career field or job-specific AFOQT qualification standards for USAF/USSF officers. (Air Force Enlisted Classification Directory (AFECD), 2022; Air Force Officer Classification Directory (AFOCD) 2022; Carretta, 2014; Walsh et al., 2021). The SDI is a personality assessment administered as part of the AFOQT. When the SDI was added to the AFOQT in 2005 (AFOQT Form S), it was not used to qualify officers for jobs (Weissmuller & Schwartz, 2007). The Air Force recently adopted PSMs that combine AFOQT cognitive subtests and SDI facets for manned aircraft pilot, Combat Systems Officer, and Air Battle Manager training (Woolley et al., in press). These PSMs were implemented in March 2023 as alternative aptitude qualification composites to the current AFOQT cognitive-based composites.

The objective of the current effort was to develop and validate PSMs that combine scores on the aforementioned cognitive aptitude tests (ASVAB for enlisted, AFOQT for officers) and personality assessments (TAPAS for enlisted, SDI for officers) to predict training success and ensure diversity of qualified applicants. We developed and validated the PSMs in the current effort as follows: (1) preparing analysis samples and predictor-criterion data, (2) developing PSMs by career field, and (3) evaluating the projected impact of using PSMs, operationally, for job qualification on the selected set of USSF career fields.

For Task One (*prepare analysis samples and predictor-criterion data*), we identified criterion measures aligned with USAF/USSF job requirements. Subsequently, we defined analysis inclusion criteria and constructed criterion composites to reflect success in technical training. Lastly, we analyzed relationships between predictor subtests and criterion composites and estimated subgroup differences in predictors by gender and race/ethnicity. Predictor subtests included the ASVAB subtests, special tests (Cyber Test, Electronics Data Processing Test), and TAPAS facets for enlisted personnel. The AFOQT was used for officers, including the cognitive subtests and the SDI. The TAPAS is a forced-choice personality instrument designed to provide resistance to faking and is customized for use in military contexts (Stark et al., 2014). The SDI is a self-report, Likert-type personality assessment developed for officers (Weissmuller & Schwartz, 2007; Woolley et al., 2022).

For Task Two (*develop PSMs by career field*), we applied a multi-objective optimization technique, Pareto Optimization (De Corte et al., 2007), to develop and validate PSMs by career field. The PSMs were intended to improve upon the current qualification standards in predicting training success and opening up qualifications to more diverse applicants with the right talents for success. It is important to note that this optimization occurs according to the two aforementioned objectives (training success, diversity)—and a range of solutions are estimated that balance those two objectives at varying levels; there is not a single solution that maximizes both objectives. Raw PSM scores were subsequently normed to the USAF applicant population, and the predictive validity and subgroup differences of scores on the developed PSMs were evaluated.

For Task Three (*evaluate projected impact of candidate PSMs*), we identified the most promising candidate PSMs to evaluate for potential operational use in job qualification. We simulated the projected impact of the newly developed PSMs to screen and qualify enlisted and officer applicants into targeted AFSCs/SFSCs, benchmarking findings against current job qualification

standards. Finally, we identified candidate qualification standards on the candidate PSMs for each of the 16 targeted enlisted AFSCs and three targeted officer career fields.

This technical report is organized according to the task sequence outlined above, with separate sections for Tasks One, Two, and Three.

2.0 PREPARING ANALYSIS SAMPLES AND PREDICTOR-CRITERION DATA

The analysis preparation phase consisted of three relatively distinct subtasks. First, criterion measures were identified that were aligned with USAF-USSF job requirements for consideration in future research. We note that this task was unrelated to preparing and analyzing current study data; these potential criterion measures were intended to inform future job performance measurements within USSF and USAF. Second, we prepared analysis samples, screening out data that failed to meet analysis inclusion criteria, and constructed criterion composites. Third, we analyzed relationships between predictor subtests and criterion composites in the enlisted and officer analysis samples. Each of the three subtasks are outlined below.

2.1 Identifying Criterion Measures Aligned with Job Requirements

We began this subtask by identifying potential sources of job requirements and potential sources of criterion measures. After identifying available sources of job requirements on the targeted jobs, the AFECDD/AFOCD “Duties and Responsibilities” sections were selected as the source of job requirements.² The AFECDD/AFOCD duties and responsibilities were in a standard format across AFSCs, were available for all targeted AFSCs, and were at a sufficient level of detail to allow for cross walking of requirements onto criterion dimensions.

The primary goal of identifying a source of criterion measures was usefulness in classification since the primary intended application of the proposed PSMs was job qualification or person-job matching. Therefore, we focused on identifying criteria that would show varying levels of importance across USAF/USSF AFSCs/SFSCs and career fields, thereby differentiating applicants’ performance potential between career fields. Career field- or job-specific performance behaviors were viewed as preferable to general criterion dimensions for classification purposes, as career- or job-specific performance behaviors are more likely to differentiate among jobs. For example, the broad dimension of technical task performance (Campbell, 2012) is a well-established dimension of performance. However, a “general technical task performance” dimension is unlikely to be useful in classification because this dimension is important for all jobs³, unless one specifies the constituent behaviors that differentiate across jobs.

Barelka et al. (2019) outline 27 competencies that comprise the Air Force Foundational Competency Model. These competencies represent a “measurable set of knowledge, skills,

² USSF versions of the enlisted and officer classification directories were not produced at the time of this project. Therefore, we relied on the USAF versions because the requirements and technical training pipelines were expected to remain the same.

³ We acknowledge that specific individual performance behaviors indicative of technical task performance will vary across jobs. However, a general technical task performance dimension is likely to be similarly important across jobs.

abilities, and other characteristics (KSAOs) that manifest in observable behaviors needed to perform successfully” (Barelka et al., 2019, p. 1) and apply to all USAF Airmen. The competencies are relatively broad, with examples including Accountability, Achievement Orientation, Communication, and Analytic Thinking. Upon review of these competencies, we decided a finer-grained taxonomy of performance components was preferable, with the rationale being that more specific components will better differentiate across similar jobs in the KSAOs needed to perform effectively. After further review of available measures, we ultimately selected the Air Force Work Interest Navigator (AF-WIN) Work Activities as the recommended criterion taxonomy. The Work Activities were selected because they represent a wide range of specific and measurable performance behaviors that systematically vary across jobs. Furthermore, the Work Activities have been successfully used in prior research to profile enlisted USAF career fields (Johnson et al., 2020). The associated AF-WIN Functional Communities or Job Contexts were not selected because these components did not represent performance behaviors.

To determine the alignment of job requirements with the Work Activities, we conducted a crosswalk task, rating the degree of alignment of each AFSC’s job tasks with each work activity. Tasks were rated as 0 (*Task is not related to performance component*; “Not related”), 1 (*Task may be related to performance component*; “Potentially related”), or 2 (*Task is related to performance component*; “Related”). One researcher rated all AFSCs across enlisted and officer career fields. A small number of ratings (< 2%) were flagged for review if there was uncertainty of a job task-to-work activity relationship. In these instances, a second researcher rated the pairing, then both raters discussed and resolved discrepancies. Overall, we found that the AF-WIN Work Activities were aligned with job requirements for both enlisted and officer AFSCs. Furthermore, the Work Activities captured differences in performance requirements across jobs. However, we did suggest refinements and additions to the Work Activities to tailor measurement for spacepower disciplines, which we outline below. Attachment A contains histograms for enlisted personnel and officers presenting the number of Work Activities rated as “Related” across AFSCs.

Subsequently, USAF subject matter expert (SME) data was used to assess the convergence of our crosswalk ratings. USAF Career Field Managers (CFMs) distributed the AF-Win Work Activities to enlisted Airmen and officers who were identified as SMEs. SMEs then rated the applicability of the Work Activities for specific AFSCs (Johnson et al., 2020). The SME data was used as a convergence check rather than for primary examination because it was reported at the AFSC level, whereas our crosswalk ratings were made at the task level (i.e., tasks within AFSCs). To quantify convergence, we calculated average crosswalk ratings across all tasks within a given AFSC. In essence, each source of data (crosswalk ratings, SME data) had Work Activity profiles by AFSC that we correlated to assess the stability of our crosswalk ratings. For enlisted, our task-averaged ratings correlated $r = .79$ with SME ratings across AFSCs, indicating convergent validity for our task ratings. For discriminant validity, mismatched AFSCs (specifically, AFSCs outside of the AFSC in the questions’ career field) had task-average ratings that correlated $r = .67$ with SME ratings, on average. For Officer, our task-averaged ratings correlated $r = .73$ with SME ratings across AFSCs, whereas mismatched AFSCs correlated $r = .67$. While convergent validity estimates were higher than discriminant validity estimates, the discriminant validity correlations were still relatively high—which is at least partially reflective of the similarity in job profiles across career fields (i.e., jobs generally heavy in data, analysis,

and documentation). Overall, we view the crosswalk exercise as displaying reasonably high convergence with SME ratings of the same jobs.

Crosswalk ratings indicated that the sampled USSF SFSCs tended to be heavy on data, technology, and analysis-related activities. Patterns were similar across enlisted and officer crosswalk ratings. We found these results to be intuitive with respect to the job tasks summarized in the AFECD and AFOCD. Figures 1 and 2 show the most common work activities for enlisted and officer AFSCs.

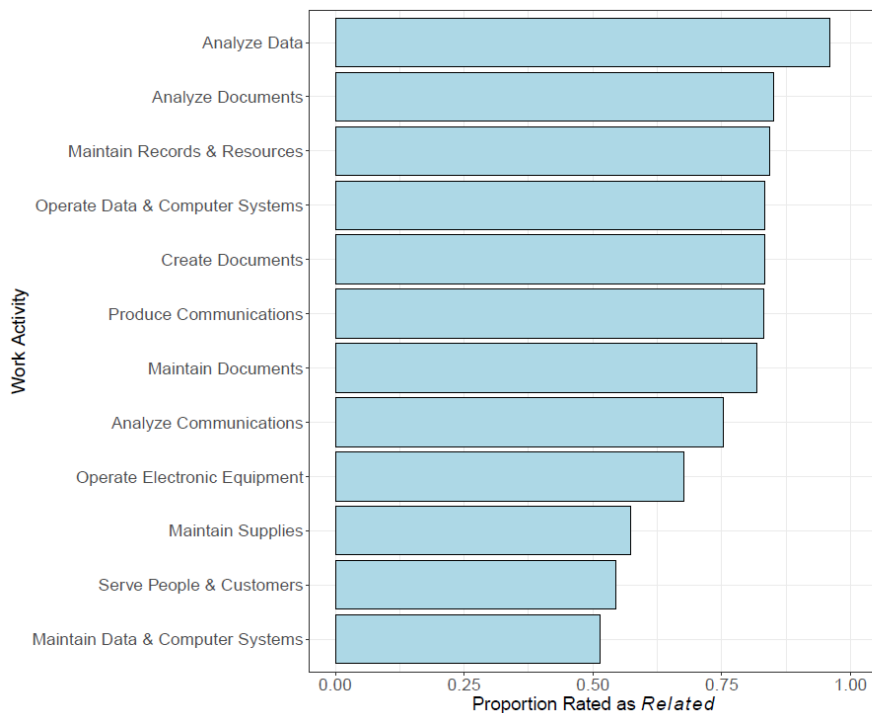


Figure 1. Most common work activities for enlisted AFSCs

To assess how well the breadth of all Work Activities captured job requirements, we examined how many Work Activities were rated as related to at least one task for each AFSC. In other words, how many work activities were aligned with each AFSC? A low number of tasks rated as Related for a given AFSC may indicate that the Work Activity taxonomy did not sufficiently capture job requirements. However, crosswalk ratings suggested the Work Activities sufficiently captured job requirements. No AFSC had fewer than 13 AF-WIN Work Activities rated as Related to at least one task. For enlisted Airmen, an average of 17 Work Activities (out of 28 total; Standard Deviation (*SD*) = 3) were rated as Related across all job tasks in an AFSC. For officers, an average of 24 Work Activities (out of 38 total; *SD* = 4) were rated as Related across all job tasks in an AFSC. These results indicated that job requirements for the sampled USSF SFSCs are aligned with many Work Activities across AFSCs.

Finally, after reviewing the AF-WIN Work Activity taxonomy concerning the spacepower disciplines, we identified recommended additions and potential refinements to the Work

Activities. These recommendations were focused on tailoring performance measurement to the context of USSF job requirements. The following additional work activities were identified as highly relevant and were recommended to be added to the taxonomy: (a) Work in a Team, (b) Plan Projects and Operations, (c) Manage Projects and Operations, (d) Develop Software (or Develop Code), and (e) Maintain Cybersecurity. Recommendations are found in Attachment A.

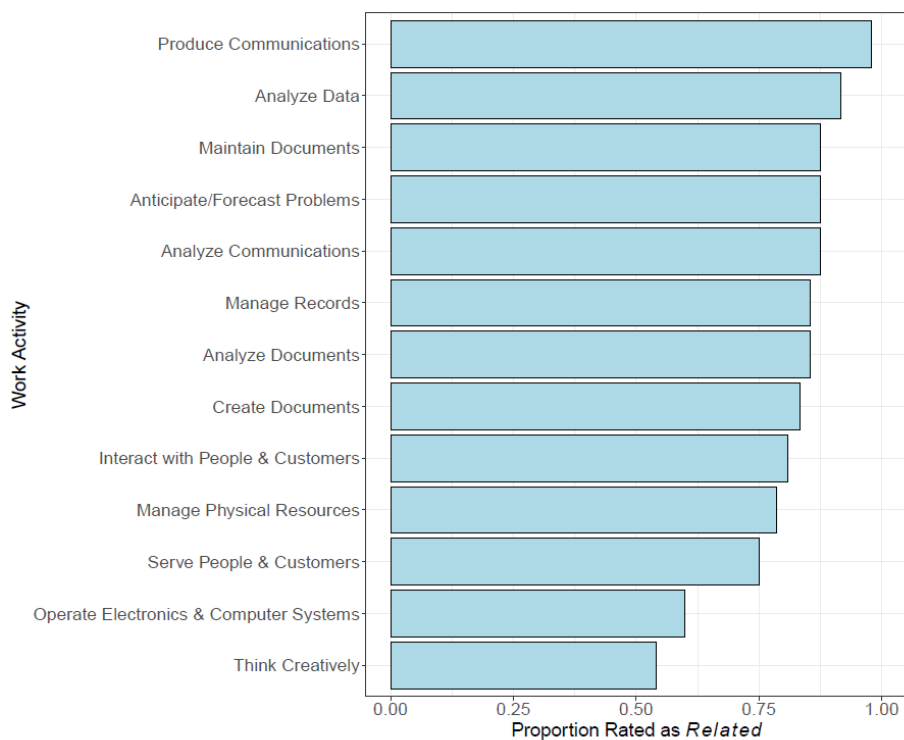


Figure 2. Most common work activities for officer AFSCs

2.2 Preparation of Analysis Samples and Construction of Criterion Composites

There were four sources of data provided by the USAF for this study: (a) enlisted trainee, (b) officer trainee, (c) enlisted applicant, and (d) officer applicant. The trainee datasets consisted of both pre-accession data (enlisted: ASVAB, TAPAS; officer: AFOQT, SDI) and post-accession data (e.g., technical training course pass/fail, washbacks, FSGs) on USAF active-duty members. The trainee datasets contained data on 16 enlisted AFSCs (comprising four career fields) and four officer AFSCs (comprising three career fields). The targeted enlisted and officer AFSCs are listed in Table 1, along with corresponding SFSC identifiers and analysis sample sizes. The applicant datasets consisted of USAF enlisted and officer applicants from 2015 to 2021 and contained pre-accession data, including demographic and qualification testing data. Attachment B provides documentation of the enlisted and officer AFSCs, their career field mappings, and current entry standards.

Table 1. Enlisted and officer targeted AFSCs/SFSCs and sample sizes

Enlisted AFSCs/SFSCs		
Job Title (<i>Career Field</i>)	AFSC / SFSC Identifier	Analysis <i>n</i>
<i>Space Systems Operations</i>		238
Space Systems Operations	1C6X1 / 5S0X1	238
<i>Intelligence-All Source Analysts¹</i>	5I0X1	1,199
All Source Intelligence	1N0X1 / 5I0X1	647
Geospatial Intelligence	1N1X1 / 5I1X1	436
Targeting Analyst	1N8X1 / 5I8X1	116
<i>Intelligence-Cryptologic Analysts¹</i>	5I2X1	2,162
Signals Intelligence Analyst – Electronic	1N2X1A / 5I2X1	378
Signals Intelligence Analyst – Communications	1N2X1C / 5I2X1	407
Intelligence Analyst – Cyber Intelligence Analysis	1N4X1A / 5I4X1	557
Intelligence Analyst – Cryptologic Analysis & Reporting	1N4X1B / 5I4X1	820
<i>Cyber Operations²</i>		4,786
Knowledge Management	3D0X1 / 5C0X1	376
Cyber Systems Operations	3D0X2 / 5C0X1	2,094
Cyber Surety	3D0X3 / 5C0X1	762
Computer Systems Programming	3D0X4 / 5C0X1	217
Client Systems	3D1X1 / 5C0X1	401
Cyber Transport Systems	3D1X2 / 5C0X1	450
RF Transmission Systems	3D1X3 / 5C0X1	333
Cable and Antenna Systems	3D1X7 / 5C0X1	153
Officer AFSCs/SFSCs		
<i>Space Operations</i>		
Space Operations	13SX	204
<i>Intelligence Operations</i>		
Intelligence	14NX	394
<i>Cyber Operations</i>		
Warfighter Communications Operations	17DX	407

Notes. ¹ Since this work was conducted, USSF has merged all SFSCs falling under Intelligence-All Source Analysts into a single SFSC (5I0X1). Similarly, USSF has merged all SFSCs falling under Intelligence-Cryptologic Analysts into a single SFSC (5I2X1). ² Since this work was conducted, USSF has merged all SFSCs falling under Cyber Operations into a single SFSC (5C0X1).

2.2.1 Preparation of Analysis Samples

We used predetermined screening criteria to define the enlisted and officer analysis samples of interest, removing records that were not aligned with analysis goals or records that were anomalous. First, we only included applicants and trainees with no indicator of prior military service. Prior service applicants typically constitute a minority of military applicants and represent a different population than non-prior service applicants (prior service applicants are typically older with more work experience, particularly military experience, than non-prior services applicant). Excluding prior service applicants enables a cleaner evaluation of the use of scores on the ASVAB and other predictor measures to make accessioning or job qualification decisions, as well as more appropriate inferences about the value of predictor scores as they apply to the primary population of U.S. military applicants. Second, we only included data on applicants processed for the U.S. Air Force Regular, Guard, or Reserve, excluding records processed for the U.S. Navy, U.S. Army, or U.S. Marine Corps. Subsequently, we used a series of rules to screen the predictor subtest data to valid records. First, we only included records with non-missing cognitive (ASVAB/AFOQT)⁴ and non-cognitive (TAPAS/SDI) predictor data. For the enlisted samples, we only included records with data on the most recent TAPAS versions, version USAF 5.0 or 5.1.⁵ We restricted to current TAPAS versions because the goal was to develop qualification standards using the TAPAS. Second, we only included enlisted records with ASVAB scores normed to the current Profile of American Youth (PAY) 97 (Segall, 2004) norming sample, removing cases that were normed (or were flagged as suspected to be normed) using the now-dated PAY80 sample. Third, when records had multiple testing attempts, we selected the most recent record (i.e., the score of record) for cognitive tests, whereas we selected the first testing attempt for non-cognitive tests (TAPAS, SDI). We selected the most recent cognitive record because faking was not a concern on the cognitive tests. Therefore, the most recent score reflected the applicant or trainees' most current aptitude levels. We selected the first-attempt non-cognitive scores because faking is a concern for the personality inventories, as there are not clear correct/incorrect answers on these inventories. Applicants that did not qualify for service on their first testing attempt are likely to show response distortion on the TAPAS on subsequent attempts (Roberts et al., 2019). Finally, all records with a testing date prior to 17-August-2016 were excluded for officers due to an SDI scoring issue that affected scores before this date.

Finally, for the enlisted and officer trainee samples, we further restricted analysis samples based on training criterion data. First, we only included a trainee's first technical training attempt, excluding trainees that attempted technical training in a second AFSC. The rationale for this decision was to tailor qualification models to identify trainees that will be successful during their first attempt at technical training. Second, we restricted analysis samples to trainees that had at least one criterion composite measure available. Criterion composites are discussed in the next section. Third, we excluded cases with an ASVAB testing date greater than awarding course completion date, which signaled a potential issue with a record.

⁴ For enlisted samples, we allowed missingness on the Cyber Test, the Electronic Data Processing Test (EDPT), and Assembling Objects (AO), as not all applicants complete these tests.

