



AFRL-RH-WP-TR-2023-0014

**TAILORED ADAPTIVE PERSONALITY ASSESSMENT
SYSTEM (TAPAS)
PRE-IMPLEMENTATION DOCUMENTATION**

**Fritz Drasgow
Oleksandr S. Chernyshenko
Stephen Stark
Christopher D. Nye
Drasgow Consulting Group**

April 2023

Interim Report

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

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

NOTICE AND SIGNATURE PAGE

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

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

AFRL-RH-WP-TR-2023-0014 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION.

//signature//

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

//signature//

LOGAN A. WILLIAMS, DR-III, PhD
Human Performance Product Area Lead
Operational Product Section
Product Development Branch
Airman 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, 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 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) 18-04-23		2. REPORT TYPE Interim		3. DATES COVERED (From - To) 13 July 2021 to 3 Nov 2022	
4. TITLE AND SUBTITLE Tailored Adaptive Personality Assessment System (TAPAS) Pre-Implementation Documentation				5a. CONTRACT NUMBER FA8650-21-F-4104	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Fritz Drasgow, Oleksandr S. Chernyshenko, Stephen Stark, and Christopher D. Nye				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER H12Q	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Drasgow Consulting Group (DCG) 3508 N. Highcross Rd. Urbana, IL 61802				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Materiel Command Air Force Research Laboratory 711 th Human Performance Wing Airman Systems Directorate Airman 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-2023-0014	
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution Statement A: Approved for public release. Distribution unlimited.					
13. SUPPLEMENTARY NOTES Report contains color. AFRL-2023-2122, cleared 2 May 2023.					
14. ABSTRACT The Tailored Adaptive Personality Assessment System (TAPAS) was originally developed by the Drasgow Consulting Group (DCG) under the Army's Small Business Innovation Research (SBIR) grant program with work beginning in 2004. As with any new assessment, careful documentation is needed; this report provides some of the necessary documentation. Included are the history of TAPAS and TAPAS testing, a description of the facets, composites, configurable specifications and details of the adaptive algorithm, scoring and norming, reliability and conditional standard errors.					
15. SUBJECT TERMS Enlistment testing, personality assessment, item response theory, multi-unidimensional pairwise preference model, two-alternative forced choice, computer adaptive testing, marginal reliability, conditional standard error					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT: SAR	18. NUMBER OF PAGES 144	19a. NAME OF RESPONSIBLE PERSON (Monitor) Thomas R. Carretta 19b. TELEPHONE NUMBER (Include Area Code) NA
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			

TABLE OF CONTENTS

ACKNOWLEDGEMENT	v
EXECUTIVE SUMMARY	1
1.0 INTRODUCTION	3
2.0 BRIEF HISTORY OF THE TAPAS BEFORE MEPS IMPLEMENTATION	3
2.1 Reasons for TAPAS Development.....	4
2.2 TAPAS Initial Validation in U.S. Army: Expanded Enlistment Eligibility Metrics (EEEM) Research Project (2006-2009)	8
2.2.1 TAPAS-95s Description	8
2.2.2 TAPAS-95s Construct Validity Results.....	11
2.2.3 TAPAS-95s Criterion Validity Results.....	13
2.2.4 TAPAS-95s Adverse Impact Results.....	14
2.3 Summary and Conclusions from the Initial TAPAS Validation Efforts and Approval for Use at Military Entrance Processing Stations.....	15
3.0 TAPAS MEPS TESTING.....	16
3.1 Chronological Order of TAPAS Version Releases.....	16
3.1.1 Initial Deployment: Phase 1	16
3.1.2 New Forms with MEPS-Only Items: Phase 2.....	17
3.1.3 Reduced Dimensionality: Phase 3	18
3.1.4 New Versions: Phase 4	20
3.2 TAPAS MEPS Version Creation.....	23
3.3 TAPAS MEPS Test Administration Procedures.....	27
4.0 TAPAS FACETS.....	27
4.1 Initial 22-facet TAPAS Taxonomy	27
4.2 2011 Revision of TAPAS Facet Taxonomy	28
4.3 Addition of 6 New Experimental TAPAS Facets to Support Enhanced Suitability Screening	29
4.4 Summary	31
5.0 STATEMENT POOLS	32
5.1 General Steps for Statement Pool Development.....	32
5.2 Development of the Initial 22-facet TAPAS Research Statement Pool	33
5.3 Development of the 23-facet DoD Exclusive Statement Pool.....	37
5.4 Developing Statement Pools for Five New Experimental Facets and Virtue.....	39
5.5 Additional TAPAS Forms to Support DoD Personality Research Projects	41
5.6 Summary	46
6.0 TAPAS COMPOSITE SCORES	46
7.0 DETAILED DESCRIPTIONS OF TAPAS VERSION CONFIGURATION SPECIFICATIONS AND COMPUTER ADAPTIVE TESTING ALGORITHM.....	51
7.1 Specifying Facets, Statement Pools, and Test Blueprints.....	51
7.2 TAPAS CAT Administration Algorithm and Configurable Test Specification Variables	56
7.2.1 Logic of the TAPAS CAT Algorithm	56
7.2.2 Configurable Specifications for Item Construction and Selection Constraints.....	57
7.3 Mathematical Details for Calculating Item and Test Information Values	

During CAT Administration	60
8.0 TAPAS SCORING, NORMING, AND REPORTING TEST RESULTS AND DIAGNOSTICS	66
8.1 TAPAS Trait and Standard Error Estimation	66
8.2 TAPAS Norming and Raw Score Transformation	68
8.3 TAPAS Composite Scores	71
8.4 TAPAS Diagnostic Flags for Unmotivated Responding	72
8.5 Summary of TAPAS Test Results and Guidance on TAPAS Score Interpretation.....	77
9.0 RELIABILITY OF TAPAS SCALE SCORES ACROSS TAPAS MEPS VERSIONS.	78
10.0 TEST-RETEST RELIABILITY	96
10.1 Study 1	97
10.1.1 Method	97
10.1.2 Results	97
10.2 Study 2	98
10.2.1 Method	98
10.2.2 Results	99
10.3 Study 3	100
10.3.1 Method	100
10.3.2 Results	100
10.4 Study 4	102
10.4.1 Method	102
10.4.2 Results	102
10.5 Study 5	103
10.5.1 Method	103
10.5.2 Results	104
10.6 Discussion	105
11.0 SUMMARY AND DISCUSSION.....	106
12.0 REFERENCES	108
APPENDIX A: TAPAS MEPS IMPLEMENTATION AUTHORIZATION	116
APPENDIX B: TEST BLUEPRINT FOR TAPAS VERSION V9.....	118
APPENDIX C: EXAMPLE OF PRE-TEST FORM FOR ESTIMATING GGUM PARAMETERS OF TAPAS STATEMENTS	129
APPENDIX D: EXAMPLE OF PRE-TEST FORM FOR ESTIMATING SOCIAL DESRABILITY PARAMETERS OF TAPAS STATEMENTS	132
LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS	135

LIST OF FIGURES

Figure 1: APFT scores, 6-Month Attrition, Army Life Adjustment, and Disciplinary Incidents by AFQT category (from Nye et al., 2012b)	49
Figure 2: Example Specification for a Tolerance Statement	52
Figure 3: Example of V12 Test Blueprint Specifications for the Linking Subtest.....	55
Figure 4: Item Construction and Selection Constraints for V12.....	57
Figure 5: Example of an Item Administered on Version 12	60
Figure 6: MUPP Item Response Functions (IRFs) for an Item Composed of Stimuli Representing Different Facets, and Composed of Stimuli Representing the Same Facet	63
Figure 7: MUPP Item Information Functions (IIFs) for the Items Having IRFs shown in Figure 6	65
Figure 8: Example of TAPAS Score-Related Output for the Achievement Facet	69
Figure 9: Example of Composite Scores Produced by TAPAS Version 12	72
Figure 10: TAPAS Diagnostic Flags	73
Figure 11: Example of a Markov Chain Transition Matrix	75
Figure 12: Example of Response Check Item.....	77