⁵ We focused analyses and model development on the 13 core TAPAS facets shared between TAPAS version USAF 5.0 and 5.1, along with the two new facets from version 5.1 (Responsibility and Situational Awareness).

Applying the screening rules outlined above resulted in analysis samples of the following sizes: enlisted applicant ($N = 194,859$), enlisted trainee ($N = 8,385$), officer applicant ($N = 46,887$), and officer trainee ($N = 1,005$). All database preparation and subsequent analyses were conducted in the *R* statistical software.

2.2.2 Construction of Criterion Composites

We selected two focal technical training criteria for analysis: (a) technical training pass-fail and (b) FSG. We constructed a first-pass pipeline success (FPPS) criterion measure using the technical training pass-fail data. FPPS measured whether a USAF trainee graduated from the initial technical training pipeline on their first attempt without experiencing a washback or failure (i.e., initial technical training pipeline leading to an awarding AFSC and skill level 3). The FPPS criterion was constructed such that trainees experiencing an academic washback or failure were coded as 0, and trainees who graduated without experiencing a washback failure were coded as 1. Importantly, technical training across targeted AFSCs ranged from one to three courses, and FPPS values encompassed all courses (i.e., the complete training “pipeline”). Therefore, if a trainee experienced a washback or failure in any technical training course within their AFSC, they were coded as a 0, even if they ultimately passed technical training. This decision rule allowed for the identification of trainees that experienced any training issues during their training pipeline. Mean FPPS rates for enlisted trainees varied across AFSCs from 41.9% to 97.9% ($M = 78.2\%$, $SD = 16.4\%$). Mean FPPS rates for the three officer career fields (comprising one AFSC each) were: 86.8% (Cyber Operations), 94.6% (Intelligence Operations), and 100% (Space Operations).

The second criterion targeted was FSG from a trainee’s AFSC awarding course. FSG is the percent grade for the final (awarding) course in a trainee’s technical training sequence. Awarding course was selected because trainees with a passing FSG from their awarding course were successful in their technical training and because the awarding course displayed the most variability in FSG in relation to the initial and intermediate courses, where censoring of non-passing FSGs was greater. Valid FSG values ranged from 70 to 100. Mean FSG rates for enlisted trainees varied across AFSCs from 85.8% to 95.0% ($M = 89.9\%$, $SD = 3.2\%$). Mean FSG rates for the three officer career fields (comprising one AFSC each) were: 92.0% (Cyber Operations), 94.9% (Intelligence Operations), and 96.4% (Space Operations).

While predictive validity and subgroup differences were estimated on both criteria, we needed to select a single criterion for which to develop qualification models. We made different criterion decisions for enlisted and officer because of differing data characteristics for each sample. For enlisted, we selected FPPS as the primary criterion over FSG. First and foremost, FPPS was selected over FSG for enlisted personnel because of data limitations on FSG. The sample available to model FPPS was $N = 8,271$, compared to $N = 4,373$ for FSG. The reduction in N for FSG occurred because there were five AFSCs either missing FSG or with FSG scores almost all out of the expected range. Therefore, modeling on FPPS allowed for developing models for all 16 AFSCs (rather than 11 for FSG) and avoided deleting almost half of the enlisted analysis data. Secondly, FPPS summarizes information across multiple technical training courses, whereas trainees that failed an earlier training course were systematically missing from FSG for the final awarding course. The officer data did not have the same FSG data missingness issue for two

reasons. First, virtually the same number of officers had both FPPS and FSG scores available. Second, two of the three officer career fields only contained one training course, and the preliminary course in the third career field displayed a very high pass rate ($> 95\%$). Therefore, there were not the same systematic missingness issues in FSG for officer data as there were for enlisted data. Furthermore, the FPPS data for officers had very little variability because almost all officer trainees passed technical training. Therefore, for officers, we decided to model based on FSG.⁶

2.3 Analyzing Relationships Between Predictors and Criteria

The primary goal of this phase of analyses was to analyze the predictive validity and adverse impact potential of all predictor subtests. Table 2 outlines all predictor subtests examined for enlisted and officer analysis samples. Note that the names of the TAPAS and SDI facets are not presented because that information is restricted. A restricted distribution version of this report contains TAPAS and SDI subtest-level information and results. Next, we describe the methodology used to estimate applicant population-level predictive validity and subgroup differences, followed by enlisted and officer subtest results.

2.3.1 Validity and Subgroup Difference Estimation Approach

Population-level predictive validity for applicants was estimated in the enlisted and officer trainee samples by examining the correlations between predictor subtest scores, FSG (Pearson correlations), and FPPS (point-biserial correlations), corrected for multivariate range restriction (MVRR) on the predictor subtests since scores on the subtests were used for service screening and job qualification. Predictive validities and intercorrelations were estimated within each AFSC and then aggregated to the career field level via sample size weighting.

The observed (restricted) predictor-predictor and predictor-criterion covariance matrices among the trainee samples were estimated for each AFSC using Full Information Maximum Likelihood (FIML) estimation. We used FIML estimation because it makes use of all available data, whereas listwise deletion removes all cases with any missing values. Furthermore, FIML is generally less biased than listwise deletion under different missing data mechanisms (Newman, 2014). For seven of 16 enlisted trainee samples, a high rate of missingness on the two TAPAS facets new to version 5.1 (Responsibility, Situational Awareness) caused convergence issues when estimating FIML covariances. For these AFSCs, we used multiple imputations to estimate the covariance matrices. Specifically, within each of the seven AFSCs, we imputed missing values to form a complete dataset and calculated covariance matrices for the observed-imputed datasets. We repeated this imputation procedure 50 times per AFSC, averaging across covariance matrices to form a final predictor-criterion covariance matrix per AFSC. These covariance matrices were used in range restriction corrections. FIML was used as the preferred choice because FIML directly estimates covariances and does not involve imputing individual-level data. However, both FIML and multiple imputations show comparable performance under data missingness and are both less biased than listwise deletion under different conditions

⁶ We also used FSG for one enlisted AFSC (1C6X1/5S0X1) due to insufficient variability in FPPS.

Table 2. Enlisted and Officer Predictor Subtests/Facets

Subtest/Facet Name	Percent Missing – Trainee Sample	Percent Missing – Applicant Sample
<i>Enlisted Predictor Subtests/Facets</i>		
<i>ASVAB Subtests</i> (Plus Cyber and Electronic Data Processing Test)		
Assembling Objects (AO)	9.2%	7.0%
Arithmetic Reasoning (AR)	0.0%	0.0%
Auto-Shop (AS)	0.0%	0.0%
Electronics Information (EI)	0.0%	0.0%
General Science (GS)	0.0%	0.0%
Mechanical Comprehension (MC)	0.0%	0.0%
Math Knowledge (MK)	0.0%	0.0%
Paragraph Comprehension (PC)	0.0%	0.0%
Word Knowledge (WK)	0.0%	0.0%
Cyber Test (CT)	25.9%	27.9%
Electronic Data Processing Test (EDPT)	92.4% ¹	98.6% ¹
<i>TAPAS Facets</i> ²		
TAPAS 01 - 15	0.0% (13 core facets) 2.1% (2 facets new to v5.1)	0.0% (13 core facets) 83.1% for (2 new facets)
<i>Officer Predictor Subtests/Facets</i>		
<i>AFOQT Subtests</i> ²		
Aviation Information (AI)	0.0%	0.0%
Arithmetic Reasoning (AR)	0.0%	0.0%
Block Counting (BC)	0.0%	0.0%
Instrument Comprehension (IC)	0.0%	0.0%
Math Knowledge (MK)	0.0%	0.0%
Physical Science (PS)	0.0%	0.0%
Reading Comprehension (RC)	0.0%	0.0%
Table Reading (TR)	0.0%	0.0%
Verbal Analogies (VA)	0.0%	0.0%
Word Knowledge (WK)	0.0%	0.0%
<i>SDI Facets</i>		
SDI 01 – SDI 30	0.0%	0.0%

Notes. ¹ EDPT has a high missingness rate because scores on the test are used operationally for only one AFSC within the Cyber Operations career field. ² TAPAS and SDI facet names are not presented in the main body of the report because that information is restricted.

(Newman, 2014). Both FIML and multiple imputations relied on the assumption that data were Missing at Random (MAR), meaning that observed variables were used to predict missing data values. The officer trainee sample was estimated with FIML only, as there were no variables with high rates of missingness that caused estimation issues.

Scores on the predictor measures for the enlisted or officer trainee samples were impacted by direct range restriction in two ways. First, the scores were restricted because enlisted and officer applicants needed to meet minimum scores on AFQT (enlisted) and AFOQT (officer) to qualify to access the service. Second, the scores for enlisted trainee samples were further impacted by direct range restriction at the career- or job-level because accessing applicants needed to meet minimum scores to qualify for an AFSC (based on AFSC-specific MAGE and PSM-based composites). Unlike the enlisted trainee samples, scores for the officer trainee samples were not similarly impacted by this same source of direct range restriction because none of the officer career fields targeted here used additional scores on the AFOQT to qualify officer accessions to career fields. We expected that scores among the enlisted and officer trainee samples were also impacted by indirect range restriction, in addition to the aforementioned sources of direct range restriction. The primary source of indirect range restriction was applicants or accessions self-selecting to career fields or jobs that matched their aptitudes and personality profile (i.e., the variability in scores on aptitudes or personality facets that were not important to a career field or job were similarly lower because higher scoring accessions gravitated to alternative career fields or jobs where the aptitudes or personality facets were important). In summary, there was direct range restriction among scores on the AFQT, AFOQT, and MAGE (within specific AFSCs), and indirect range restriction among scores on the ASVAB, TAPAS, and SDI subtests. Accordingly, the observed (restricted) predictor-criterion covariance matrices were corrected using the Lawley MVRR correction procedure (Aitken, 1934; Lawley, 1944) to the USAF applicant population. Lawley's MVRR correction accounts for both direct and indirect range restriction. Applying this correction resulted in a corrected correlation matrix for each AFSC that provided USAF population-level correlations between the ASVAB/AFOQT and TAPAS/SDI subtests as well as the FPPS and FSG criteria.⁷ We aggregated correlation estimates to the career field level by sample size (n) weighting individual AFSC correlation matrices. No aggregation was needed for officers because each officer career field in our sample corresponded to only one AFSC. As a benchmark for subtest-level predictor validities, we also computed predictive validity estimates of MAGE and AFOQT operational composites. To accomplish this, we applied subtest-to-composite weight matrices to the corrected covariance matrices via matrix algebra to generate composite-level corrected validities for the FPPS and FSG criteria.

Adverse impact potential was quantified using subgroup score differences on the predictor subtests. Specifically, we evaluated adverse impact potential by examining standardized group mean score differences in ASVAB/AFOQT and TAPAS/SDI subtest scores. Group mean score

⁷ For enlisted personnel, we also considered a second stage of MVRR corrections due to the possibility that there may be selection effects related to which trainees had data available on four scores with higher rates of missingness (TAPAS-Responsibility, TAPAS-Situational Awareness, Cyber Test, and EDPT). As such, the subset of trainees with data on these scores may be further range-restricted than the full AFSC samples. Therefore, we tested a second stage of RR correction, correcting the further-restricted AFSC samples out to the full AFSC samples. We found U-ratios were $\geq .98$ in these samples, and correlation estimates differed by $r = .00$. Thus, there was very little range restriction in this sample, and the relatively trivial restriction was likely accounted for via FIML/Multiple Imputation (MI) estimation. Therefore, we did not incorporate this second stage of MVRR corrections in final estimates.

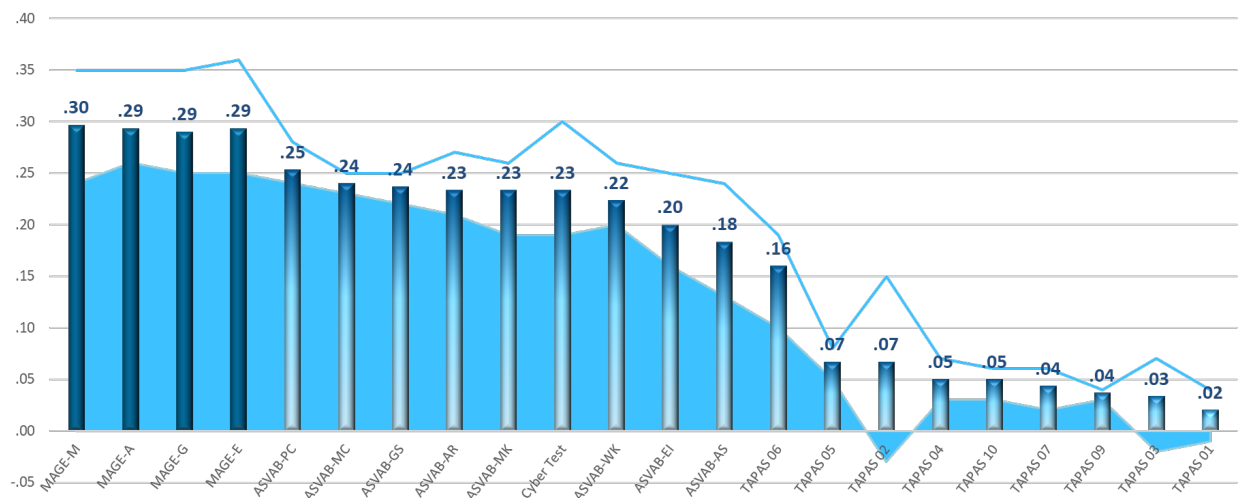
differences in predictor subtest scores were estimated by (a) gender (female-male), (b) race (Black, Non-Hispanic—White, Non-Hispanic; Asian, Non-Hispanic—White, Non-Hispanic; and Multiracial, Non-Hispanic—White, Non-Hispanic), and (c) ethnicity (White, Hispanic—White, Non-Hispanic). Standardized group mean score differences (d 's) in ASVAB/AFOQT and TAPAS/SDI subtest scores were computed using Formula (1) below, within the observed (restricted) trainee samples, as well as the USAF applicant population (unrestricted). Positive d values indicate that the minority group scored higher on the selected subtest, on average than the majority group, while negative d values indicate that the minority group scored lower, on average than the majority group. Consistent with our examination of the subtests' predictive validities, we focused our evaluation on the population-level group score differences.

$$d = \frac{\bar{X}_{Minority} - \bar{X}_{Majority}}{\sqrt{([\sigma_{Minority}^2 * n_{Minority}] + [\sigma_{Majority}^2 * n_{Majority}]) / (n_{Minority} + n_{Majority} - 1)}}$$

2.3.2 Enlisted Validity and Subgroup Difference Results

Figure 3 presents the MAGE composite and ASVAB/TAPAS subtest predictive validity estimates for the USAF applicant population on FPPS across the three career fields with sufficient variability to model FPPS (Cyber Operations, Intelligence-All Source Analysts, and Intelligence-Cryptologic Analysts). Individual ASVAB subtests generally predicted technical training success more strongly than did individual TAPAS facets. The ASVAB subtests that were the most predictive of technical training success were: (a) Paragraph Comprehension ($\rho = .25$), (b) Mechanical Comprehension ($\rho = .24$), and (c) General Science ($\rho = .24$). The most predictive TAPAS facets were: (a) TAPAS 06 ($\rho = .16$), (b) TAPAS 05 ($\rho = .07$), and (c) TAPAS 02 ($\rho = .07$).

Critically, there was variability in the determinants of technical training success by career field, suggesting that career field-specific qualification models may be useful in differentiating applicants' predicted training performance across jobs. The most predictive subtests within a career field displayed intuitive, rational linkages to the job tasks for constituent AFSCs, as listed in the AFECD. For example, the Cyber Test emerged as the best predictor of training success for the Cyber Operations career field ($\rho = .30$), which consists of proportionally more cyber-related job tasks (e.g., maintaining networks and computer systems, developing software, cybersecurity risk management) than the other enlisted career fields. Similarly, ASVAB-Math Knowledge ($\rho = .25$) and ASVAB-Paragraph Comprehension ($\rho = .24$) were the best predictors of training success in the Intelligence-Cryptologic Analysts career field. Tasks in this career field are focused on cryptologic and signals intelligence analysis, which utilize math and reading skills.



Notes. $N = 8,147$ total enlisted trainees across three career fields (Cyber Operations, Intelligence-All Source Analysts, Intelligence-Cryptologic Analysts). Middle estimates = Mean validity. Top estimates = Maximum validity. Bottom estimates = Minimum validity. All FPPS validity estimates were corrected for multivariate range restriction. Only the top nine TAPAS facet validities are presented due to space considerations.

Figure 3. ASVAB/TAPAS subtest FPPS validity estimates across career fields

The top three ASVAB subtest predictors of FSG were almost identical across all four career fields—(a) Math Knowledge, (b) Paragraph Comprehension, and (c) Arithmetic Reasoning. Although these three subtests were still among the top predictors of FPPS, the prediction profiles were more differentiated by career field for FPPS. This finding suggests that FPPS may capture a broader range of performance behaviors than FSG, including non-academic performance behaviors that predict persistence through the initial training pipeline. Persistence behaviors may be better captured by FPPS because this measure includes training information from multiple courses (when applicable), whereas FSG data is missing on trainees that did not complete a course or dropped out in a preliminary course.

Figure 4 presents a summary of subgroup difference estimates by gender and race-ethnicity in the USAF applicant population for MAGE composites, ASVAB subtests, and top TAPAS facets. Overall, MAGE composites and ASVAB subtests displayed greater subgroup differences than TAPAS facets, with TAPAS facets displaying minimal subgroup differences. Despite the TAPAS facets' relatively lower predictive validity estimates, our expectation was that including TAPAS in job qualification models (i.e., PSMs) would provide incremental validity beyond operational qualification standards in developing models that increased predicted job/training success while reducing adverse impact potential. This was our expectation because (a) previous research shows that TAPAS displays incremental validity over the ASVAB in predicting military job performance (Kirkendall et al., 2020; Nye et al., 2012), (b) TAPAS facets had positive validities on FPPS with TAPAS profiles varying by career field and likely to predict distinct variance from the ASVAB (i.e., non-academic performance behaviors), and (c) TAPAS facets displayed markedly lower subgroup differences than ASVAB subtests. Furthermore, the pattern

of subgroup differences was quite similar across the different subgroups evaluated, indicating that the use of a subtest that reduces adverse impact for females or a racial-ethnic minority group is also likely to reduce adverse impact for other minority groups.



Notes. Positive d -values indicate higher female/minority group scores. All estimates are from the USAF applicant population. Male $n = 140,076$. Female $n = 54,783$. White $n = 106,811$. Black $n = 36,569$. Hispanic $n = 30,791$.

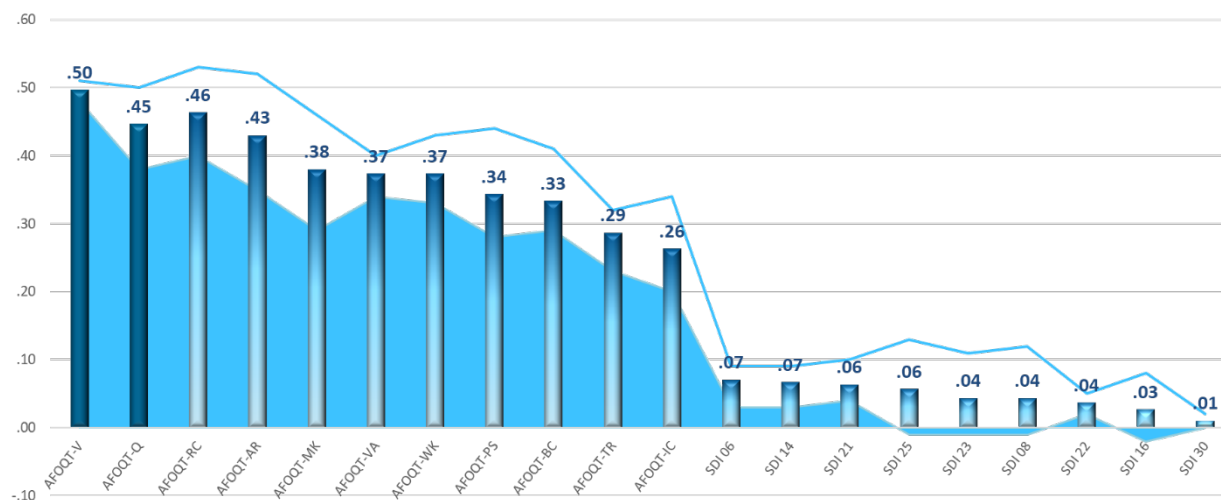
Figure 4. ASVAB/TAPAS subtest subgroup difference estimates in USAF enlisted applicant population

2.3.3 Officer Validity and Subgroup Difference Results

Figure 5 presents the AFOQT composite and AFOQT/SDI subtest predictive validity estimates for officer applicants on FSG across the three targeted officer career fields. We note that the officer predictive validity results focus on FSG because extremely high training pass rates precluded modeling FPPS empirically. Consistent with enlisted results, scores on the cognitive subtests predicted technical training performance (FSG, for officers) more strongly than scores on the non-cognitive subtests, although several SDI facets did display significant positive validities. The strongest overall AFOQT subtest-level predictors of FSG were Reading Comprehension ($\rho = .46$) and Arithmetic Reasoning ($\rho = .43$), which displayed validities comparable to or approaching the AFOQT composites of Verbal ($\rho = .50$) and Quantitative ($\rho = .45$).