LIST OF TABLES

Table 1. Initial 22 TAPAS Facet Taxonomy: Trait Names, Markers, and Descriptions	5
Table 2. TAPAS-95s Facet Descriptions and Basic Statistics	10
Table 3. Correlations between TAPAS-95s Facets and Selected Dimensions from the AIM, RBI, and ASVAB	12
Table 4. Incremental Validity Results for TAPAS-95s Facets and Four Training Criteria.....	14
Table 5. Subgroup Comparisons of AFQT and TAPAS-95s Scores	15
Table 6. Army Phase 1 Facet Summary.....	17
Table 7. Army Phase 2 Facet Summary.....	18
Table 8. Army Phase 3 Facet Summary.....	19
Table 9. Army Phase 4 Facet Summary.....	20
Table 10. TAPAS Versions by Administration Date and Sample Size	22
Table 11. TAPAS Facets Assessed at MEPS by Version	24
Table 12. 2011 Updates to the TAPAS Facet Taxonomy (DoD Exclusive Only)	29
Table 13. Description of Six New Experimental TAPAS Facets	31
Table 14. Breakdown of Samples Used to Estimate GGUM Parameters for TAPAS Statements	34
Table 15. Number of Statements Available for Each of the 22 TAPAS Facets	35
Table 16. Samples Used to Estimate GGUM Parameters for the Second TAPAS Statement Pool.....	37
Table 17. Numbers of Statements Representing Each of the 23 Facets in the Second	

TAPAS Statement Pool	39
Table 18. Numbers of Statements Representing Each of the Six New TAPAS Item Pools.....	40
Table 19. Additional TAPAS Forms and Facets Assessed.....	42
Table 20. Statement-Level Test Blueprint Specifications for Version 12	54
Table 21. Example Norm Table for TAPAS Raw Score Conversions	70
Table 22. IRT Reliabilities for TAPAS MEPS Versions.....	80
Table 23. Marginal Reliabilities and Conditional Standard Errors for Version 4	82
Table 24. Marginal Reliabilities and Conditional Standard Errors for Version 5	83
Table 25. Marginal Reliabilities and Conditional Standard Errors for Version 7	84
Table 26. Marginal Reliabilities and Conditional Standard Errors for Version 8	85
Table 27. Marginal Reliabilities and Conditional Standard Errors for Version 9	86
Table 28. Marginal Reliabilities and Conditional Standard Errors for Version 10	87
Table 29. Marginal Reliabilities and Conditional Standard Errors for Version 11	88
Table 30. Marginal Reliabilities and Conditional Standard Errors for Versions 12 13, and 14 (Part 1, 133 items)	89
Table 31. Marginal Reliabilities and Conditional Standard Errors for Version 12 (Parts 1 and 2, 176 Items).....	91
Table 32. Marginal Reliabilities and Conditional Standard Errors for Version 13 (Parts 1 and 2, 176 Items).....	92
Table 33. Marginal Reliabilities and Conditional Standard Errors for Version 14 (Parts 1 and 2, 176 Items).....	93
Table 34. Summary of Marginal Reliabilities Across All TAPAS MEPS Army Versions	95
Table 35. TAPAS Test-Retest Reliabilities in Study 1.....	98
Table 36. TAPAS Test-Reliabilities in Study 2.....	99
Table 37. TAPAS Test-Retest Reliabilities for the 10D Paper Form in Study 3.....	101
Table 38. TAPAS Test-Retest Reliabilities for the 10D Computerized Form in Study 3.....	101
Table 39. TAPAS Test-Reliabilities in Study 4.....	103
Table 40. TAPAS Test-Reliabilities in Study 5.....	104

ACKNOWLEDGEMENT

We would like to sincerely thank Thomas R. Carretta, PhD (Air Force Research Laboratory), Annette Rizer (Infoscitex), Bobbie Ann Dirr, PhD (Air Force Personnel Center's Strategic Research and Assessment Branch), and Matthew Palomo (Air Force Personnel Center's Strategic Research and Assessment Branch) for their help and valuable contributions to this work.

EXECUTIVE SUMMARY

Drasgow Consulting Group (DCG) began development of the Tailored Adaptive Personality Assessment System (TAPAS) in 2004 under the Army's SBIR grant program. Rooted in the Big Five theory of personality (RBI), TAPAS was designed to support military selection and classification decisions. Rather than the Big Five, TAPAS measures their underlying narrow facets that are important for predicting military performance. To date, more than two million military enlistment applicants have taken TAPAS.

To mitigate faking, TAPAS utilizes a two-alternative forced choice format where the options are balanced in social desirability. To reduce testing time, TAPAS is adaptive with statements drawn from large item pools, which have the additional benefit of reducing the possibility of test compromise. Importantly, adaptive testing yields more precision in scores with fewer items.

As with any new assessment, careful documentation is needed. A blue-ribbon panel (Roberts, Arthur, Reckase, Sackett, & Zenisky, 2019) detailed required documentation. This report provides some of the necessary documentation.

We begin with the history of the TAPAS. The initial phase of work produced a conceptual model consisting of 22 facets that underlie and constitute the Big Five personality dimensions. The initial version of the assessment instrument, TAPAS-95s, measured only 12 personality facets (not the complete conceptual model). Researchers collected predictive and construct-related validity for TAPAS-95s during the U.S. Army Expanded Enlistment Eligibility Metrics (EEEM) research project between 2006 and 2009. Key findings from this work included substantial improvements in the prediction of 6-month attrition, Army Physical Fitness Test (APFT) scores, and adjustment to Army life. Importantly, gender differences and racial/ethnic differences were much smaller for TAPAS-95s facets than for the Armed Forces Qualification Test (AFQT).

In May 2009, testing started at Military Entrance Processing Stations (MEPS). Since then, more than a dozen versions have been deployed at MEPS, with all but one version being adaptive. Each MEPS version measures between 13 and 19 facets. The Army, Air Force, and Marine Corps are currently administering TAPAS at MEPS. In addition to MEPS versions, there are several other forms to support Department of Defense (DoD) personality research.

This report provides a description of the original 22 facets in the conceptual model, 5 additional facets added in 2011, and 5 more facets added in the 2015-2017 time period. For each facet, a large number of statements were written, edited, and pretested. Ultimately, pools of 40 to 60 statements per facet are used during adaptive testing.

As with the Armed Services Vocational Aptitude Battery (ASVAB), career decisions should be based on composites rather than individual facets. Army composites have been developed to predict Can-Do performance, Will-Do performance, Attrition/Adaptation (both terms have been used), and misconduct attrition. Additionally, the Air Force creates Predictive Success Model (PSM) composites that are tailored to certain Air Force Specialty Codes (AFSCs).

This report also contains a detailed description of configurable specifications and the adaptive algorithm. There are many constraints governing the construction and selection of items in TAPAS tests to ensure accurate measurement for all test takers.

We also describe TAPAS scoring, norming, and score diagnostics. Scoring is based on item response theory (IRT), and specifically the multi-unidimensional pairwise preference (MUPP) model. Norming is a critical step that ensures that scores can be interpreted easily. TAPAS also computes several diagnostics to identify individuals who may not be responding effortfully.

The three sections following the scoring section provide information on the reliability, conditional standard errors of scores, and test-retest correlations. Because TAPAS is adaptive, traditional reliability measures such as coefficient alpha cannot be computed. Instead, marginal reliability is computed based on IRT. In addition, IRT was used to compute conditional standard errors, which characterize measurement precision throughout the trait continuum. We show that the adaptive algorithm of TAPAS yields similar measurement precision across trait values. We review five studies that examined test-retest correlations.

In sum, this report provides detailed information about the origin of TAPAS, the facets it measures, how it measures the facets, the process of computing facet and composite scores, and the accuracy of the facet scores.