The best cognitive predictors of FSG were relatively consistent across career fields. For Cyber Operations, the best predictors of FSG were the AFOQT subtests of Reading Comprehension, Arithmetic Reasoning, and Verbal Analogies (ρ s = .40, .35, .34). For Intelligence Operations, the best predictors of FSG were Reading Comprehension, Word Knowledge, and Arithmetic Reasoning (ρ s = .46, .43, .42). For Space Operations, the best predictors of FSG were Reading Comprehension, Arithmetic Reasoning, and Math Knowledge (ρ s = .53, .52, .46). For the non-

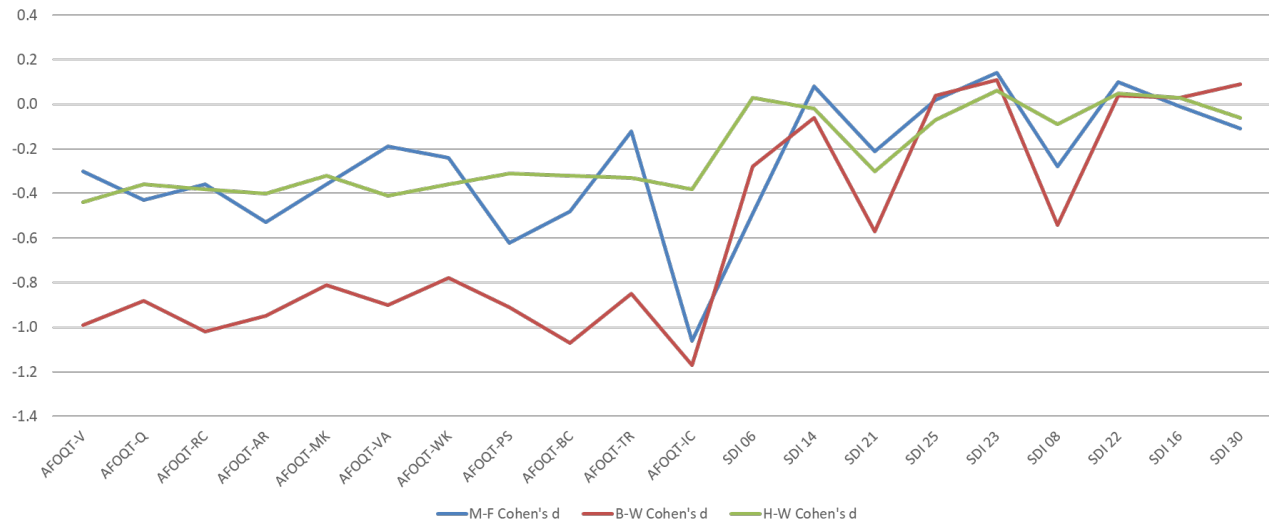
cognitive SDI facets, there was much more differentiation across career fields in rank-ordering of predictors. For example, the best non-cognitive predictors of FSG for Space Operations were SDI 25, SDI 08, and SDI 21 (ρ s = .13, .12, .10), whereas the best non-cognitive predictors for Intelligence Operations were SDI 23, SDI 06, and SDI 14 (ρ s = .11, .09, .08); with no top predictors shared across career fields. We note that there were also negatively valenced SDI facets (i.e., lower levels of trait = more desirable) from the Machiavellianism domain, which predicted FSG negatively. However, we chose not to use the negatively valenced facets in PSMs, which we discuss in section 3.1.



Notes. AFOQT-V = AFOQT-Verbal Composite. AFOQT-Q = AFOQT-Quantitative Composite. $N = 1,005$ total officer trainees across three career fields (Cyber Operations, Intelligence Operations, Space Operations). Middle estimates = Mean validity. Top estimates = Maximum validity. Bottom estimates = Minimum validity. All FSG validity estimates were corrected for multivariate range restriction. Only the top nine SDI facet validities are presented due to space considerations.

Figure 5. AFOQT/SDI subtest FSG validity estimates, across career fields

Figure 6 below summarizes subgroup difference estimates among officer applicants by gender and race-ethnicity for AFOQT composites, AFOQT subtests, and SDI facets. As with enlisted, scores on the AFOQT cognitive composites and subtests displayed substantially larger subgroup differences than the non-cognitive SDI facets, with SDI facets displaying minimal subgroup differences. We held the same expectation as enlisted about the incremental contribution of the SDI facets in job qualification models, given our hypothesis that the SDI predicts distinct variance in training success relative to the AFOQT cognitive subtests and given minimal subgroup differences on the SDI facets. One caveat to this expectation is that the FSG criterion used for officers is a more cognitively-loaded criterion than FPPS.



Notes. AFOQT-V = AFOQT-Verbal Composite. AFOQT-Q = AFOQT-Quantitative Composite. Positive d -values indicate higher female/minority group scores. All estimates are from the USAF applicant population. Male $n = 34,427$. Female $n = 12,440$. White $n = 28,824$. Black $n = 4,561$ Hispanic $n = 4,511$.

Figure 6. AFOQT/SDI subtest subgroup difference estimates in USAF officer applicant population

3.0 DEVELOPING PREDICTIVE SUCCESS MODELS (PSMs) BY CAREER FIELD

Next, we developed and validated multi-objective PSMs tailored to enlisted and officer career fields. The goal of PSM development was to potentially implement these models for use in job qualification to supplement or replace the operational MAGE (enlisted) and AFOQT (officer) qualification standards. The PSMs were keyed to predict technical training success. Accordingly, high scores on the estimated PSMs reflect an increased probability of completing technical training or performing successfully in technical training. We note that PSMs can also be interpreted to predict technical training failure, such that low scores on the subtests of interest predicts a greater likelihood of failure in technical training.

3.1 PSM Modeling Methodology

The PSMs we developed were intended to optimize predicted training success and the diversity of qualified applicants in combination and relative to current qualification standards, in line with the Secretary of Defense's (2020) recommendations on Department of Defense (DoD) diversity and inclusion. Accordingly, we used a multi-objective optimization technique called Pareto optimization (De Corte et al., 2007) to estimate models that consider two important organizational objectives: (a) predicting training success and (b) minimizing adverse impact potential by gender and race-ethnicity (i.e., expanding access to qualified and diverse candidates). A range of solutions were estimated for each career field that balanced the two objectives of predicting training success versus minimizing adverse impact at varying levels. The estimated multi-objective PSMs were subsequently evaluated relative to the current qualification standards. Pareto optimization differentially weights scores on two or more predictors and can be used to form composite scores—in this case, PSMs. Pareto optimization is advantageous because the method identifies weighted predictor score combinations that optimize two objectives (criteria) simultaneously, in contrast to standard unit-weighting or regression-weighting of predictor scores keyed to a single criterion (De Corte et al., 2007; De Corte et al., 2011). Computationally, Pareto optimization estimates Pareto-optimal (PO) weights that optimize one objective (predictive validity), given a specified level on another objective (group score differences that may contribute to adverse impact) or vice versa (estimate PO weights that minimize group score differences at a specified level of predictive validity).⁸ Pareto optimization provides a frontier or curve of possible solutions at varying levels of predictive validity and group score (subgroup) differences.

Previous research highlights the advantages of using PO weighting over unit weighting when one's goal is to create selection systems that optimally balance multiple objectives, specifically, predictive validity with minimizing adverse impact potential as measured by subgroup differences in predictor scores by gender, race, and ethnicity (Wee et al., 2014; Song et al.,

⁸ The end user can ultimately select a Pareto solution for operational use based on a desired validity or adverse impact level (from the available options), depending on the goals and constraints of a given selection system. For example, an end user may select the PO solution with the highest validity level for an adverse impact ratio greater than .80 (i.e., "four-fifths rule" from Uniform Guidelines on Employee Selection Procedures [1978]). Alternatively, an end user can look at the Pareto curve specifications to select a solution that displays reduced adverse impact based on an acceptable level of validity decrement (or vice versa). For example, if the Pareto curve is relatively flat, then a solution can likely be selected with substantially reduced adverse impact and only minimal decreases in validity (and subsequent expected job performance).

2017). Furthermore, Pareto optimization is recognized as a promising development and is generally viewed positively by personnel selection researchers (Rupp et al., 2020).

While Pareto optimization procedures were similar across enlisted and officer samples, there was one key difference. As highlighted in subsection 2.2.2, we tailored enlisted models to the FPPS criterion⁹, whereas the officer models were tailored to the FSG criterion. FPPS is a dichotomous pass-fail metric, which necessitates a different modeling/optimization approach than the continuous FSG metric – analogous to logistic versus linear regression modeling. There are existing tools available to implement Pareto optimization with a continuous criterion; we used the *ParetoR* R package (Song et al., 2017) to calibrate FSG models. However, there was no existing tool to implement Pareto optimization with dichotomous criteria. Upon inquiry, an expert in the domain of Pareto optimization, Dr. Wilfried De Corte, adapted his continuous Pareto optimization program to model dichotomous criteria in the Cygwin environment (De Corte, 2022). The inputs and outputs of this program are nearly identical to the continuous-criterion implementations of Pareto optimization. The development of this tool for dichotomous criteria represents an advancement in personnel selection, given a common goal in the selection of minimizing turnover or maximizing training success formulated as pass/fail (Hom et al., 2017). Furthermore, this tool represents an advancement in modeling dichotomous outcomes because the tool allows one to develop range restriction-corrected Pareto models. Logistic regression models for dichotomous criteria require raw data as inputs, meaning logistic models calibrated on incumbent samples are affected by range restriction. However, De Corte's (2022) dichotomous Pareto optimization tool uses a correlation matrix to develop models, which allows one to correct input correlation matrices for range restriction to develop population-level Pareto models.

We used the same general inputs for both the continuous and dichotomous Pareto modeling tools. First, we selected the 10 most promising predictor scores (AFOQT subtests and SDI facets) for each career field for use in PSMs. Pareto optimization ultimately selected less than 10 scores in the final models. Two raters reviewed corrected predictive validity estimates, relative importance weights (Johnson & LeBreton, 2004), subgroup mean differences and rational review of subtest-job linkages to select 10 predictors for a career field. For each career field, we selected six cognitive subtests and four personality facets. For officers, we did not include negatively valenced SDI facets due to concerns that the predictive validity estimates for these facets may be confounded by cognitive ability (i.e., smarter officer candidates may be more likely to or more skilled at responding to these items in a socially desirable manner) and concerns that these facets were more susceptible to faking or coaching effects. There were no negatively-valenced facets within the TAPAS.

We selected the following Pareto modeling inputs with the two related goals of (1) developing PSMs to be generalizable to the USAF enlisted and officer applicant populations and (2) developing PSMs that will be robust to cross-validation in new samples. Accordingly, we used population-level subtest intercorrelations because doing so provided more stable estimates of predictor interrelationships than career field-specific estimates. Furthermore, there was an expectation that cognitive and personality trait relationships should be stable across career fields.

⁹ There was one exception to this rule, 1C6X1/5S0X1, which did not have sufficient training pass/fail variability to model FPPS. This AFSC's PSM is tailored to FSG.

We used career field-specific predictor subtest validity estimates that were similarly corrected to the USAF applicant population. The subtest subgroup difference estimates were estimated in the USAF applicant population. Subgroup differences were sample size-weighted across all possible majority-minority combinations of race-gender-ethnicity, with White non-Hispanic males as the majority group (White non-Hispanic Male [majority group] versus White non-Hispanic Female, Black non-Hispanic Male, Black non-Hispanic Female, White Hispanic Female, White Hispanic Male, Asian non-Hispanic Female, Asian non-Hispanic Male, Multiracial non-Hispanic Female, and Multiracial non-Hispanic Male [minority groups]). For example, for a given subtest, we calculated the average of all majority-minority group d values, weighted by minority group sample size, such that larger minority groups received more weight in the final estimated subtest d value.

We generated 20 Pareto solutions per career field, which provided a variety of optimal solutions at differing levels of predictive validity and subgroup differences (adverse impact potential). Pareto solution 1 optimized validity exclusively, and was identical to standard regression modeling keyed to a single criterion (predicting training success) only. Pareto solution 20 optimized diversity exclusively, estimating the solution with the lowest subgroup score differences. Each of these Pareto solutions contained a unique set of standardized weights. The Pareto solutions with lower numbers placed proportionally more weight on cognitive (ASVAB, AFOQT) than non-cognitive (TAPAS, SDI) scores, whereas the higher Pareto solution numbers placed proportionally more weight on non-cognitive facets.

Once target (standardized) weights were generated by the Pareto programs, we rounded the target weights to integer weights (rounded to the nearest 0.5) to avoid over-capitalizing on sample-specific variance. We converted standardized weights (resulting from the Pareto estimation program) into unstandardized weights to account for differing subtest scales, as inputs to the Pareto estimation procedure (i.e., correlation matrix, subgroup d values) did not account for subtest scaling. We estimated raw PSM scores for each career field's set of 20 Pareto solutions by applying the unstandardized weights to applicant or trainee subtest scores. Subsequently, all raw PSM scores were converted to normed (percentile) PSM scores using the USAF enlisted and officer applicant populations as the norming samples, respectively.

For each PSM, we estimated composite validity and subgroup mean score differences. PSM composite validity and subgroup differences were benchmarked against the operational MAGE and AFOQT composites for enlisted and officers, respectively. We also estimated predictive validity and subgroup differences on any operational PSMs that had been previously developed and implemented within career fields. PSM composite validity estimates were calculated using the same matrix multiplication procedure as for the MAGE/AFOQT composites, which involved applying a subtest PSM weight matrix to MVRR-corrected predictor subtest covariance matrices. PSM subgroup mean differences were estimated by calculating Cohen's d 's on the normed PSM scores.¹⁰

¹⁰ Raw PSM scores were used to estimate validity and normed PSM scores to estimate subgroup differences. We acknowledge that using the same PSM score type across estimations is ideal. However, raw PSM scores were used in composite validity estimation because of practical limitations of the range restriction correction method. In our view, the most accurate correction method given the available data was the matrix multiplication procedure

Finally, two analysts reviewed all PSM results and ultimately identified three to four PO solutions per career field to test when evaluating operational job qualification (discussed in Section 4). The PO solutions identified were based on careful consideration of the trade-offs between validity and diversity overall and relative to the operational MAGE/AFOQT composites. PO solutions identified were intended to reflect a reasonable range of solutions that may be of interest to USSF/USAF for possible operational implementation—some with greater weight placed on validity and others with differing levels of weight placed on validity versus reduction of subgroup differences on composites. Identification of PO solutions considered the slope of the Pareto curve—steeper Pareto curve slopes indicated greater performance costs for reducing subgroup differences, whereas flatter slopes indicated lower performance costs for reducing subgroup differences. Of the three to four PO solutions identified per career field, the first solution identified was weighted most heavily toward validity, whereas the other solutions reflected varying balances between validity and reduction of subgroup differences. This range of solutions for each career field was presented to USSF for consideration in operational job qualification. The solutions identified are referred to as “candidate PSMs” and represent composites that may be better performing than current qualification composites in terms of validity or reducing subgroup differences.

The PSMs estimated in the current sample have not been cross-validated in a separate sample. Consequently, the estimated PSM weights were subject to shrinkage. In comparison, the MAGE composites were developed in earlier samples and have been validated with subsequent samples, multiple times (e.g., Beatty et al., 2020; Carretta, 2014; Diaz & Ingerick, 2008). To evaluate the level of validity estimated for PSMs in future samples, we adopted a technique recently developed by Song et al. (2023) for estimating shrinkage in Pareto frontiers. Critically, at the endpoint of the Pareto curve where validity is maximized (i.e., Pareto solution #1), the predictor weights are overfitted for the validity objective, whereas overfitting is negligible for the diversity objective (because diversity is not assigned any weight in the model for Pareto #1). Conversely, predictor weights are overfitted for diversity at the endpoint of the Pareto curve where diversity is maximized (Song et al., 2017). Given this observation, Song et al. (2023) used formula-based shrinkage estimates at both endpoints and developed a formula to interpolate shrunk R values for each Pareto solution between the two endpoints. Song et al. compared this method to empirical cross-validation (i.e., assessing model prediction in a holdout sample), and found that the formula-based approach accurately approximated empirical cross-validation in calibration sample sizes of 500 and 1,000 with 10 predictors (cognitive and non-cognitive), and slightly underestimated cross-validated shrinkage under the same conditions with a calibration sample size of 200. We adopted the Song et al. formula-based approach using the recommended Claudy-Browne shrinkage formula. Pareto curves figures presented show estimates adjusted for shrinkage (in grey) in each of the Pareto frontiers (Figures 7, 9, 11, 13, 15, 17, 19). We note that

described above. This procedure used predictor subtest/facet scores and weights, meaning we could not incorporate the normed PSM scores into composite validity estimation. We estimated subgroup differences on the normed PSMs because that is the metric proposed for operational use. Therefore, we decided to use the best possible estimates of validity and subgroup differences, even though one used raw scores and the other used normed scores. We estimated d -values both ways (raw, normed) and found differences to be trivial. We could not estimate validity estimates both ways due to psychometric limitations. However, we note that the raw and normed PSM scores all correlated $\geq .97$. Therefore, we are confident the raw-to-percentile conversion would not have made an appreciable difference between raw and normed PSM scores' validity estimates.

diversity shrinkage was minimal throughout due to subgroup difference estimation in the large applicant enlisted and officer applicant populations.¹¹ Similarly, validity shrinkage tended to be small ($\leq .03$ for FPPS, $\leq .05$ for FSG), and was smallest in the career fields with larger trainee samples.

Another source of shrinkage is the pre-selection of 10 subtests that were fed into each Pareto model, selected from the larger battery of subtests (26 for enlisted personnel, 40 for officers). Validity estimates were one of three sources of information used to select the top 10 subtests per career field, and validities were estimated on the same sample as Pareto models were calibrated – which is also likely to contribute to shrinkage. Therefore, we conducted a simulation to evaluate the extent of shrinkage due to pre-selection and to get a sense of the accuracy of formula-based shrinkage estimates in our data compared to empirical cross-validation. We treated the full, observed predictor-criterion data as the population for a given career field. Then, we simulated 1,000 multivariate normal samples, specifying the population multivariate normal distributions (i.e., mean and covariance matrix) from the observed predictor-criterion data for a career field, with each sample containing size n (where n = sample size of trainee sample for a given career field). Each simulated sample represented a multivariate normal draw of size n from a career field's predictor-criterion data. Within each sample, we identified the top 10 subtests (six cognitive, four non-cognitive) based on zero-order predictor criterion r 's in that sample. Next, we regressed the FSG criterion on the top 10 subtests within a sample and then calculated (a) the observed Multiple R and (b) the formula-based shrinkage-adjusted R (using the Claudy-Browne formula). Finally, we applied the regression coefficients from each sample back to the career field population (i.e., observed data) and calculated empirical cross-validated Multiple R 's. The difference between the formula-based shrinkage-adjusted R and the empirical cross-validated R quantifies the degree to which the shrinkage formula overestimates cross-validated prediction after accounting for predictor pre-selection since the top 10 predictors were selected in the same simulated samples in which the regression models were calibrated. This comparison allowed for evaluation of the overall accuracy of the formula-based shrinkage methods in the USAF/USSF-specific data.