1.0 INTRODUCTION

TAPAS is a personality assessment measure rooted in the Big Five theory of personality. Rather than focusing on the broad Big Five dimensions, TAPAS assesses their more focused underlying facets that are expected to predict performance in military specialties. To date, there are many versions of TAPAS and more than two million US military applicants have taken TAPAS.

TAPAS incorporates features that address problems with traditional Likert scale personality measures, including faking, limitations of classical test theory (CTT), and test compromise, as well as ipsative scoring for other forced-choice personality measures, which yield scores with an intrapersonal interpretation rather than normative meaning. Specifically, TAPAS uses a two-alternative forced-choice (2AFC) response format, computer adaptive test (CAT) administration, and IRT scoring.

As with any new assessment, careful documentation is needed. A blue-ribbon panel (Roberts, Arthur, Reckase, Sackett, & Zenisky, 2019) detailed required documentation and areas for further research. This report provides some of the documentation recommended by the committee. Specifically, this report provides information about:

- the history of TAPAS
- personality facets assessed by each TAPAS version
- composites computed from the facet scores (e.g., Army's Can-Do)
- samples used to calibrate the psychometric properties of TAPAS statements
- norming and equating procedures
- score interpretation
- underpinnings of the TAPAS adaptive algorithm
- marginal reliability and conditional standard errors (SEs) of scores
- test-retest reliability

A second report, Drasgow et al. (in press), has also been written in response to the recommendations of the blue-ribbon panel.

2.0 BRIEF HISTORY OF TAPAS BEFORE MEPS IMPLEMENTATION

Personality constructs predict performance across diverse civilian and military occupations (e.g., Barrick & Mount, 1991; Campbell & Knapp, 2001; White & Young, 1998) and provide incremental validity beyond general cognitive ability (Schmidt & Hunter, 1998). When studying technical or task performance, personality was thought to only marginally correlate with job performance. However, researchers have found stronger personality-performance relationships for other aspects of job performance, such as citizenship and counterproductive work behaviors (Campbell, 1990, Hough et al., 1990).

In the 1980s, the Army developed the Assessment of Background and Life Experiences (ABLE) to measure personality dimensions. ABLE personality dimensions correlated ($r = .15-.35$) with job

criteria such as effort, leadership, personality discipline, physical fitness, military bearing, first-term attrition, and citizenship performance (Hough et al., 1990). For Air Force mechanics, ABLE personality dimensions had higher correlations with supervisory ratings of citizenship performance than cognitive ability tests (Motowidlo & Van Scotter, 1994).

However, the Likert scale format for ABLE was easily faked. When respondents purposely faked responses in research studies, the predictive validity of ABLE for job-performance was near zero (White et al., 2001; Young, White, & Oppler, 1992). To address potential faking in operational settings, White and Young (1998) built the Army's Assessment of Individual Motivation (AIM).

AIM utilizes a forced-choice response format to mitigate faking. Specifically, each AIM item includes four statements, each assessing a different dimension; two statements are positive on their respective trait continuum and social desirability, and two are negative on their respective trait continuum and social desirability. Respondents then select one statement as "Most like me" and one statement as "Least like me." See Knapp, Heggstad, and Young (2004; page 2-3) for an example item.

If an AIM respondent selects a positive statement as "Most like me" or a negative statement as "Least like me," the respective trait is given +1 point. If a respondent selects a negative statement as "Most like me" or a positive statement as "Least like me," the respective trait is given -1 point. AIM and similar forced-choice personality inventories do not produce normative scores, allowing for direct comparison across respondents, and instead produce intra-personal scoring. Interpretation of scores only allows for comparison between facets for each person (e.g., Person A is more conscientious than agreeable), but not for comparison of respondents on the same facet (i.e., it should not be asserted that Person A is more conscientious than Person B).

2.1 Reasons for TAPAS Development

Drasgow Consulting Group (DCG) originally developed the TAPAS under the Army's SBIR grant program (Drasgow, Stark, & Chernyshenko, 2006; Stark, Drasgow, & Chernyshenko, 2008). The development work began in 2004 and continues to this day.