Overall, the average difference between the formula-based shrinkage-adjusted R and empirical cross-validated R was .008 ($SD = .006$ across career fields), indicating that the formula-based shrinkage adjustment slightly overestimated empirical cross-validity after accounting for predictor pre-selection. We note that this finding was only applicable to Pareto solution #1's (regression weights), and that R differences between empirical cross-validated and formula-based R 's will be smaller for subsequent Pareto solutions (i.e., #2 on). Furthermore, our rounding of raw weights to integer weights (rounded to the nearest 0.5) is likely to further decrease shrinkage and improve cross-validity. Given these factors, as well as the magnitude of the R difference between formula-based and empirical methods (less than .01), we view the formula-

¹¹ Final PO weights may also display less diversity shrinkage than weights presented in figures because the Pareto optimization process averaged subgroup difference estimates across multiple subgroups rather than optimizing for a single subgroup.

based shrinkage-adjustments presented in the subsequent sections to be relatively accurate estimates of empirically cross-validated R^2 s.¹²

3.2 Enlisted Predictive Success Model (PSM) Results

Attachment C provides composite-level predictive validity and subgroup estimates for enlisted candidate PSM composites, operational MAGE composites, and operational PSMs (developed by USAF previously), along with shrinkage-adjusted estimates of validity and subgroup differences. To facilitate the interpretation of the candidate Pareto PSMs, we provide figures illustrating Pareto curves for each career field's Pareto solutions (i.e., all potential PSMs), comparing Male-Female and Black-White subgroup differences and benchmarked against operational composites. We also present subtest weighting schemes for one candidate PSM per career field, which were ultimately presented as options for implementation. Importantly, across PO curve figures, one can observe multiple candidate PSMs identified per career field, representing differing emphasis placed on validity versus reduction of subgroup differences. For the PSMs, all text below focuses on interpreting shrinkage-adjusted validity and subgroup difference estimates. Figure 7 presents Pareto frontiers for the PSMs developed within the Space Systems Operations (SSO) career field relative to the composite used in operational qualification for this career field, MAGE-Electronics (E). Points toward the upper-right portion of the graph are desirable because they represent composites that have high validity and low subgroup differences. All candidate PSMs (or "better-performing options") that were identified for potential implementation are numbered in red, where the number represents the PO solution number. The operational composites currently used in job qualification are highlighted in blue. For example, upon review of the PSM results, SSO's Pareto solution 3 was identified as one candidate PSM for potential operational use in job qualification. This solution displayed an estimated shrinkage-adjusted validity (ρ) .01 lower than the regression-weighted solution (Pareto solution 1/candidate PSM 1; ρ s = .53 versus .52), while displaying reduced shrinkage-adjusted subgroup differences of +.20 for Female-Male differences (d s = -.15 versus -.35) and +.18 for Black-White differences (d s = -.63 versus .81). This PSM displays both validity and diversity advantages relative to the current operational composite (MAGE-E). Specifically, the candidate PSM displayed a similar level of prediction of FSG as the operational MAGE-E composite (ρ s = .52 versus .51), a .51 reduction in Female-Male subgroup differences (d = -.15 versus -.66), and a .35 reduction in Black-White subgroup differences (d s = -.63 versus -.98). Therefore, implementation of this PSM is expected to offer similar validity and substantial reduction of adverse impact relative to the operational standard.

¹² Implementing the simulation for the three enlisted career fields using the FPPS criterion was not feasible due to challenges associated with simulating multivariate normal-distributed data with one dichotomous variable (FPPS) and other continuous variables. Therefore, the simulation analyses were conducted on one enlisted career field using FSG (Space Systems Operations) and all three officer career fields.

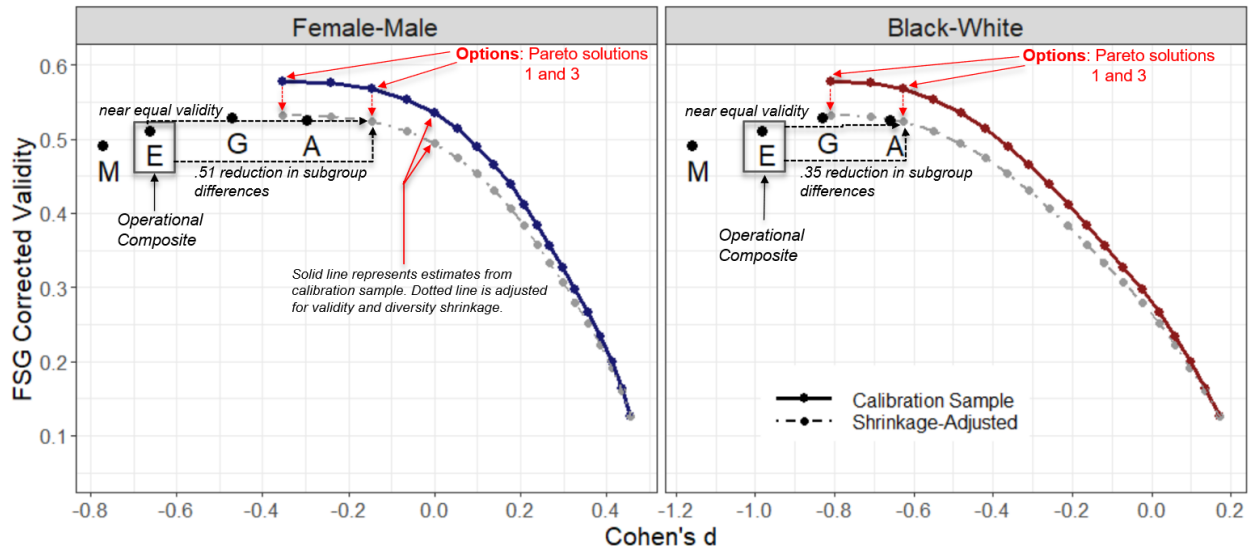


Figure 7. Space Systems Operations Pareto frontiers for candidate PSMs and operational composites

Figure 8 presents the subtest/facet weighting scheme for the highlighted candidate PSM for SSO. For conciseness, we only graphically depict and interpret one PO weighting scheme per career field. In this composite, the ASVAB subtests receive most of the weight (79%), although three TAPAS facets do receive weight. The ASVAB subtests of Arithmetic Reasoning and Math Knowledge are used in the MAGE-E composite, although Paragraph Comprehension and Assembling Objects are not currently used for qualification into this career field. Overall, the candidate PSM suggests that a profile for training success in this career field is a mixture of reading skills, math skills, and spatial reasoning.¹³

¹³ We do not interpret TAPAS facet composition because PSM composition information is restricted with respect to TAPAS facets.

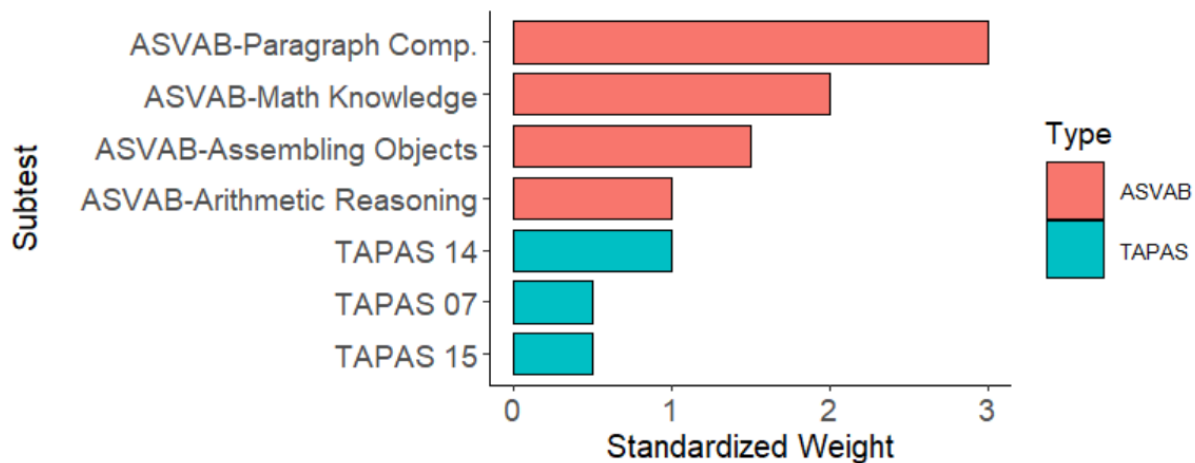


Figure 8. Space Systems Operations candidate PSM #3 composition

Figure 9 presents Pareto frontiers for the candidate PSMs developed within the Intelligence–All Source Analysts career field, relative to the composites used in operational qualification for this career field, MAGE-General (G), MAGE-Administrative (A), and pre-existing PSMs. Pareto solution’s 2, 5, and 9 were identified as candidate PSMs (or “better-performing options”) for use in job qualification. We highlight candidate PSM #5, although candidate PSM #2 and #9 were also presented as better-performing options. Candidate PSM #5 displayed a shrinkage-adjusted validity (ρ) .01 lower than the regression-weighted solution (Pareto solution 1/candidate PSM 1; ρ s = .32 versus .31) while displaying reduced subgroup differences of +.27 for Female-Male differences (d s = -.26 versus -.53) and +.26 for Black-White differences (d s = -.71 versus -.97). Similarly, this PSM displays both validity and diversity advantages relative to the current operational composites (MAGE-G and MAGE-A). This candidate PSM displays +.05 better shrinkage-adjusted prediction of FPPS than the operational MAGE-G composite (ρ s = .31 versus .26), a .21 reduction in Female-Male subgroup differences (d s = -.26 versus -.47), and a .12 reduction in Black-White subgroup differences (d s = -.71 versus -.83). The candidate PSM also displays +.05 better prediction of FPPS than the operational MAGE-A composite (ρ s = .31 versus .26), a .04 reduction in Female-Male subgroup differences (d s = -.26 versus -.30), but a .05 increase in Black-White subgroup differences (d s = -.71 versus -.66). Therefore, implementation of this candidate PSM is expected to offer validity improvements and slight overall improvements in adverse impact relative to the operational standard.

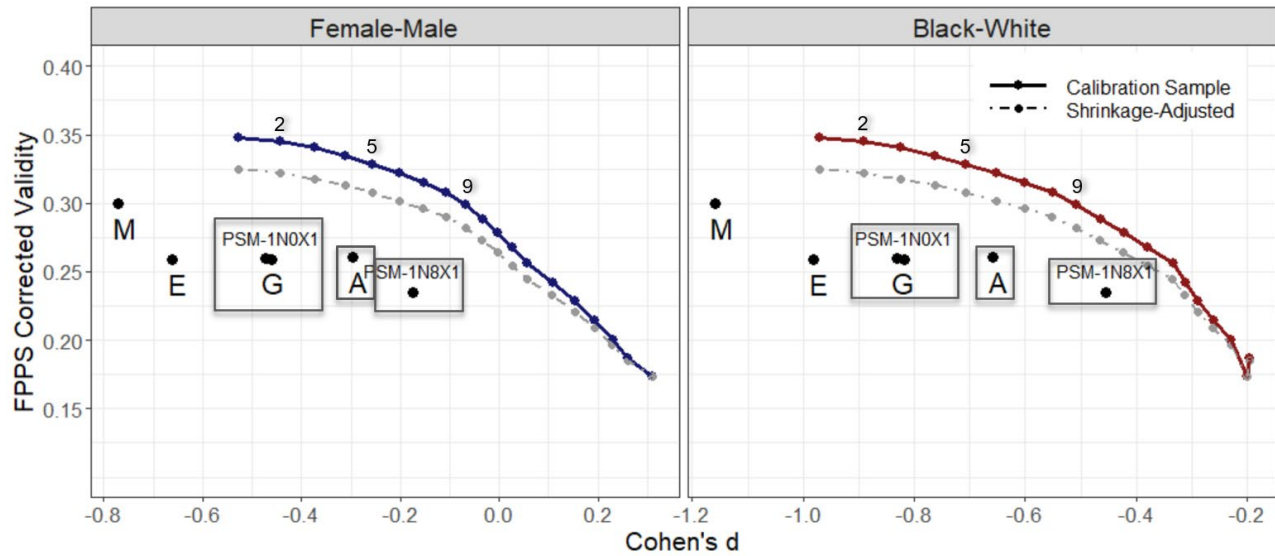


Figure 9. Intelligence – All Source Analysts Pareto frontiers for candidate PSMs and operational composites

Figure 10 presents the subtest weighting scheme for Intelligence-All Source Analysts' candidate PSM #5. The ASVAB subtests and TAPAS facets receive similar proportions of weight (55% versus 45%, respectively). The ASVAB subtest of Paragraph Comprehension is currently used in the MAGE-A and MAGE-G composites, while the Math Knowledge subtest is currently used in the MAGE-A composite. However, Assembling Objects and Auto-Shop are not currently used for qualification into this career field. The PSM suggests that a profile for training success in this career field is a mixture of reading skills, math skills, spatial reasoning, and knowledge of machinery systems/repairs.



Figure 10. Intelligence – All Source Analysts candidate PSM #5 composition

Figure 11 presents Pareto frontiers for the candidate PSMs developed within the Intelligence-Cryptologic Analysts career field, relative to the composite used in operational qualification for this career field, MAGE-G or MAGE-A. The better-performing options for SSO were identified as Pareto solutions 1 and 4. Candidate PSM #4 displayed non-trivial diversity advantages relative to the regression-weighted solution (candidate PSM #1) and the current operational MAGE composites (MAGE-G and MAGE-A). Candidate PSM #4 displayed a shrinkage-adjusted validity (ρ) .01 lower than the regression-weighted solution (ρ s = .30 versus .29), while displaying reduced shrinkage-adjusted subgroup differences of +.29 for Female-Male differences (d s = .01 versus -.28) and +.25 for Black-White differences (d s = -.46 versus -.71). Relative to MAGE composites, this candidate PSM also shows +.02 better prediction of FSG than MAGE-G (ρ s = .29 versus .27), a .48 reduction in Female-Male subgroup differences (d s = .01 versus -.47), and a .37 reduction in Black-White subgroup differences (d s = -.46 versus -.83). This candidate PSM displays the same level of prediction of FSG as MAGE-A (ρ s = .29 for both), a .29 reduction in Female-Male subgroup differences (d s = .01 versus -.30), and a .20 reduction in Black-White subgroup differences (d s = -.46 versus -.66). Therefore, the implementation of candidate PSM #4 is expected to offer slight validity improvements and substantial reduction of adverse impact relative to the operational standards.

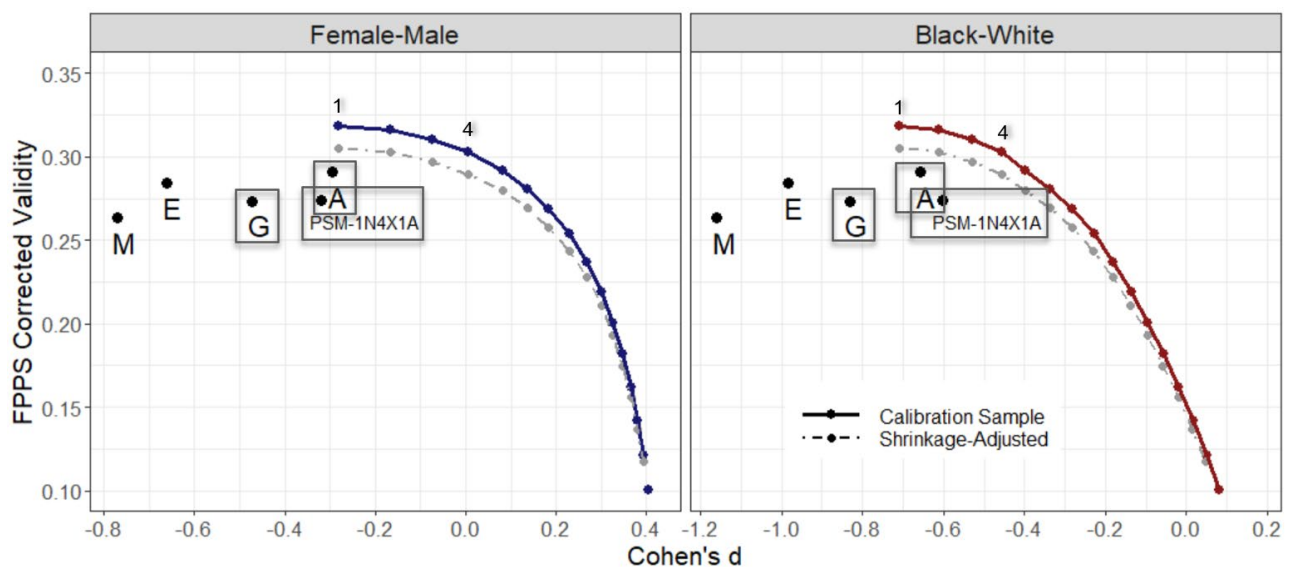


Figure 11. Intelligence – Cryptologic Analysts Pareto frontiers for candidate PSMs and operational composites

Figure 12 presents the subtest weighting scheme for Intelligence-Cryptologic Analysts' candidate PSM #4. The ASVAB and TAPAS receive similar weight in this composite (57% versus 43%, respectively). The ASVAB subtest of Paragraph Comprehension is currently used in the MAGE-A and MAGE-G composites, and the Math Knowledge subtest is currently used in the MAGE-A composite. Additionally, the Cyber Test is used for qualification to two of the four AFSCs comprising this career field. Therefore, the cognitive subtests used in this PSM are all used in operational job qualification within this career field. However, the advantages for the candidate PSM come from (a) incorporating scores on personality facets into models and (b) re-

weighting scores on the aptitude subtests to improve the prediction of training success. Overall, the candidate PSM suggests that a profile for training success in this career field is a mixture of reading, math, and cyber skills but is most heavily related to math skills-knowledge.

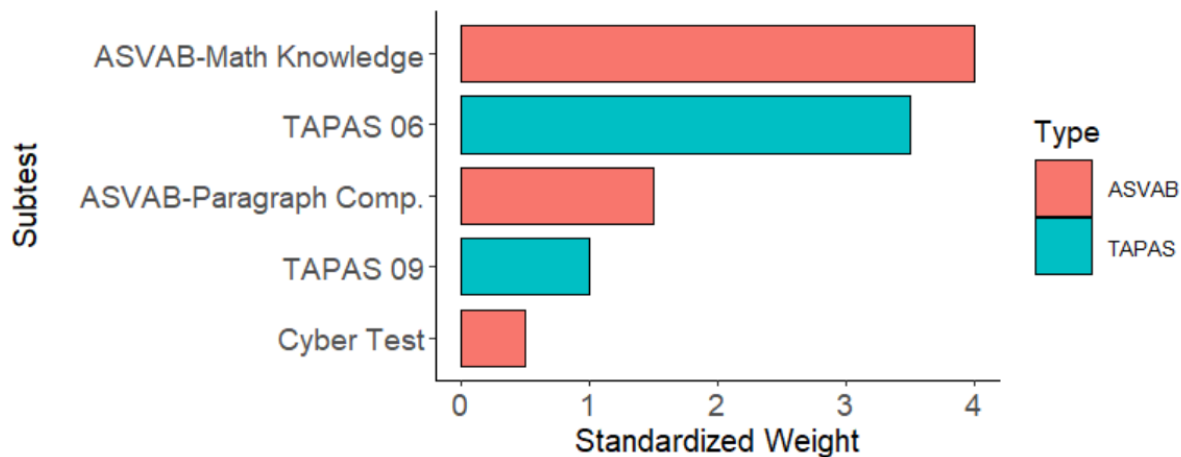


Figure 12. Intelligence – Cryptologic Analysts candidate PSM #4 composition

Figure 13 presents Pareto frontiers for the candidate PSMs developed within the Cyber Operations career field relative to the composite used in operational qualification for this career field, MAGE-Mechanical (M), MAGE-E, and MAGE-G. Upon a review of PSM results, Cyber Operation's Pareto solutions #2 and #5 were identified as better-performing options for potential operational use. With respect to candidate PSM #5, this composite displayed a shrinkage-adjusted validity (ρ) .03 lower than the regression-weighted solution (ρ s = .41 versus .38), while displaying reduced subgroup differences of +.26 for Female-Male differences (d s = -.23 versus -.49) and +.20 for Black-White differences (d s = -.53 versus -.73). The candidate PSMs displayed both validity and diversity advantages relative to the current operational composites (MAGE-M, MAGE-E, and MAGE-G). Candidate PSMs #2 and #5 yield a +.01 to +.06 improved prediction of FPPS than MAGE-M (ρ s = .41 and .38 for PSMs 2 and 5, respectively, versus .35), MAGE-E (ρ = .37), and MAGE-G (ρ = .36). Candidate PSMs #2 and #5 yield at least a .05 reduction in Female-Male subgroup differences compared to MAGE-M (d s = -.42 and -.23, respectively, versus -.77), MAGE-E (d = -.66), and MAGE-G (d = -.47). PSMs #2 and #5 yield at least a .17 reduction in Black-White subgroup differences compared to MAGE-M (d s = -.66 and -.53, respectively, versus -1.16), MAGE-E (d = -.98), and MAGE-G (d = -.83). Therefore, implementation of these PSMs is expected to offer slight validity improvements and reduction of adverse impact relative to the operational standards.