TAPAS was created with the intention of providing the U.S. military with a modern, computer administered personality assessment for use in selection and classification of military recruits, in conjunction with the ASVAB and other specialized accession testing. The work began by identifying 22 facets underlying and composing the well-known Big Five personality framework (Goldberg, 1993); Table 1 presents these facets. This initial taxonomy was developed using the results of several large-scale factor-analytic studies; results of this effort have been published in peer reviewed journals and books (Chernyshenko, Stark, & Drasgow, 2011; Roberts, Chernyshenko, Stark, & Goldberg, 2005; Woo, Chernyshenko, Longley et al., 2014; Woo, Chernyshenko, Stark, & Conz, 2014).

Table 1. Initial 22 TAPAS Facet Taxonomy: Trait Names, Markers, and Descriptions

Big Five Factor	Current (Initial) Facet Name	Key Adjectives	Brief Description
Agreeableness	Consideration	compassionate, warm, cold, insensitive	High scoring individuals are affectionate, compassionate, sensitive, and caring.
	Cooperation	agreeable, cordial, trusting, uncooperative	High scoring individuals are pleasant, trusting, cordial, non-critical, and easy to get along with.
	Selflessness (Generosity)	charitable, helpful, generous, stingy, selfish	High scoring individuals are generous with their time and resources.
Conscientiousness	Achievement	ambitious, industrious, aimless	High scoring individuals are seen as hard working, ambitious, confident, and resourceful.
	Non-Delinquency	rule-following, lawful, delinquent	High scoring individuals tend to comply with rules, customs, norms, and expectations, and they tend not to challenge authority.
	Order	organized, neat, sloppy	High scoring individuals tend to organize tasks and activities and desire to maintain neat and clean surroundings.
	Responsibility	prompt, irresponsible, unreliable	High scoring individuals are dependable, reliable, and make every effort to keep their promises.
	Self-Control	controlled, deliberate, inconsistent	High scoring individuals tend to be cautious, levelheaded, able to delay gratification, and patient.
	Virtue	honest, frank, misleading	High scoring individuals strive to adhere to standards of honesty, morality, and “good Samaritan” behavior.
	Adjustment	relaxed, certain, insecure, nervous	High scoring individuals are well adjusted, worry free, and handle stress well.
Emotional Stability	Even Tempered	calm, composed, moody, hot-headed	High scoring individuals tend to be calm and stable. They don’t often exhibit anger, hostility, or aggression.
	Optimism (Well-Being)	happy, optimistic, depressed, dejected	High scoring individuals have a positive outlook on life and tend to experience joy and a sense of well-being.

Table 1. (continued)

Big Five Factor	Current (Initial) Facet Name	Key Adjectives	Brief Description
Extraversion	Attention Seeking (Excitement Seeking)	loud, entertaining, dull, unexciting, shy	High scoring individuals tend to engage in behaviors that attract social attention. They are loud, loquacious, entertaining, and even boastful.
	Dominance	assertive, direct, submissive, helpless	High scoring individuals are domineering, “take charge” and are often referred to by their peers as “natural leaders.”
	Sociability	sociable, gregarious, talkative	High scoring individuals tend to seek out and initiate social interactions.
Openness to Experience	Aesthetics	aesthetic, artistic, unsophisticated, unrefined	High scoring individuals appreciate various forms of art and music and participate in art-related activities more than most people.
	Curiosity	curious, perceptive, unobservant	High scoring individuals are inquisitive and perceptive; they are interested in learning new information and attend courses and workshops whenever they can.
	Depth	introspective, reflective, shallow	High scoring individuals tend to examine their lives and exhibit behaviors associated with self-improvement.
	Ingenuity	creative, inventive, unimaginative	High scoring individuals are inventive and can think “outside of the box.”
	Intellectual Efficiency	intelligent, analytical, knowledgeable	High scoring individuals believe they process information and make decisions quickly; they see themselves (and they may be perceived by others) as knowledgeable, astute, or intellectual.
	Tolerance	tolerant, broadminded, biased	High scoring individuals scoring are interested in other cultures and opinions that may differ from their own.
Military Specific	Physical Conditioning	active, vigorous, fit, inactive, brisk	High scoring individuals tend to engage in activities to maintain their physical fitness and are more likely participate in vigorous sports or exercise.