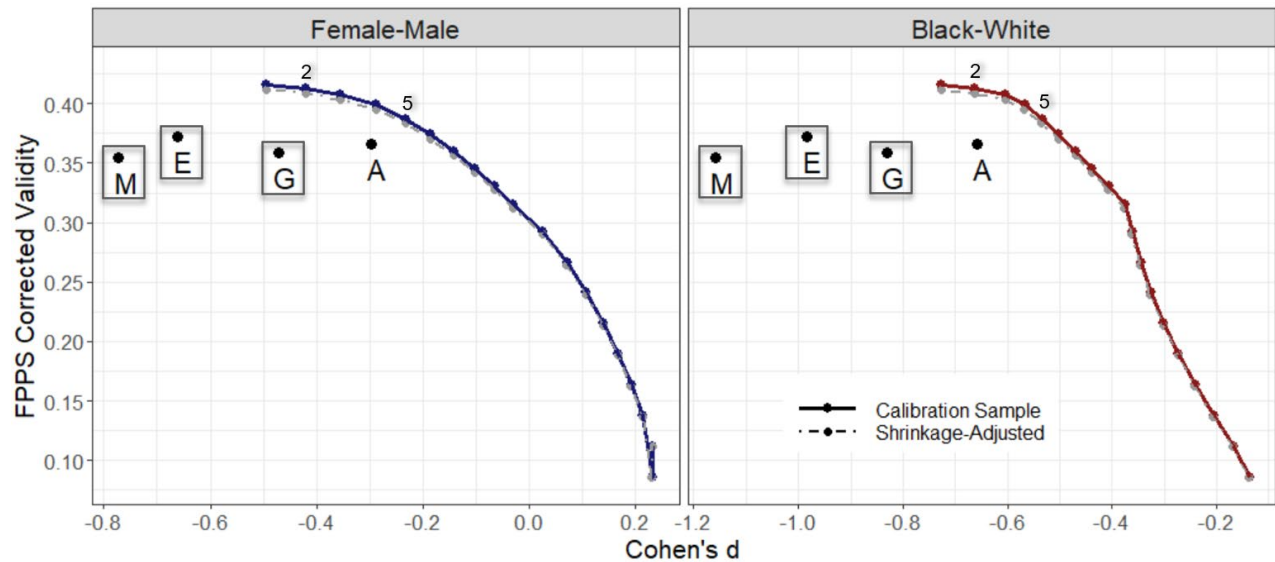


Figure 13. Cyber Operations Pareto frontiers for candidate PSMs and operational composites

Figure 14 presents the subtest weighting scheme for Cyber Operations' PSM #5. The ASVAB subtests receive most of the weight (63%) in this composite, although four TAPAS facets do receive weights. The ASVAB subtests and the Cyber Test are all used in operational job qualification for this career field. However, the advantages for the candidate PSM come from (a) incorporating scores on personality facets into models and (b) re-weighting scores on the aptitude subtests to improve training success prediction. Overall, the PSM suggests that a profile for training success in this career field is heavily related to math skills/knowledge, along with cyber and reading skills.

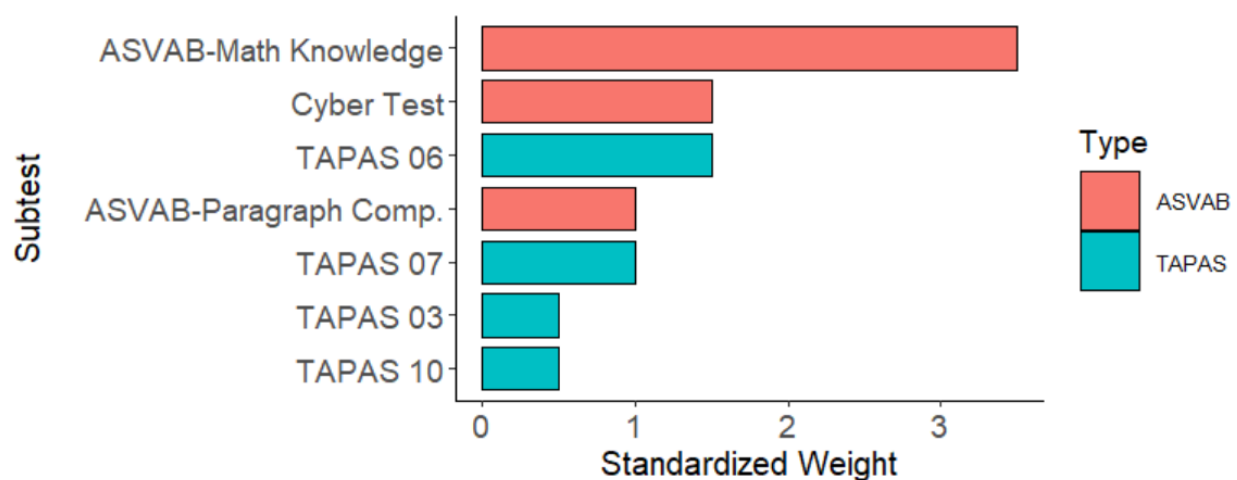


Figure 14. Cyber Operations candidate PSM #5 composition

3.3 Officer Predictive Success Model (PSM) Results

Attachment C provides composite-level predictive validity and subgroup estimates for officer PSM composites and operational AFOQT composites, along with shrinkage-adjusted estimates of validity and subgroup differences. Figures 15, 17, and 19 present Pareto frontiers for each of the officer career fields, relative to operational AFOQT standards. We note that Pareto curves are steeper for officers than for enlisted, which indicates there is a greater performance cost associated with a reduction in adverse impact for officers. Of note, several solutions identified as candidate PSMs displayed non-trivial decreases in validity relative to the regression-weighted solution (Pareto #1). However, when it comes to operationally implementing the PSMs, the impact of lower validities is much smaller because of the higher targeted qualifying rates for officers, the high base rate of training success (mean FSG), and the low variability in training success (*SD* FSG). We discuss this finding further in section 4.

Figure 15 presents Pareto frontiers for the candidate PSMs developed within the Space Operations (SO) career field relative to the operational AFOQT-Q and AFOQT-V composites. Upon review of PSM results, SO's Pareto solutions #4, #6, and #9 were identified for potential use in job qualification. The candidate PSMs ranged from performing slightly better to worse in predicting FSG, adjusting for shrinkage (ρ s = .55, .49, and .42 for PSM 4, 6, and 9, respectively) compared to AFOQT-Q and AFOQT-V (ρ = .51 for both). The candidate PSMs range from slightly better to slightly worse concerning Female-Male subgroup differences (d s = -.43, -.37, and -.32 for PSMs 4, 6, and 9, respectively) compared to AFOQT-Q (d = -.43) and AFOQT-V (d = -.30). Likewise, the candidate PSMs range from performing slightly better to slightly worse in terms of Black-White subgroup differences (d s = -1.03, -.92, and -.77 for PSMs 4, 6, and 9, respectively) compared to AFOQT-Q (d = -.88) and AFOQT-V (d = -.99). Therefore, implementation of these PSMs is likely to offer similar performance as the AFOQT.

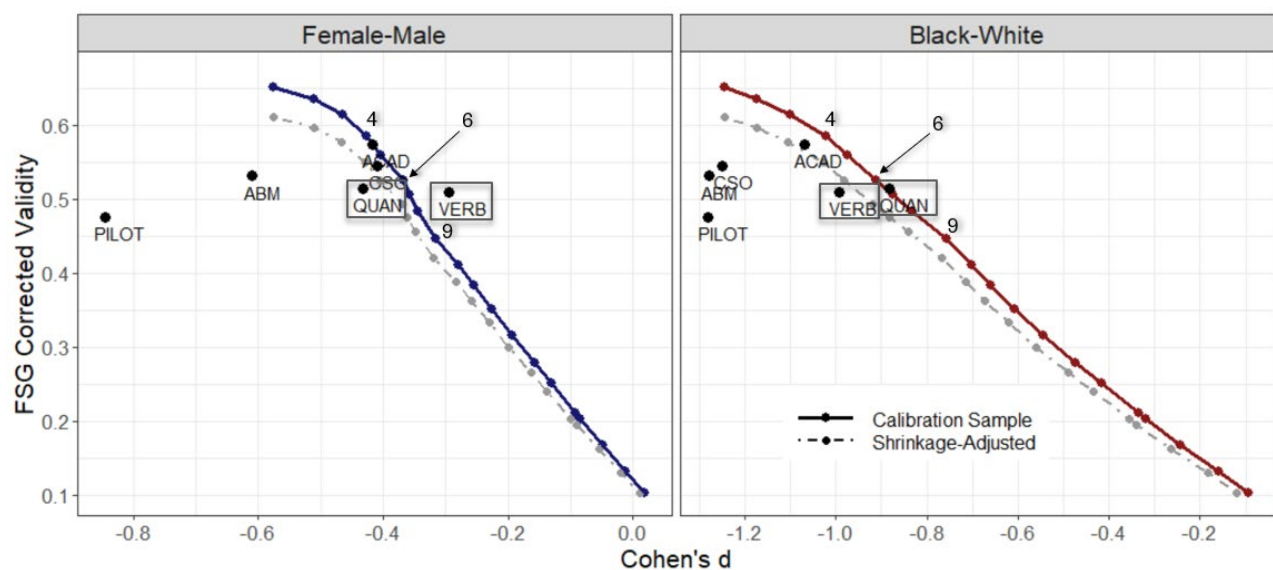


Figure 15. Space Operations Pareto frontiers for candidate PSMs and operational composites

Figure 16 presents the subtest weighting scheme for Space Operations' candidate PSM #6. AFOQT-Math Knowledge receives the largest proportion of weight, although Reading Comprehension and two SDI facets are weighted at smaller levels. The Math Knowledge and Reading Comprehension subtests are respectively used in the AFOQT-Verbal and Quantitative composites. Overall, the PSM suggests that a profile for training success in this career field is strongly weighted towards math knowledge/skills.

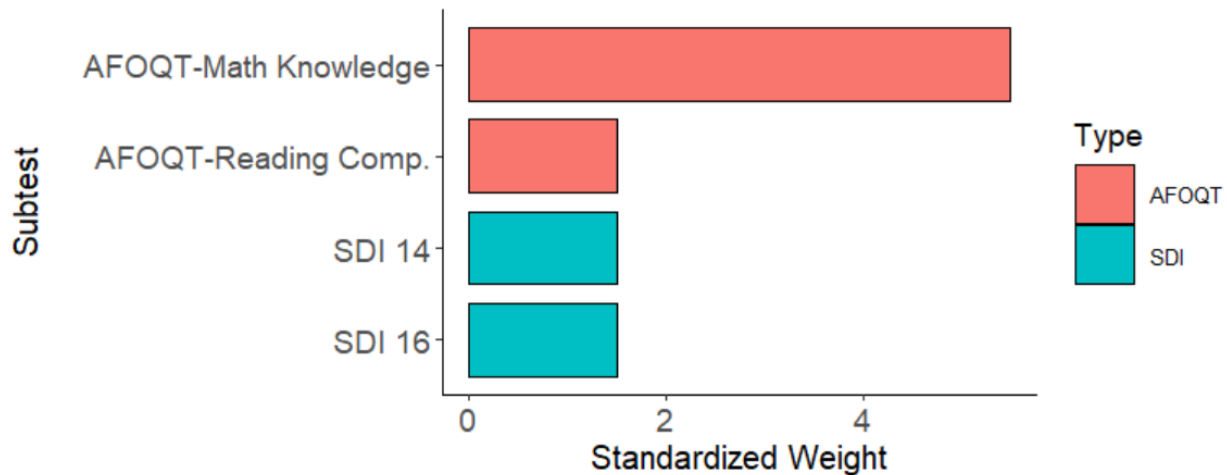


Figure 16. Space Operations candidate PSM #6 composition

Figure 17 presents Pareto frontiers for the candidate PSMs developed within the Intelligence Operations career field, relative to the AFOQT-Q and AFOQT-V composites. Upon review of PSM results, Intelligence Operation's Pareto solutions #4 and #8 were identified for potential operational use. The candidate PSMs range from slightly better to slightly worse in predicting FSG, adjusting for shrinkage (ρ s = .52 and .42 for PSM 4 and PSM 8, respectively) compared to AFOQT-Q (ρ = .45) and AFOQT-V (ρ = .49). The candidate PSMs also performed about as well as or better in reducing Female-Male subgroup differences (d s = -.28 and -.19 for PSMs 4 and 8, respectively) compared to AFOQT-Q (d = -.43) and AFOQT-V (d = -.30). Likewise, the candidate PSMs performed as well as or better in reducing Black-White subgroup differences (d s = -.88 and -.49 for PSMs 4 and 8, respectively) compared to AFOQT-Q (d = -.99) and AFOQT-V (d = -.88). Therefore, implementation of these candidate PSMs may offer slight improvements in either training success or the diversity of qualified applicants.

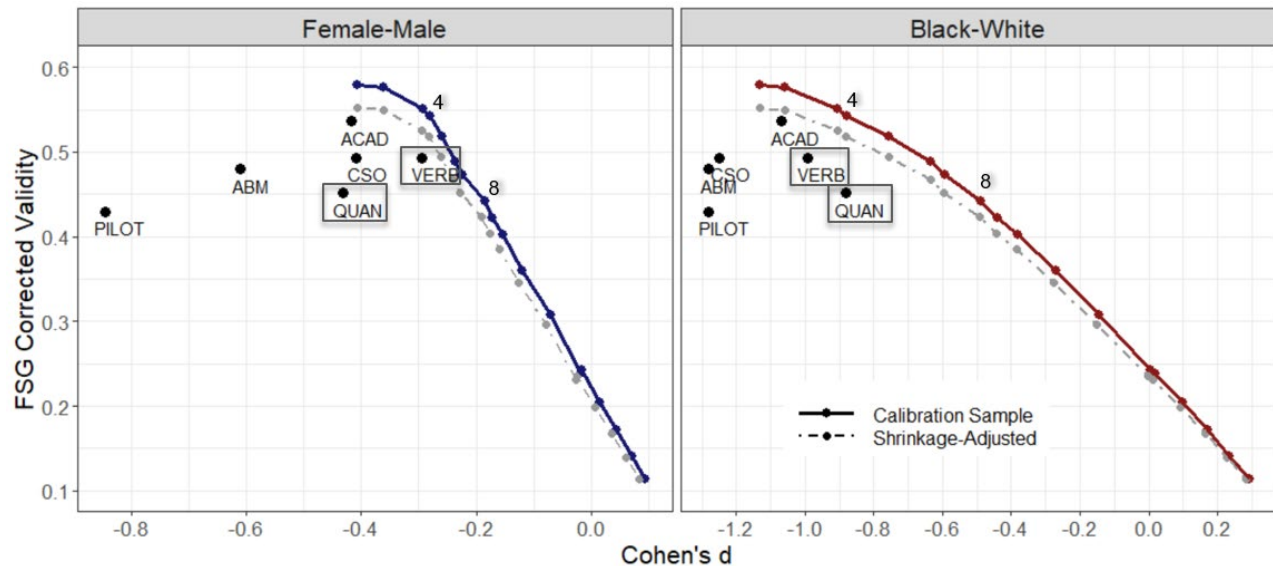


Figure 17. Intelligence Operations Pareto frontiers for candidate PSMs and operational composites

Figure 18 presents the subtest weighting scheme for Intelligence Operations' candidate PSM #4. The AFOQT subtests receive much of the weight (70%) in this composite, although three SDI facets also receive weight. The AFOQT subtests of Math Knowledge and Word Knowledge currently contribute to the operational AFOQT-Q and V composites. However, Table Reading is not currently used in officer commissioning. Overall, the candidate PSM suggests that a profile for training success in this career field is a mixture of reading and math knowledge, along with the ability to read and interpret tables.



Figure 18. Intelligence Operations candidate PSM #4 composition

Figure 19 presents Pareto frontiers for the PSMs developed within the Cyber Operations career field relative to the AFOQT-Q and AFOQT-V composites. Upon review of PSM results, Cyber Operation's Pareto solutions #4 and #9 were identified for potential use in operational qualification. The candidate PSMs perform about as well in predicting FSG, adjusting for shrinkage ($\rho_s = .47$ and $.32$ for PSMs 4 and 9, respectively) as AFOQT-Q ($\rho = .40$) and AFOQT-V ($\rho = .49$). The candidate PSMs ranged from slightly worse to slightly better in Female-Male subgroup differences ($ds = -.37$ and $-.27$ for PSMs 4 and 9, respectively) compared to AFOQT-Q ($d = -.43$) and AFOQT-V ($d = -.30$). Likewise, the candidate PSMs varied in reduction in Black-White subgroup differences ($ds = -1.03$ and $-.71$ for PSMs 4 and 9, respectively) compared to AFOQT-Q ($d = -.88$) and AFOQT-V ($d = -.99$). Therefore, implementation of these PSMs is expected to offer similar performance as AFOQT-V and may offer slight improvements on AFOQT-Q.

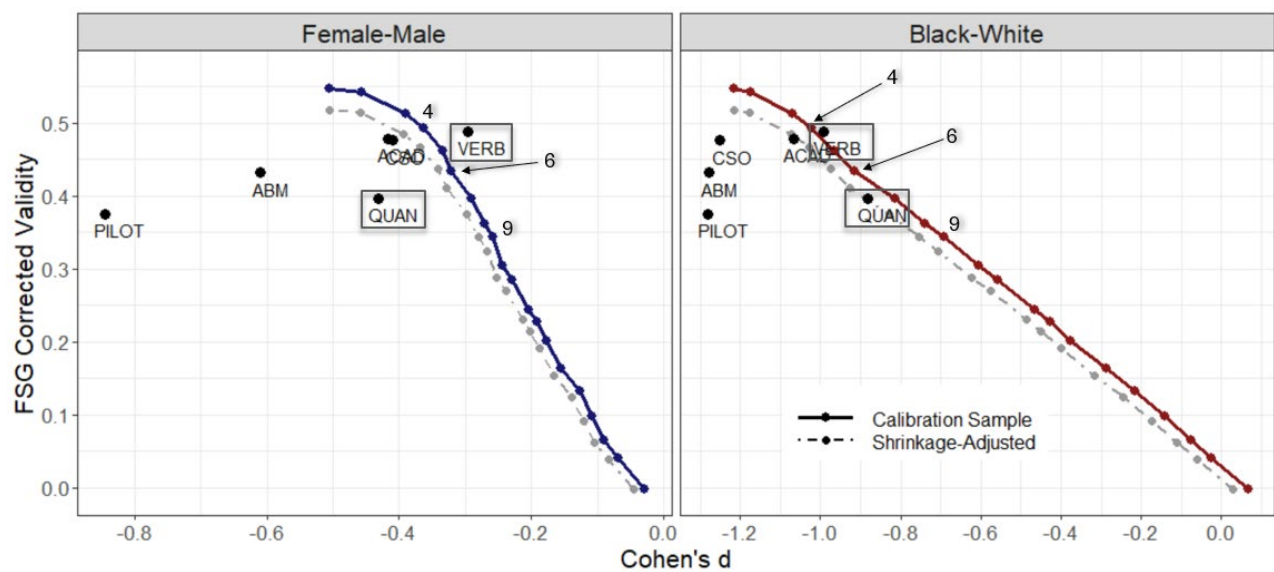


Figure 19. Cyber Operations Pareto frontiers for candidate PSMs and operational composites

Figure 20 presents the subtest weighting scheme for the Cyber Operations' candidate PSM #4. The AFOQT subtests receive much of this composite's weight (76%), although two SDI facets also receive weight. The AFOQT subtests of Math Knowledge, Reading Comprehension, Word Knowledge, and Verbal Analogies are all used in the current AFOQT-Verbal and Quantitative composites. The similarity in subtest weighting schemes between the operational AFOQT composites and the candidate PSM will likely explain the similar validity-diversity results observed in the Pareto frontier graph. The PSM suggests that a profile for training success in this career field is a mixture of math, reading, and verbal skills.

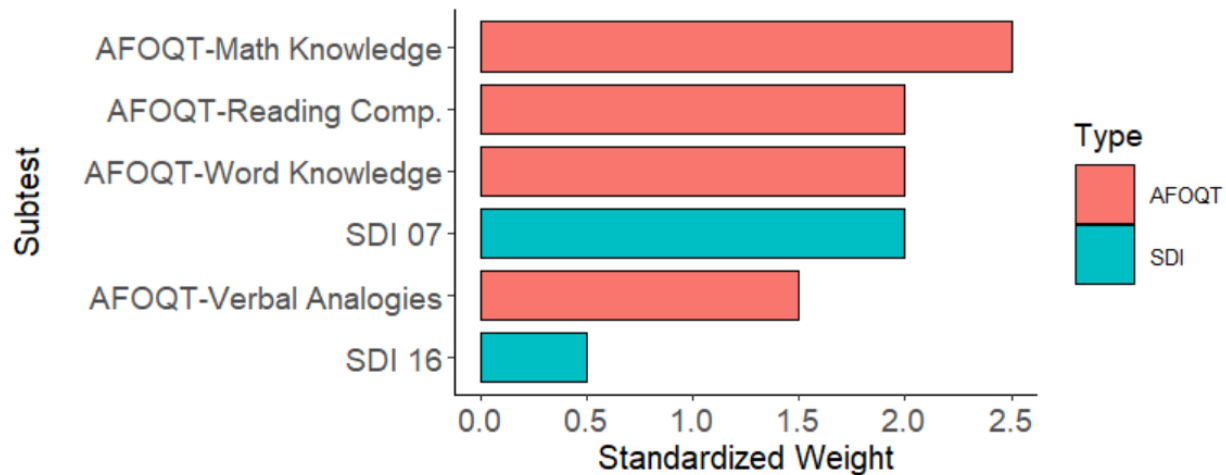


Figure 20. Cyber Operations candidate PSM #4 composition

4.0 EVALUATING PROJECTED IMPACT OF CANDIDATE PSMs

To conclude, we simulated the projected impact of implementing the “best bet” PSMs to qualify enlisted and officer recruits into the targeted AFSCs. To simulate the projected impact of PSMs on job qualification for enlisted, we tested a series of experimental PSM and joint MAGE-PSM qualification standards to qualify applicants for each SFSC/AFSC. We refer to the PSM composites as candidate PSMs (or composites), while single or multiple candidate PSMs with associated cut-scores (e.g., $\text{PSM } 1 \geq 60$) are referred to as candidate standards. The projected impact results were used to evaluate the potential benefits of using PSM scores for job qualification into each of the targeted SFSCs/AFSCs, either to supplement the operational MAGE/AFOQT composites or to replace the operational MAGE/AFOQT composites entirely.

A series of metrics were then calculated to assess how well the experimental qualification standards were functioning. Metrics calculated to inform standards’ performance included: (a) percentage of applicants qualifying (qualification rates, QRs), (b) ratio of the group-level percentages of applicants qualifying by gender and race-ethnicity, (adverse impact ratios, AIRs), (c) predicted training success rates for qualifying applicants, and (d) classification efficiency metrics (e.g., F-1 score, Matthew’s Correlation Coefficient (MCC)). Table 3 summarizes the primary metrics that were used in this evaluation.

We implemented a different approach for the officer projected impact analyses because job qualification functions differently between enlisted Airmen/Guardians and officers. At a high level, enlisted career fields qualify applicants using existing SFSC/AFSC-specific qualification standards, whereas there are no existing job qualification standards on the targeted career fields for officers. Therefore, for enlisted, we tested directly altering or replacing the SFSC/AFSC-specific standards. Conversely, all officers must meet a general (not job-specific) qualification standard to access ($\text{AFOQT-Q} \geq 10$ & $\text{AFOQT-V} \geq 15$). Therefore, we tested implementing two different qualification models for officers: (a) implementing new PSM-based standards by career field and (b) using PSMs to inform person-job matching without explicit cut-scores on PSMs. We outline the enlisted and officer projected impact approaches and results separately below.

Table 3. Projected Impact Metrics Evaluated

Metric	Definition	Formula
<i>Qualification Rate (QR)</i>	Percentage (%) of applicants that would qualify for the selected AFSC when using the specified qualification standard.	$\frac{N_{\text{meeting cut threshold}}}{N_{\text{applicants}}}$
<i>Adverse Impact Ratio (AIR)</i>	Ratios of the QRs for one group of qualifying applicants relative to a second group ($QR_{\text{Group1}} / QR_{\text{Group2}}$). AIRs were computed for: • Gender ($QR_{\text{Female}} / QR_{\text{Male}}$) • Race ($QR_{\text{Black, Asian, Multiracial}} / QR_{\text{White, Non-Hispanic}}$) • Ethnicity ($QR_{\text{White, Hispanic}} / QR_{\text{White, Non-Hispanic}}$)	$\frac{QR_{\text{Group 1}}}{QR_{\text{Group 2}}}$
<i>Predicted First Pass Pipeline Success (FPPS) Rate</i>	Predicted FPPS rate (percentage, %, successfully completing technical training on the first pass) among qualifying applicants, at a specified standard.	--
<i>Precision</i>	The proportion of predicted positive cases that are actual positives.	$\frac{TP}{TP + FP}$
<i>Recall (or sensitivity)</i>	The proportion of actual positives that were correctly classified.	$\frac{TP}{TP + FN}$
<i>F1 score</i>	Combines precision and recall into a single metric, using the harmonic mean of the two metrics.	$\frac{2 * (Precision + Recall)}{Precision + Recall}$
<i>Accuracy</i>	The proportion of cases that were correctly classified as either a positive or negative.	$\frac{TP + TN}{TP + FP + TN + FN}$
<i>Matthew's Correlation Coefficient (MCC)</i>	Overall measure of classification quality. MCC is also the phi correlation coefficient, ranging from -1 (when cases are all incorrectly classified) to +1 (when cases are all correctly classified).	$\frac{TP \times TN - \sqrt{(TP + FP)(TP + FN)}}{\sqrt{(TP + FP)(TP + FN)}}$

Notes. TP = True positive. FP = False positive. TN = True negative. FN = False negative. An applicant was considered a predicted positive (negative) if they met (failed) the standard of interest; in other words, the standard predicted that this applicant would succeed (fail) in technical training. An applicant was considered an actual positive (negative) if they were projected to pass (fail) technical training via our regression modeling.

4.1 Enlisted Projected Impact Approach

For enlisted career fields, our goal was to identify PSM-based solutions that could supplement or replace the existing MAGE or PSM qualification standards (operational standards). For each enlisted SFSC/AFSC, there was a single operational standard in use, which applicants must meet to be admitted into the job. For example, Space Systems Operation's (1C6X1/5S0X1) operational standard is currently MAGE-E ≥ 60 , meaning applicants must achieve a MAGE-E score of 60 to qualify for SSOs. We tested a series of experimental qualification standards (candidate standards) that were benchmarked against the operational MAGE (or PSM) standard.

Candidate standards were benchmarked according to (a) overall qualification, (b) diversity in qualification (i.e., QRs and AIRs by subgroup), (c) predicted training success (i.e., predicted FFPS rate among qualifying applicants), and (d) classification accuracy (the standards' accuracy in classifying training successes or failures). The same enlisted applicant analysis sample was used for projected impact analyses, although we restricted analyses to those meeting the current USAF/USSF minimum standard of AFQT ≥ 31 to serve. The AFQT screen was applied to evaluate the job qualifications of applicants who were eligible to access USAF/USSF. To model PSMs without substantially restricting sample size, we used multiple imputation on the enlisted applicant sample for the four subtests/facets with significant missingness (EDPT, Cyber Test, two TAPAS facets). Missing values were imputed 100 times and then averaged to form final scores for the four predictors with missingness.

We tested different candidate standard scenarios to supplement or replace the operational MAGE (or PSM) qualification standards. They were:

1. Pairing a PSM with an operational qualification standard, with an “or” (compensatory) decision rule (e.g., MAGE-A ≥ 72 OR PSM ≥ 65). We chose a compensatory (“or”) decision rule rather than a non-compensatory (“and”) because we expected compensatory candidate standards to display superior validity/diversity outcomes since they enabled applicants to qualify for an SFSC/AFSC based on the PSM alone. For example, an applicant with a MAGE-A score of 65 and a PSM of 66 would qualify for the hypothetical SFSC/AFSC based on the PSM alone (which is optimized to maximize validity/diversity outcomes).
2. Using a PSM alone, replacing the MAGE standard entirely (e.g., PSM ≥ 65).
3. Reducing the operational MAGE standard (e.g., MAGE-A ≥ 70), to evaluate if simply lowering the existing operational cut-score achieved more favorable validity/diversity outcomes than implementing a PSM.
4. For a few AFSCs where the incremental validity of PSMs was relatively small, we tested a more complex decision rule that required either: (a) the original MAGE cut-score or (b) a reduced MAGE cut-score paired with a PSM (e.g., MAGE-A ≥ 72 OR [MAGE-A ≥ 67 & PSM ≥ 65]). We tested this more complex decision rule to account for potential non-linear relationships between standards and training success that a non-compensatory model may better capture. We varied the candidate cut score tested by 20-30 percentile for each tested model. For example, when evaluating the “or” decision rule, we would test MAGE-A ≥ 72 or PSM $\geq 50 - 80$.

We estimated predicted FFPS probabilities and FSGs among applicants without using scores on the PSMs or the operational MAGE or AFOQT composites to mitigate the predicted FFPS rates or mean FSGs from favoring one or more of these composites. Specifically, the predicted FFPS regression equations for each SFSC/AFSC were estimated using a binary logistic regression model, while predicted FSG equations were estimated using an ordinary least squares regression model. Scores on the ASVAB subtests and TAPAS facets were used as predictors in the regression models. We used best subsets regression to mitigate overfitting in predictor selection, using the Akaike Information Criterion (AIC) to select the best model. Best subsets regression

iterates across all possible additive combinations of predictors, selecting the best subset of predictors for inclusion in the final model that minimizes prediction error (Furnival & Wilson, 2000). Accordingly, best subsets regression avoids the statistical issues associated with alternative methods, particularly stepwise or hierarchical regression. We then applied the SFSC/AFSC-specific regression equations to the enlisted applicants. For each applicant, we obtained a predicted probability (y-hat) of experiencing training success (FPPS) or a predicted FSG, for each SFSC/AFSC. For the models keyed to FPPS, we then identified a target first-pass rate for each SFSC/AFSC, which was calculated as the observed FPPS rate for trainees in the selected AFSC who met the operational qualification standard. Using this target first-pass rate, continuous predicted FPPS probabilities (y-hats) were rounded to dichotomous scores (0/1) such that the mean FPPS rate in the applicant sample matched the observed FPPS rate in the trainee sample at the current operational MAGE (or PSM) qualification standard. Stated differently, we adjusted the predicted FPPS probabilities to 0's (projected fails) and 1's (projected passes) such that the mean applicant predicted FPPS rate matched the observed FPPS rate at the current operational job qualification standards. We conducted this intercept adjustment to anchor predicted training success in the applicant sample to more closely approximate training success rates observed in the trainee sample for each selected SFSC/AFSC. Confidence intervals around the mean predicted FPPS estimates did not factor in prediction error, although they did factor in sampling error.

Once all scenarios had been implemented and all qualification metrics had been calculated, multiple analysts reviewed the projected impact output. We ultimately identified the most advantageous candidate standards to present as options for implementation after testing a wide range of standards for each SFSC/AFSC. Candidate standards were identified that showed: (a) a qualifying rate greater than or equal to the operational standard's qualifying rate and (b) a predicted training success level greater than or equal to the predicted success level observed with the operational standard's lower bound confidence interval (i.e., same predicted performance level as the operational standard, within an estimated margin of error). Identification of a single standard among the candidate standards typically involves trade-offs between qualifying rates (and AIRs) versus predicted training success levels. Also influencing this decision is whether there was a desire to pair a PSM with the primary MAGE standard or implement PSM-only qualification. To the extent possible, we identified both paired MAGE-PSM and PSM-only solutions for each AFSC. We also identified options at multiple qualifying rates to the extent possible. Overall, our focus was on identifying a range of candidate standards for each SFSC/AFSC to enable USSF flexibility when deciding which candidate standard(s) to implement, operationally, depending on their accession targets and preferences.

4.2 Enlisted Projected Impact Results

Attachment D documents the identified PSM-based standards for each AFSC. Figure 21 displays the predicted training success levels for candidate PSM-based standards (in light blue) across the targeted SFSC/AFSCs, using the operational standard (in red) as the baseline.¹⁴ Figure 21 shows

¹⁴ This AFSCs operational standard has a very stringent cut-score, specifically related to the requirement of EDPT $\geq$ 71. The PSM models tested did not match the predicted performance levels offered by the current operational standard unless PSM cut scores were set to the 99th percentile. Therefore, we took a different approach to

that all identified candidate standards were either at or above (or within the margin of error of) the training success rate set by the operational standard. In other words, our candidate standards match or exceed current observed training success rates. Specifically, many options have predicted FPPS rates 1%-5% higher than current FPPS rates.

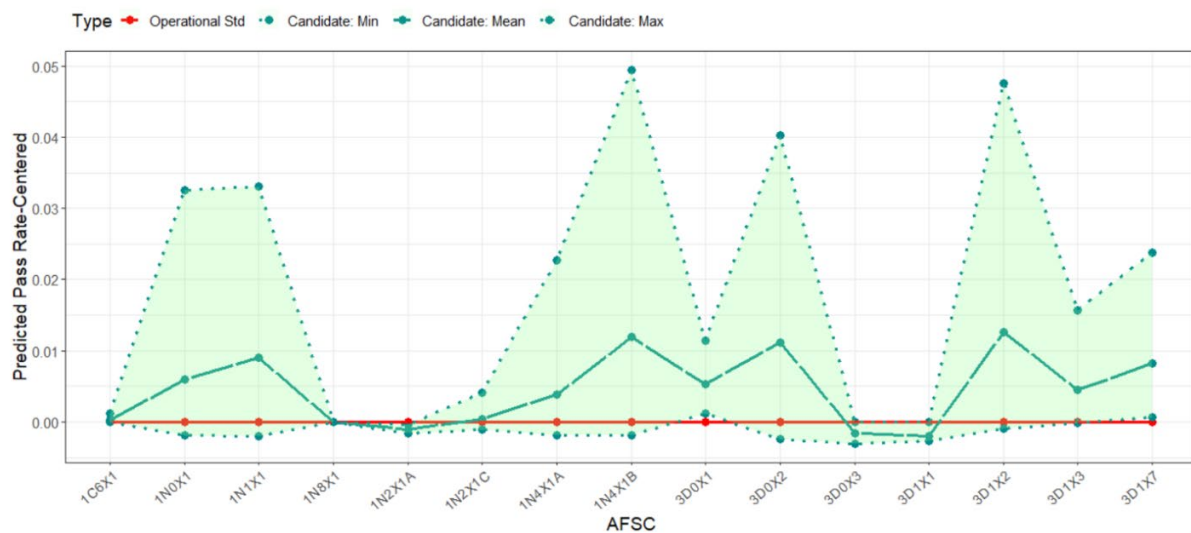
Figure 22 displays the overall qualifying rates by SFSC/AFSC, while Figure 23 displays qualifying rates for selected subgroups. Likewise, Figure 24 displays overall adverse impact ratios by SFSC/AFSC for selected subgroups. The minority group QRs were generally 5%-10% higher when using the candidate standards than the operational standard, on average, with some candidate standards producing minority QR improvements of 15% or greater. While the percentage differences may seem small, the differences translate into qualifying a larger number of female and minority group applicants over time. For example, a candidate standard for 3D1X2 (Cyber PSM 5 $\geq$ 47) results in a 6.6% increase in overall qualification rate (50.5% versus 57.1%), relative to the operational standard, while maintaining a similar level of predicted training success. This candidate standard results in a female qualification rate improvement of 22.4% (30.9% versus 53.3%), meaning over 11,000 more females would have been qualified to join this SFSC/AFSC between 2015-2021. The candidate standard also results in a Black qualification rate improvement of 18.1% (26.6% versus 44.7%), meaning 5,800 more Black applicants would have qualified to join this SFSC/AFSC between 2015 and 2021. Similarly, adverse impact ratio improvements were substantial—varying from small improvements (i.e., .01-.05 points) to over +.20 AIR improvement across SFSCs/AFSCs. For example, a candidate standard for 1N2X1C (Intelligence-Cryptologic PSM 4 $\geq$ 57) showed a similar overall QR relative to the operational standard (45.0% for the operational standard, 46.4% for the candidate standard). However, the candidate standard showed an improved AIR of +.30 for Female applicants (.73 versus 1.03) and +.15 for Black applicants (.54 versus .69). Overall, identified PSM-based options that were beneficial for one minority group were generally advantageous for the other groups as well. This is because PSMs were intentionally developed to increase the diversity of qualified applicants with respect to all minority groups with a sufficient sample size (Female, Black, Hispanic, Asian, and Multiracial).¹⁵

Notably, there were trade-offs among candidate standards. Specifically, candidate standards that showed the highest predicted training success rates typically were the most selective (i.e., higher cut-scores) and exhibited smaller diversity advantages. Conversely, the candidate standards with the largest improvements in subgroup QRs and AIRs typically showed smaller improvements in predicted training success rates. A range of potential solutions was identified for each SFSC/AFSC. Overall, the candidate standards identified offer improvements over the existing operational MAGE (or PSM) qualification standard(s), with some options weighted more heavily towards improving diversity, others weighted more heavily toward improving predicted training success rates and others balanced between the two. None of the candidate standards identified were expected to negatively impact training

identifying candidate standards. Namely, we asked: what if one was willing to accept a decrease in technical training pass rates to: (a) open the AFSC up to more applicants and more diverse applicants, and (b) offer an option to access this AFSC without taking the EDPT. Given that this AFSC followed a different model of options identified, we do not plot its results.

¹⁵ A minority of candidate standards contain lower QRs for the majority groups (Male, White).

success (within a small margin of error) or diversity, with respect to the current operational standards.



Notes. Performance is centered on the training success rate set by each AFSC’s operational standard (displayed by the red line). The range in blue reflects the short list of PSM-based candidate standards that were identified for each AFSC.

Figure 21. Mean Predicted Performance Across Enlisted AFSCs on Operational and Candidate Standards

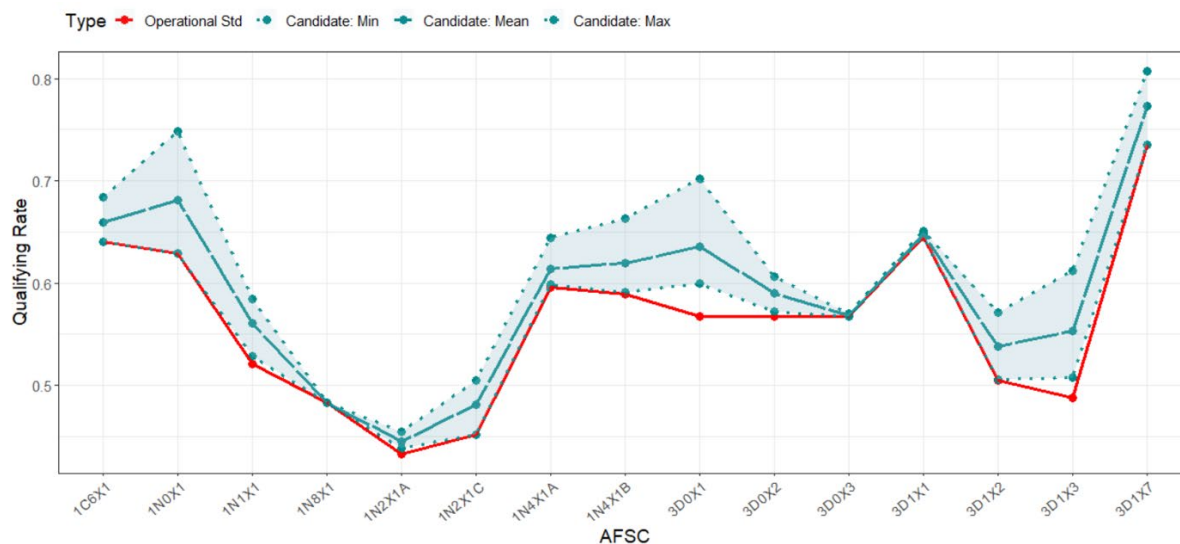


Figure 22. Mean Qualifying Rates Across Enlisted AFSCs on Operational and Candidate Standards