Note. As the TAPAS taxonomy was being developed, various names were used for facets. Initially, Optimism was labeled Well-being, Selflessness was labeled Generosity, and Attention Seeking was labeled Excitement Seeking. Since 2010, these three facets have been named Optimism, Selflessness, and Attention Seeking. Physical Conditioning corresponds to the Energy facet of Extraversion customized for the Army to enhance prediction of performance and consequently it is listed as Military Specific.

TAPAS development continued by combining this new taxonomy with advances in psychometric methods and computing technology to create a new generation of personality measures (Dragow et al., 2012). Four features make TAPAS advantageous over past Likert scale and forced-choice personality testing for military selection and classification purposes:

1. ***computer adaptive test (CAT) administration***: increases test efficiency and security;
2. ***force-choice response format with statements matched on social desirability***: mitigates faking good;
3. ***item response theory (IRT) scoring*** allows for normative scoring for forced-choice responses, more score precision, and comparisons across TAPAS versions;
4. ***narrow personality facets***: stronger personality to military performance linkages.

Computer adaptive test (CAT) administration dynamically tailors content to each individual by drawing statements from large pools. Items are created by pairing statements based on current estimated trait levels (based on choices on previous items) to maximize precision with less items and time than non-adaptive testing. Because each person receives a different set of items, this reduces test comprise.

Because statements in each TAPAS pair are matched on social desirability, the likelihood that respondents can identify the correct response and artificially inflate their scores is reduced. Hence, *TAPAS was expected to demonstrate validity even in high-stakes settings where applicants may be motivated to respond dishonestly*. And, in fact, a recent meta-analysis by Cao and Dragow (2019) found substantial resistance to faking for forced-choice formats (see also Trent, Barron, Rose, & Carretta, 2020). Moreover, research has shown that forced-choice scales maintain validity when they are transitioned from low stakes research settings to operational use in high stakes selection settings (e.g., Bartram, 2007; Hirsh & Peterson, 2008; Lee et al., 2018, O'Neill et al., 2017, Zhang et al., 2020).

Traditional scoring of forced-choice formats produces ipsative scores, which do not allow for between-person comparisons. TAPAS IRT scoring (see section 8) has overcome this limitation and is capable of recovering normative scores (Stark, Chernyshenko, Dragow, & White, 2012). Moreover, to the extent that IRT assumptions are satisfied, trait estimates for a given facet are directly comparable across versions (see section 3).

The current version of TAPAS at MEPS is capable of measuring up to 32¹ non-redundant, narrow personality facets, so assessments can be easily customized to meet the assessment needs of various military services with diverse military occupations. Typically, TAPAS test versions assess between 12 and 17 personality facets, and the resulting scores are then combined into various composites designed to predict military job performance or other important outcomes (e.g., attrition, adjustment to military life). Composite scores can then be used to inform personnel selection or classification decisions by supplementing cognitive ability scores (see section 5).

¹ 29 facets have DoD-exclusive item pools. 19 of the 29 facets also have non-exclusive research item pools. 3 facets have non-exclusive research item pools only.

In sum, TAPAS was originally envisioned as a next generation personality instrument. Its ambitious agenda included computer adaptive assessment for an enhanced personality framework using a fake resistant format that yields normative scores comparable across forms. This report summarizes nearly two decades of TAPAS research and development.

2.2 TAPAS Initial Validation Effort in U.S. Army: Expanded Enlistment Eligibility Metrics (EEEM) Research Project (2006-2009)

The initial version of TAPAS, TAPAS-95s, is a 12-facet version of TAPAS, with the 95 referring to the number of items, and the “s” meaning “static” (non-adaptive). This version was static to allow paper-and-pencil administration for initial testing.