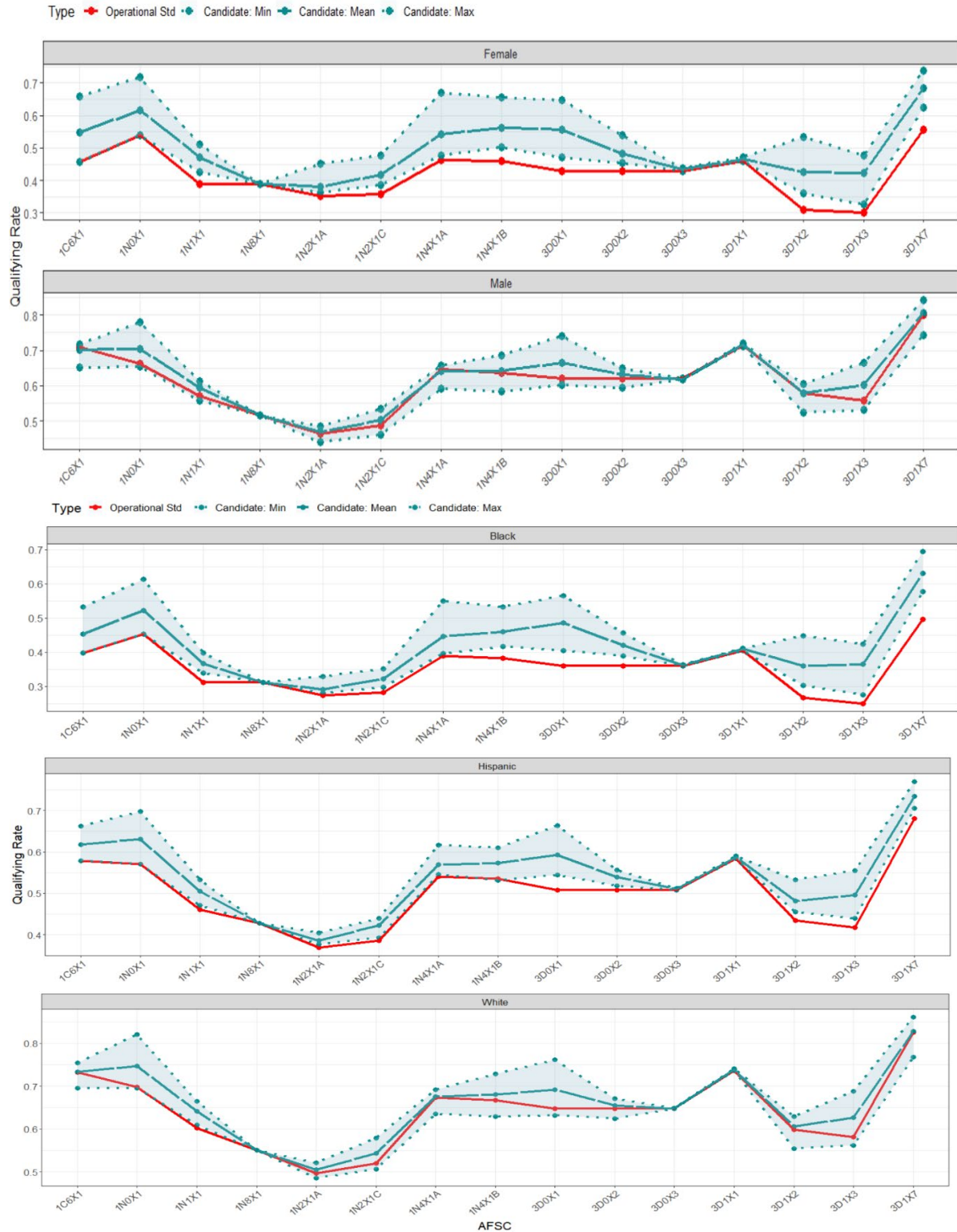


Figure 23. Mean Qualifying Rates by Subgroup Across Enlisted AFSCs on Operational and Candidate Standards

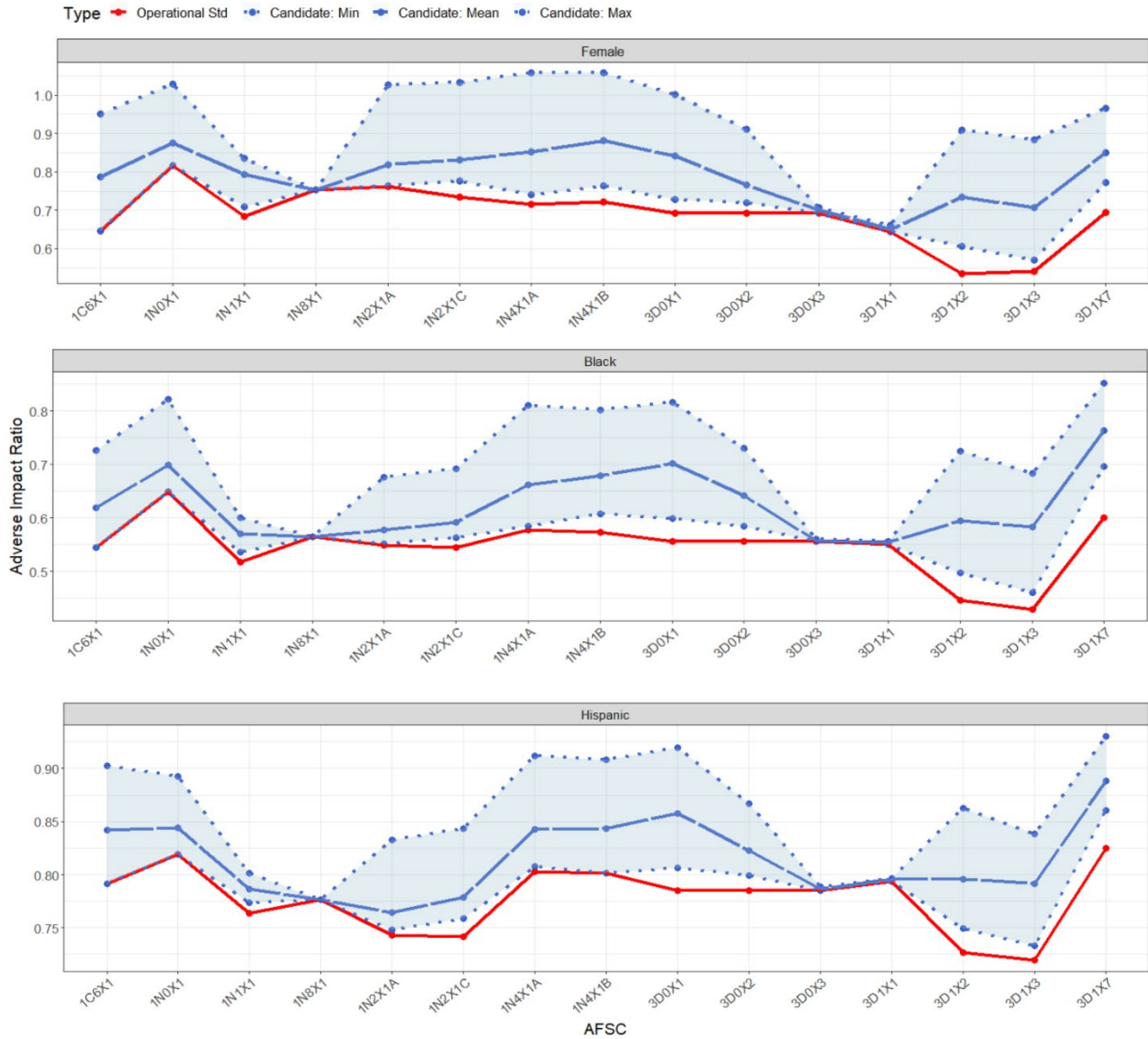


Figure 24. Mean Adverse Impact Ratios by Subgroup Across Enlisted AFSCs on Operational and Candidate Standards

To provide an illustration, we present candidate and operational standards for two example SFSCs/AFSCs. The first SFSC/AFSC is 1N4X1B/514X1B (Intelligence Analyst—Cryptologic Analysis & Reporting). Primary duties for this job involve analyzing and reporting intelligence information, performing analysis on target network communications, and producing intelligence for strategic, operational, and tactical customers (AFECD, 2022). Table 4 provides key metrics on qualification using the current operational MAGE (or PSM) standard versus four identified candidate standards.

For 1N4X1B/5I4X1B, two of the candidate standards use the PSM to supplement the MAGE (i.e., MAGE OR PSM). These two standards varied in terms of their overall selectivity—overall QRs are 61.1% and 62.2%. Both produced predicted FPPS rates that were comparable to the current FPPS rate. The PSM-only solutions achieved even higher QRs at the same or higher training success rate. Effectively, the higher predictive validity of the PSM permits lowering the qualifying standard, enabling more applicants to qualify for the SFSC/AFSC without negatively impacting predicted training success. Notably, the PSM-only solutions offered substantially higher Female QRs (+15.7%, +19.6%) and Black QRs (+10.8%, +15.0%), relative to the operational standard. When scaled as adverse impact ratios, these differences were also substantial for Female (+.34, +.34) and Black (+.21, +.23) applicants, relative to the operational standard. The PSM-only solutions were also expected to have slightly higher training success rates than the operational standard. This finding was consistent with a general pattern of results across SFSCs/AFSCs—that PSM-only solutions typically offered more distinct validity-diversity advantages than paired PSM-MAGE solutions.

The second example SFSC/AFSC is 3D0X1/5C0X1 (Knowledge Management). The primary duties for this job consist of managing processes for organizations to coordinate and share information, assisting organizations with meeting statutory records management requirements, performing IT project management duties, and conducting defensive cyber operations (AFECD, 2022). See Table 5.

Similarly, for 3D0X1/5C0X1, two candidate standards use the PSM to supplement the MAGE (i.e., MAGE OR PSM). In contrast, two other candidate standards use the PSM to replace the MAGE (i.e., PSM only). The two joint MAGE-PSM standards varied by overall selectivity, with QRs of 62.0% and 70.0%. Both standards showed similar predicted FPPS rates. However, the less selective standard yielded substantial increases in Female (+21.8%) and Black (+20.6%) QRs. The two candidate PSM-only standards have overall QRs closer to the operational QR (60.2% and 62.5% for candidate standards; 56.8% or operational standards) but still generated notable QR and AIR improvements by gender and race-ethnicity. The candidate PSM-only solutions displayed higher Female (+17.1%, +19.7%) and Black (+14.8%, +17.5%) applicant QRs, relative to the operational MAGE standard. When scaled as AIRs, the PSM-only solutions also produced higher Female (+.31, +.31) and Black (+.24, +.26) applicant AIRs relative to the operational MAGE standard. For this SFSC/AFSC, the candidate PSM (Cyber PSM 8) showed lower gender and race-ethnicity score differences than the operational composite (MAGE-G); therefore, qualification on the PSM resulted in improved diversity outcomes relative to the operational composite.

Table 4. Candidate Standards for 1N4X1B/5I4X1B

Standard Type	Qualification Standard	Overall Qual. Rate (CI L/U-95 th)	Female Qual. Rate (AIR)	Black Qual. Rate (AIR)	Hispanic Qual. Rate (AIR)	Asian Qual. Rate (AIR)	Predicted FPPS Rate (CI L/U-95 th)	F1 Score	MCC
Operational Standard	MAGE-G $\geq$ 62	58.9% (58.7%, 59.1%)	46.0% (.72)	38.2% (.57)	53.5% (.80)	62.2% (.93)	.786 (.783, .788)	0.79	0.50
MAGE OR PSM	MAGE-G $\geq$ 62 OR Intelligence-C PSM 4 $\geq$ 73	61.1% (60.9%, 61.3%)	50.3% (.77)	41.7% (.61)	56.0% (.82)	65.0% (.95)	.786 (.783, .788)	0.81	0.52
MAGE OR PSM	MAGE-G $\geq$ 62 OR Intelligence-C PSM 4 $\geq$ 68	62.2% (61.9%, 62.4%)	52.4% (.80)	43.4% (.63)	57.2% (.83)	66.3% (.96)	.784 (.781, .786)	0.81	0.53
PSM Only	Intelligence-C PSM 4 $\geq$ 45	59.3% (59.1%, 59.6%)	61.7% (1.06)	49.0% (.78)	56.3% (.90)	65.5% (1.04)	.809 (.807, .812)	0.82	0.56
PSM Only	Intelligence-C PSM 4 $\geq$ 41	63.0% (62.8%, 63.2%)	65.6% (1.06)	53.2% (.80)	60.2% (.91)	68.4% (1.03)	.790 (.788, .792)	0.82	0.55

Notes. Shading is based on Candidate outcomes relative to the Operational standard. For QRs, light green = Candidate QRs are 5%-10% higher; medium green = 10%-20% higher. For FPPS rates, light green = Candidate FPPS rates are 1%–2% higher; medium green = 2%–5% higher. AIR = Adverse impact ratio. CI L/U= Confidence Interval Lower/Upper Bounds -95th = 95% confidence interval, lower and upper bounds. MCC = Matthew’s Correlation Coefficient. Intelligence-C = Intelligence-Cryptologic Analysts.

Table 5. Candidate Standards for 3D0X1/5C0X

Standard Type	Qualification Standard	Overall Qual. Rate (CI L/U-95 th)	Female Qual. Rate (AIR)	Black Qual. Rate (AIR)	Hispanic Qual. Rate (AIR)	Asian Qual. Rate (AIR)	Predicted FPPS Rate (CI L/U-95 th)	F1 Score	MCC
Operational Standard	MAGE-G $\geq$ 64 OR [MAGE-G $\geq$ 54 & Cyber $\geq$ 60]	56.8% (56.6%, 57.0%)	42.9% (.69)	35.9% (.56)	50.8% (.79)	60.8% (.94)	.917 (.915, .918)	0.70	0.02
MAGE OR PSM	MAGE-G $\geq$ 64 OR Cyber PSM 8 $\geq$ 64	62.0% (61.8%, 62.2%)	53.0% (.81)	44.7% (.65)	57.2% (.83)	67.3% (.98)	.920 (.918, .921)	0.74	0.04
MAGE OR PSM	MAGE-G $\geq$ 64 OR Cyber PSM 8 $\geq$ 47	70.0% (69.8%, 70.2%)	64.7% (.90)	56.5% (.75)	66.3% (.89)	75.9% (1.02)	.921 (.919, .922)	0.80	0.05
PSM Only	Cyber PSM 8 $\geq$ 44	60.2% (59.9%, 60.4%)	60.0% (1.00)	50.7% (.80)	57.6% (.91)	69.3% (1.10)	.928 (.927, .930)	0.74	0.07
PSM Only	Cyber PSM 8 $\geq$ 42	62.5% (62.3%, 62.8%)	62.6% (1.00)	53.4% (.82)	60.2% (.92)	71.1% (1.09)	.928 (.926, .929)	0.76	0.07

Notes. Shading is based on Candidate outcomes relative to the Operational standard. For QRs, light green = Candidate QRs are 5%-10% higher; medium green = 10%-20% higher. For FPPS rates, light green = Candidate FPPS rates are 1%–2% higher; medium green = 2%–5% higher. AIR = Adverse impact ratio. CI L/U-95th = 95% confidence interval, lower and upper bounds. MCC = Matthew’s Correlation Coefficient

4.3 Officer Projected Impact Approach

We followed the same general approach with officers, evaluating the impact of using candidate PSM-based scores among the officer applicant population. However, for officers, there were not operational job-specific qualification standards for which to benchmark potential candidate standards. Accordingly, we tested two different PSM-based score uses for officers.

1. Using PSM scores as a tool for career field matching. With this approach, we placed each officer applicant in the career field of their highest-scoring PSM for a given set of PSMs. For example, if an applicant scored at the 60th percentile on Cyber Ops PSM 2, 65th percentile on Intelligence Operations PSM 2, and 70th percentile on Space Ops PSM 2, they were placed in Space Operations. We note that this placement model may depart from observed placements due to officer applicant interests and career field production requirements. However, this analysis was intended to roughly simulate use of PSMs in career field matching; that is, an officer could be provided their PSM scores and suggested to go into the career field for which they have the highest potential.
2. Using PSM scores to screen in officers by career field at fixed qualifying rates. Absent firm qualifying rate targets, we worked to identify the best candidate standards at four fixed qualifying rates—75%, 80%, 85%, and 90%. Again, our goal for the evaluation of PSMs was to provide a range of potential options to allow USSF decisional flexibility.

We also examined three baseline conditions to compare the candidate standards against in lieu of current operational standards. First, we simulated the selection of all officer applicants into a given career field. Second, we simulated the selection of all officers meeting minimum AFOQT requirements into a given career field ($\text{AFOQT-Q} \geq 10$ & $\text{AFOQT-V} \geq 15$). Third, we tested the random placement of officers into each of the three career fields in equal proportions (i.e., 1/3 of officer applicants randomly placed in each career field). For the third condition, we repeated the random placement 100 times to avoid sampling error in any single random sample. The projected impact analyses for officers, including results for each of the baseline conditions described above, are presented in Attachment E.

The predicted training success metric was calibrated using a similar process as for enlisted (described on p. 36 above), with two key differences. First, training success was placed on the FSG metric. Since FSG is continuous, linear regression models were used within the best subsets regression framework. Second, for officers, predicted values were left as continuous predicted FSGs—not rounded to pass-fail. We adjusted the predicted FSG values by a constant for each career field, such that the mean FSG rate in the applicant sample matched the mean FSG rate for a given SFSC/AFSC. Like the enlisted analyses, this adjustment was used to calibrate mean predicted FSGs to match observed FSGs within SFSCs/AFSCs.

4.4 Officer Projected Impact Results

Table 6 displays projected impact results when officers were placed into their highest-scoring career field (according to PSM Set 4), benchmarked against random placement into the career field. Table 7 shows projected impact results for an example career field, with Space Operations

at a fixed qualifying rate. Attention is first focused on the projected impact analyses evaluating PSM use in career field matching (i.e., Table 6). We tested several PSM solutions (Pareto solutions 2, 4, 6, and 8) and ultimately found PSM 4 to be the most advantageous, yielding training success improvements beyond random placement while preserving AIRs near 1.0. Specifically, placement using PSM 4 resulted in higher FSG rates of $d = .13$, $.25$, and $.23$ relative to random placement for the Cyber, Intelligence, and Space Operations career fields, respectively.¹⁶ Using PSM 4, AIRs ranged from $.88$ to $.99$ (Cyber Operations), 1.00 to 1.20 (Intelligence Operations), and $.79$ to 1.11 (Space Operations). In summary, the candidate PSMs produced a modest advantage in expected training success while also pathing females and minority group members to each of the career fields at levels equal or close to that of males and majority group members.

With respect to the fixed qualifying rate analyses, the standards highlighted in Table 7 have QRs of 85% except for the two baseline conditions (all officers meeting AFOQT-Q and AFOQT-V standards, random career field placement). The two baseline conditions are presented to benchmark predicted training success levels. Officer applicants qualified using the PSMs performed slightly better than the All-Officers sample, with d 's ranging from $.12$ to $.21$. PSM-based standards produced AIRs that ranged from $.77$ to 1.03 , with Female, Hispanic, and Asian AIRs all above $.90$. Overall, the candidate PSM-based standards offered slightly different trade-offs with respect to validity and diversity. For example, the standard using Space Operations PSM 9 offered improved diversity-related outcomes, whereas the standard using Space Operations PSM 4 offered improved training success. Also of note is that a standard with AFOQT-V ≥ 30 performed comparably to the PSMs with respect to predicted FSG rates, QRs, and AIRs, suggesting the potential use of the AFOQT-V composite in career field-specific qualification for Space Operations.

We note that several of the candidate PSMs presented in these analyses displayed non-negligible validity reductions with respect to the regression-weighted option (Pareto solution #1) – as highlighted in the officer Pareto curve figures 15, 17, and 19. For example, for Cyber Operations, candidate PSMs #2 and #9 have a $.20$ difference in estimated validity ($ps = .54$ and $.34$). However, within the projected impact analyses, estimated differences in mean FSG rates are substantially smaller. For example, comparing the selection of candidate PSM #2 versus #9 at a fixed Qualifying Rate of 80% produces a predicted mean FSG of 92.8 (Pareto #2) versus 92.5 (Pareto #9). When scaled as predicted FSG, this difference is quite small—three-tenths of a grade percentage or a 0.3% decrease. The smaller differences between Pareto solutions for officers were found because (a) the base rate of training success (FSG) is high for officers, (b) variability in training success is relatively low, and (c) qualifying rates for officers are relatively high, resulting in smaller proportions of officers being screened out than enlisted personnel. In essence, training success rates for officers are less sensitive to changes in predictor composites than they would be if the base rate of training success was lower and more screening of applicants was applied.

Projected impact results for the remaining two officer career fields are presented in Attachment E. Overall, PSM-based standards showed smaller joint validity-diversity advantages for officers than for enlisted. However, PSM solutions were still identified that produced modest

¹⁶ We note that in raw FSG rates, the difference between PSM-based placement and random placement is relatively small (less than 0.5 percentage points). However, variability on FSG rates is more limited for officers. This is why we chose to highlight results in a standardized metric, d -values.

improvements in technical training performance, while opening the career fields to a more diverse pool of applicants than when not using PSM scores. As illustrated within the Space Operations results, the AFOQT-V composite results also displayed favorable outcomes. However, one implementation consideration is that using PSMs that were customized by career field would ultimately enable more officers to access them than would use of a general composite (i.e., AFOQT-Verb) across career fields. While scores on career field-specific PSMs were interrelated, the correlations were not perfect, meaning there were career field rank differences among applicants (Spearman's r 's for PSM 4 ranging from .77-.85). Operationally, this means that officers that did not meet qualification standards for one career field may qualify for another, different career field. For example, when screening the officer applicant population at an 85% QR on PSM 4, 31%–44% of applicants screened out for one career field qualified to serve in another career field. Overall, even though advantages were not as distinct for officer PSMs, officer PSMs demonstrated potential for improving the matching or qualification of officer applicants to career fields.

Table 6. Officer Projected Impact – Career Field Matching for Space Operations

Career Field	Standard Type	Qualification Standard	Overall Qual. Rate (CI L/U-95th)	Female Qual. Rate (AIR)	Black Qual. Rate (AIR)	Hispanic Qual. Rate (AIR)	Asian Qual. Rate (AIR)	Predicted FSG Rate (CI L/U-95th)	FSG d-value: Candidate vs Full Sample
Cyber Operations	Benchmark	Random Placement into CF	33.4% (32.9% - 33.8%)	33.4% (1.00)	33.2% (1.00)	33.4% (1.00)	33.3% (1.00)	92.20 (92.17 - 92.24)	0.00
Cyber Operations	CF Matching	Cyber Ops PSM 4	33.9% (33.5% - 34.4%)	33.7% (0.99)	33.4% (0.97)	32.4% (0.94)	30.2% (0.88)	92.46 (92.42 - 92.49)	0.13
Intelligence Operations	Benchmark	Random Placement into CF	33.3% (32.8% - 33.8%)	33.3% (1.00)	33.2% (1.00)	33.4% (1.00)	33.4% (1.00)	94.96 (94.94 - 94.98)	0.00
Intelligence Operations	CF Matching	Intelligence Ops PSM 4	37.7% (37.2% - 38.2%)	43.1% (1.20)	42.3% (1.13)	37.4% (1.00)	38.7% (1.03)	95.28 (95.26 - 95.30)	0.25
Space Operations	Benchmark	Random Placement into CF	33.3% (32.9% - 33.8%)	33.3% (1.00)	33.6% (1.01)	33.2% (1.00)	33.3% (1.00)	96.51 (96.48 - 96.53)	0.00
Space Operations	CF Matching	Space Ops PSM 4	33.3% (32.8% - 33.8%)	27.6% (0.79)	27.5% (0.82)	34.4% (1.03)	37.1% (1.11)	96.80 (96.77 - 96.82)	0.23

Notes. CI L/U-95th = 95% confidence interval, lower and upper bounds. FSG d-value = Cohen's *d* comparing FSG between sample passing candidate standard and full AFOQT-passing sample (AFOQT-Q: 10 & AFOQT-V: 15). Higher values indicate higher FSG values for candidate standard's sample. Negative values indicate higher FSG values for the full AFOQT-passing sample.

Table 7. Officer Projected Impact – Fixed QR for Space Operations

Career Field	Standard Type	Qualification Standard	Overall Qual. Rate (CI L/U-95th)	Female Qual. Rate (AIR)	Black Qual. Rate (AIR)	Hispanic Qual. Rate (AIR)	Asian Qual. Rate (AIR)	Predicted FSG Rate (CI L/U-95th)	FSG d-value: Candidate vs Full Sample
Cyber Operations	Benchmark	Random Placement into CF	33.4% (32.9% - 33.8%)	33.4% (1.00)	33.2% (1.00)	33.4% (1.00)	33.3% (1.00)	92.20 (92.17 - 92.24)	0.00
Cyber Operations	CF Matching	Cyber Ops PSM 4	33.9% (33.5% - 34.4%)	33.7% (0.99)	33.4% (0.97)	32.4% (0.94)	30.2% (0.88)	92.46 (92.42 - 92.49)	0.13
Intelligence Operations	Benchmark	Random Placement into CF	33.3% (32.8% - 33.8%)	33.3% (1.00)	33.2% (1.00)	33.4% (1.00)	33.4% (1.00)	94.96 (94.94 - 94.98)	0.00
Intelligence Operations	CF Matching	Intelligence Ops PSM 4	37.7% (37.2% - 38.2%)	43.1% (1.20)	42.3% (1.13)	37.4% (1.00)	38.7% (1.03)	95.28 (95.26 - 95.30)	0.25
Space Operations	Benchmark	Random Placement into CF	33.3% (32.9% - 33.8%)	33.3% (1.00)	33.6% (1.01)	33.2% (1.00)	33.3% (1.00)	96.51 (96.48 - 96.53)	0.00
Space Operations	CF Matching	Space Ops PSM 4	33.3% (32.8% - 33.8%)	27.6% (0.79)	27.5% (0.82)	34.4% (1.03)	37.1% (1.11)	96.80 (96.77 - 96.82)	0.23

Notes. CI L/U-95th = 95% confidence interval, lower and upper bounds. FSG d-value = Cohen's *d* comparing FSG between sample passing candidate standard and full AFOQT-passing sample (AFOQT-Q: 10 & AFOQT-V: 15). Higher values indicate higher FSG values for candidate standard's sample. Negative values indicate higher FSG values for the full AFOQT-passing sample.

5.0 DISCUSSION

The primary objective of the current research was to develop and evaluate new PSMs to qualify USSF applicants for spacepower disciplines. The current research was completed as follows.

First, we identified a performance taxonomy that was aligned with job requirements for the targeted USAF/USSF jobs to inform PSM model development and for consideration in future performance measurement. The Air Force-Work Interest Navigator Work Activities (Johnson et al., 2020) were selected because they represent a wide range of specific and measurable performance behaviors that vary across jobs and have been previously used to profile enlisted USAF career fields. We recommended additions and refinements to the Work Activities focused on tailoring performance measurement to the context of USSF job requirements.

Second, we prepared data on enlisted and officer trainee and applicant samples to develop and evaluate the PSMs (entrance testing data on cognitive aptitudes and personality attributes, demographic data, and technical training for trainee samples only). We constructed criterion measures from the trainees' technical training course pipeline data: (a) whether trainees were successful in their first attempt through an AFSC/SFSCs technical training pipeline (pass/fail), and (b) trainees' FSG from the awarding course of the technical training pipeline. These criterion measures were used to develop multi-objective PSMs that combined scores on cognitive and personality tests, to optimize predicting technical training success (predictive validity) in combination with reducing subgroup differences that negatively impact the diversity of qualified applicants. Multi-objective PSMs were developed by career fields separately for enlisted and officers.

Third, we evaluated the newly developed PSMs against existing qualification composites that use scores on cognitive aptitude tests only (ASVAB for enlisted, AFOQT for officers). Overall, the newly developed PSMs showed advantages over existing qualification composites in predicting technical training success, as well as reduction of gender and racial-ethnic group differences on resulting PSM scores. The advantages of the new PSMs over existing qualification composites varied by enlisted career field. For many enlisted AFSCs, PSMs improved predicted training success and reduced gender and racial-ethnic group score differences over existing qualification composites, whereas for other enlisted AFSCs the new PSMs reduced gender and racial-ethnic group score differences with minimal to no loss in predicted training success compared to existing qualification composites. For officers, in several cases, existing qualification composites rivaled the performance of the PSMs. Overall, the advantages of PSMs were more limited for officers relative to existing qualification composites.

In summary, multi-objective PSMs represent a promising approach to USSF job qualification because (a) the PSMs incorporate scores on personality measures in addition to scores on cognitive tests, and (b) PSMs were developed using a multi-objective optimization technique, Pareto optimization, that generates solutions that optimize training success and diversity, in combination. A short list of candidate PSM-based qualification standards was provided for all targeted enlisted and officer AFSCs/SFSCs and career fields. The PSM solutions identified provided a range of qualification standards to allow decisional flexibility based on USSF's accession targets and preferences.

One important consideration for operational implementation is the generalizability or cross-validity of the candidate PSMs to new applicant samples. Whereas the MAGE composites have been subsequently validated multiple times (e.g., Carretta, 2014; Diaz & Ingerick, 2008), the PSMs developed here have not been cross-validated in a new sample. For the current research, we adopted a recently developed technique from Song et al. (2023) that estimates shrinkage across entire Pareto frontiers to estimate the level of prediction for PSMs in future samples. We conducted a simulation to evaluate how Song's formula-based shrinkage adjustments compared to empirically cross-validated prediction estimates via a holdout sample and found Song's shrinkage adjustments to be accurate (within an estimated R of $< .01$). Diversity shrinkage was estimated to be extremely small in magnitude because the subgroup difference estimates used as input to modeling came from large USAF applicant populations. Validity shrinkage was mostly small in magnitude ($\leq .03$ for FPPS, $\leq .05$ for FSG), and was smallest in the enlisted career fields with large trainee samples. Most importantly, the shrinkage adjustments did not significantly change the conclusions made regarding the effectiveness of PSMs in relation to existing qualification composites.

Relatedly, we implemented steps in the candidate PSMs to reduce shrinkage and increase the cross-validation potential of PSMs in new or future samples as much as possible. These steps included (a) rounding target weights to integer weights (rounded to the nearest 0.5) to avoid over-capitalizing on sample-specific variance, (b) using population-level information to inform the Pareto modeling procedure to the extent possible, to maximize sample size and stability of estimates (i.e., predictor intercorrelations, subgroup difference estimates), and (c) estimating PSMs for career fields rather than individual jobs, making PSMs less sensitive to variability observed in individual jobs with smaller samples. Nevertheless, monitoring and conducting a follow-up test and evaluation post-implementation of the candidate PSMs is advised. Doing so is consistent with what USSF and USAF presently do with the current operational qualification standards.

Accordingly, based on the findings from the current work, the following recommendations are made:

Enlisted

1. Implement one or more of the candidate PSMs identified in the projected impact evaluation for the targeted spacepower disciplines operationally, in place of or as an additional compensatory standard to the current operational MAGE standard(s).
2. Monitor and conduct a follow-up test and evaluation of the implemented PSM candidates once sufficient data accumulates within USSF.
3. Use the results of the follow-up test and evaluation to refine the implemented PSM candidates as needed.

Officers

1. Revisit and replicate the current analyses to identify candidate PSMs for use in matching officers to spacepower disciplines after proposed refinements to the AFOQT have been completed.
2. Implement one or more of the candidate PSMs as a tool to match officers to spacepower disciplines operationally.
3. Conduct a follow-up test and evaluation of the implemented PSM candidates once sufficient data accumulates.
4. Use the results of the follow-up test and evaluation to refine the implemented PSM candidates as needed.

When moving to the operational implementation of PSMs, an additional important consideration is applicant faking on the personality assessments. The TAPAS is designed to minimize faking by presenting test-takers with two statements matched on social desirability levels (Stark et al., 2014), which makes the selection of a desirable response more challenging. Kantrowitz et al. (2019) analyzed Air Force enlisted trainee data collected for research purposes only, in which test-takers were instructed to either respond honestly ($N = 576$) or to “fake good” (i.e., provide socially desirable responses; $N = 513$). There were not notable differences between honest and fake-good responders on most TAPAS facets under these experimental, instructed response conditions (45% of d 's $\leq .10$, 33% $\leq .20$, and the remaining 22% fell between .30-.35). Small negative effects were observed on a few facets where honest test-takers scored slightly lower than fake-good respondents. Trent et al. (2020) found similar TAPAS facet validities on counterproductive work behaviors in honest and directed faking conditions, with validities slightly stronger in the honest condition, on average (mean $r_{\text{Honest} - \text{Fake Good}} = -.03$, when scales were coded to predict CWB negatively). In summary, this finding indicates that the predictive power of the TAPAS is generally resistant to faking effects, with specific findings of comparable to slightly smaller validities under fake good conditions. However, the SDI was not designed to be resistant to faking, which represents an important consideration for the operational implementation of the SDI. The SDI uses Likert-style responses, meaning that there is a lower cognitive burden to fake good. The current SDI data was collected for research purposes only; therefore, we expected minimal impact of faking on validity and subgroup differences. However, when considering a shift to the operational implementation of the SDI, an important consideration for future research is the fakeability of this assessment. Kantrowitz et al. (2019) analyze the same sample of Air Force trainees directed to respond honestly versus directed-faking, finding substantially larger differences between honest and fake good conditions (mean $d = .51$, $SD = .26$). However, we note that the SDI is being proposed for use in job classification, rather than selection into the USAF or USSF. Therefore, there is good justification for applicants to respond honestly—so that they can be guided to career fields that are most suitable for their personalities.

The present research also carries technical and conceptual implications for research in personnel selection and military job classification broadly. From a technical standpoint, this research further evidences the usefulness of multi-objective optimization techniques like Pareto optimization in personnel selection and classification contexts. The current research also utilized

a new tool developed for Pareto optimization on dichotomous criteria (De Corte, 2022). An ongoing challenge has been that range restriction impacts the estimates from models keyed on dichotomous criteria because said models require individual-level data. However, this new tool enabled the current research to account for range restriction when developing models on dichotomous criteria, developing PSMs that reflect data and inferences to be made on the USAF applicant population. Conceptually, the present research demonstrated the potential usefulness of personality attributes when developing job qualification models designed to open up qualifications to more diverse pools of qualified applicants.

6.0 CONCLUSION

The present research developed and evaluated multi-objective PSMs to qualify enlisted and officer applicants for USAF/USSF jobs. Overall, PSMs showed improvements over existing qualification standards in predicting technical training success of qualified applicants, combined with opening up qualifications to a larger and more diverse pool of applicants with the right talents for success. The advantages of PSMs over existing qualification standards were strongest for enlisted personnel. Analyses highlighted specific improvements in job qualification rates and improved diversity by gender and race-ethnicity among qualified applicants when using PSMs in job qualification, compared to current operational standards.

7.0 REFERENCES

- AFECD. (2022). *Air Force Enlisted Classification Directory (AFECD): The official guide to the Air Force enlisted classification codes*. United States Air Force.
- AFOCD. (2022). *Air force Officer Classification Directory (AFOCD): The official guide to the Air Force officer classification codes*. United States Air Force.
- Aitken, A. C. (1934). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society*, 4(2), 106–110.
- Barelka, A., Barron, L., Coggins, M., Hernandez, S., & Kulpa, P. (2019). *Development and validation of Air Force foundational competency model* (No. 1). United States Air Force.
- 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). 711 Human Performance Wing, Air Force Research Laboratory.
- Campbell, J. P. (2012). Behavior, performance, and effectiveness in the twenty-first century. *The Oxford handbook of organizational psychology*, 1, 159-194.
- Carretta, T. R. (2014). *Predictive validity of the Armed Services Vocational Aptitude Battery for several US Air Force enlisted training specialties* (AFRL-RH-WP-TP-2014-0046). 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
- Diaz, T. E., & Ingerick, M. (2008). *Estimating Air Force specialty final school grades and washout rates* (AFCAPS-FR-2012-0007). Air Force Personnel Center: Strategic Research and Assessment Headquarters.
- De Corte, W. (2022). *Pareto optimization tool for dichotomous outcomes*. Author.
- De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology*, 92(5), 1380–1393.
- De Corte, W., Sackett, P. R., & Lievens, F. (2011). Designing pareto-optimal selection systems: Formalizing the decisions required for selection system development. *Journal of Applied Psychology*, 96(5), 907-926.
- Furnival, G. M., & Wilson, R. W. (2000). Regression by leaps and bounds. *Technometrics*, 16, 499-511.
- Hom, P. W., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). One hundred years of employee turnover theory and research. *Journal of Applied Psychology*, 102(3), 530-545.

- Johnson, J. F., Romy, S., & Barron, L. G. (2020). Air Force Work Interest Navigator (AF-WIN) to improve person-job match: Development, validation, and initial implementation. *Military Psychology*, 32(1), 111-126.
- Johnson, J. W., & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods*, 7(3), 238-257.
- 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. 711 Human Performance Wing, Airman Systems Directorate, Collaborative Interfaces and Training Branch.
- Kirkendall, C., Bynum, B., Nesbitt, C., & Hughes, M. (2020). Validation of the TAPAS for predicting in-unit soldier outcomes. *Military Psychology*, 32(1), 24-35.
- Lawley, D. N. (1944). IV.—A Note on Karl Pearson's selection formulæ. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 62(1), 28–30.
- Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, 17(4), 372-411.
- Nye, C. D., Drasgow, F., Chernyshenko, O. S., Stark, S., Kubisiak, U. C., White, L. A., & Jose, I. (2012). *Assessing the Tailored Adaptive Personality Assessment System (TAPAS) as an MOS qualification instrument* (No. 1312). United States Army Research Institute for the Behavioral and Social Sciences.
- Roberts, J. A., Arthur, W., Reckase, M., Sackett, P., & Zenisky, A. (2019). *TAPAS Evaluation Project report on the readiness of TAPAS for use in selection and classification of military applicants: Final report* (2019 No. 076). Human Resources Research Organization.
- Rupp, D., Song, Q., & Strah, N. (2020). Addressing the so-called validity–diversity trade-off: Exploring the practicalities and legal defensibility of Pareto-optimization for reducing adverse impact within personnel selection. *Industrial and Organizational Psychology*, 13(2), 246-271.
- Secretary of Defense. (2020). *Memorandum for senior Pentagon leadership commanders of the Combatant Commands Defense Agency and Department of Defense (DoD) field activity directors*. Author.
- Segall, D. O. (2004). *Development and evaluation of the 1997 ASVAB score scale*. Defense Manpower Data Center.

- Song, Q., Wee, S., & Newman, D. A. (2017). Diversity shrinkage: Cross-validating pareto-optimal weights to enhance diversity via hiring practices. *Journal of Applied Psychology*, 102(12), 1636-1657.
- Song, Q., Tang, C., Newman, D. A., & Wee, S. (2023). Adverse impact reduction and job performance optimization via pareto-optimal weighting: A shrinkage formula and regularization technique using machine learning. *Journal of Applied Psychology*. Advance online publication. <https://doi.org/10.1037/apl0001085>
- Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology*, 26(3), 153-164.
- Trent, J. D., Barron, L. G., Rose, M. R., & Carretta, T. R. (2020). Tailored Adaptive Personality Assessment System (TAPAS) as an indicator for counterproductive work behavior: Comparing validity in applicant, honest, and directed faking conditions. *Military Psychology*, 32(1), 51-59.
- U.S. Space Force (2020). *Spacepower: Doctrine for Space Forces*. Author.
- U.S. Space Force. (2021). *Guardian Ideal*. <https://media.defense.gov/2021/Sep/21/2002858512/-1/-1/1/GUARDIAN%20IDEAL%20-%20FINAL.PDF>
- Walsh, J. L., Woolley, M. R., Brady, M. F., Melick, S. R., & Carretta, T. R. (2021). *Air Force Officer Qualifying Test (AFOQT) Form T Evaluation: The Situational Judgment Test* (AFRL-RH-WP-TR-2021-0080). 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
- Wee, S., Newman, D. A., & Joseph, D. L. (2014). More than g: Selection quality and adverse impact implications of considering second-stratum cognitive abilities. *Journal of Applied Psychology*, 99(4), 547-563.
- Weissmueller, J. J., & Schwartz, K. L. (2007). *Self-Description Inventory plus initiative: Assault on Occam's Razor* (AFCAPS-TR-2012-0003). Air Force Personnel Center: Strategic Research and Assessment Branch.
- Woolley, M. R., Walsh, J. L., Carretta, T. R., Moulon, A. M., & Deregla, A. R. (in press). *Development and psychometric evaluation of predictive success models (PSMs) for US Air Force rated career fields*, AFRL-RH-WP-TR-2023-xxxx. 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
- Woolley, M. R., Walsh, J. L., Mann, K. J., Wilson, R., Carretta, T. R., Moulon, A.N., & Deregla, R. (2022). *Self-Description Inventory-Officer (SDI-O): Item-, facet-, and domain level analyses*, AFRL-RH-WP-TR-2022-0091. 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

AFECD	Air Force Enlisted Classification Directory
AFOCD	Air Force Officer Classification Directory
AFOQT	Air Force Officer Qualifying Test
AFOQT-Q	AFOQT Quantitative Composite
AFOQT-V	AFOQT Verbal Composite
AFQT	Armed Forces Qualification Test
AFSC	Air Force Specialty Code
AF-WIN	Air Force Work Interest Navigator
AI	Adverse Impact
AIR	Adverse Impact Ratio
ASVAB	Armed Services Vocational Aptitude Battery
CI L/U	Confidence Interval Lower/Upper Bounds
<i>d</i>	Cohen's Standardized Mean Score Difference
DoD	Department of Defense
EDPT	Electronic Data Processing Test
FIML	Full Information Maximum Likelihood
FN	False Negative
FP	False Positive
FPPS	First-Pass Pipeline Success
FSG	Final School Grade
HumRRO	Human Resources Research Organization
KSAOs	Knowledge, Skills, Abilities, and Other Characteristics
M	Mean
MAGE	Mechanical, Administrative, General, and Electronics Composites
MAGE-A	MAGE-Administrative
MAGE-E	MAGE-Electronics
MAGE-G	MAGE-General
MAGE M	MAGE-Mechanical
MCC	Matthew's Correlation Coefficient

MI	Multiple Imputation
MVRR	Multivariate Range Restriction
PAY	Profile of American Youth
PO	Pareto-Optimal
PSM	Predictive Success Model
QR	Qualification Rate
SD	Standard Deviation
SDI	Self-Description Inventory
SFSC	Space Force Specialty Code
SME	Subject Matter Expert
SSO	Space Systems Operations
STEM	Science, Technology, Engineering, and Mathematics
TAPAS	Tailored Adaptive Personality Assessment System
TN	True Negative
TP	True Positive
USAF	United States Air Force
USSF	United States Space Force