During the U.S. Army’s EEEM Project (Knapp & Heffner, 2010) between 2006 and 2009, non-cognitive measures were studied for utility of use with the enlisted cognitive ability test (ASVAB). TAPAS-95s, AIM (White & Young, 1998) and the Rational Biographical Inventory (RBI; Kilcullen, Putka, McCloy, & Van Iddekinge, 2005) composed the personality measures in the project. Other non-cognitive measures included a situational judgment test and two person-environment (P-E) fit measures.

Soldiers from six Military Occupational Specialties (MOSs) were followed through basic training and several criterion measures were collected, including scores on job-specific knowledge tests, self-reported scores on the APFT, and ratings of job satisfaction and career intentions from the Army Life Questionnaire (ALQ). Soldiers were also evaluated by their peers and supervisors on several performance rating scales.

The three personality measures (TAPAS-95s, AIM and RBI) showed incremental validity over the AFQT, which is a composite of the ASVAB’s Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Math Knowledge subtests. Moreover, the personality measures had smaller subgroup differences than AFQT. Because TAPAS appears resistant to faking, yields normative scores, and is scalable via computerized testing (i.e., it was designed to enable testing of large numbers of individuals), it was selected for further research and development.

Knapp and Heffner (2010) provide detailed results for the EEEM project. Here we briefly describe the TAPAS-95s, and highlight some of the construct and criterion validity findings and subgroup difference results that ultimately led to TAPAS being tried out at MEPS.

2.2.1 TAPAS-95s Description

Statements for the TAPAS research item pools for the 22 facet taxonomy had been developed and pretested using large groups of Army recruits in 2006 and 2007. Of the 22 narrow facets in the TAPAS taxonomy, 12 facets were selected for TAPAS-95s on rational and empirical grounds (i.e., meta-analyses) as being potentially useful for predicting basic training outcomes. Next, 179 statements from the 12 targeted facets were selected from the TAPAS research statement pools and paired to form 71 multidimensional items and 24 unidimensional items. Multidimensional items had statements similar in desirability and extremity, but represented different facets; unidimensional items had statements from the same facet, similar in desirability, but different in extremity. Unidimensional items were included to facilitate scoring and to mitigate potentially

negative test taker reactions to the forced-choice item format. Eleven statements were used twice, so they appeared in two items.

Respondents were instructed to choose the statement in each pair that was “more like me” and that they must make a choice even if they found it difficult to do so. The item responses from 95 items were coded dichotomously and scored using the multidimensional IRT method described by Stark (2002) and Stark, Chernyshenko, and Drasgow (2005).

An example multidimensional item with one statement representing the Order facet and another representing the Curiosity facet is shown below.

- ☐ I hate when people are sloppy.
- ☐ I prefer informative documentaries to other TV programs.

An example unidimensional item for the Dominance facet is:

- ☐ I have hesitated in making others’ decisions for them.
- ☐ I have regularly voiced strong opinions.

For the unidimensional item, it is obvious that the second statement reflects a higher level of Dominance.

Table 2 presents summary statistics for the 12 personality TAPAS 95s facets assessed during the EEEM project. The table shows facet names, brief descriptions of a typical high scorer, facet score means and standard deviations (SDs), and the number of unique statements used to construct TAPAS-95s items.

Table 2. TAPAS-95s Facet Descriptions and Basic Statistics

Big Five Factor	TAPAS-95s Facet	Description	Mean (SD)	# of Statements
Agreeableness	Cooperation	High scoring individuals are trusting, cordial, cooperative, non-critical, and easy to live with.	-0.30 (0.86)	17
	Achievement	High scoring individuals are described as hard working, ambitious, confident, and resourceful.	0.17 (0.64)	16
Conscientiousness	Non-Delinquency	People with high scores on this facet tend to comply with current rules, customs, norms, and expectations; they dislike change and do not challenge authority.	0.09 (0.65)	17
	Order	High scoring individuals tend to organize tasks and activities and desire to maintain neat and clean surroundings.	-0.04 (0.64)	13
Emotional Stability	Even Tempered	Those scoring high tend to be calm, even tempered, and stable.	-0.46 (0.77)	13
	Optimism (Well-Being)	High scoring individuals have a positive outlook on life and tend to experience joy and a sense of well-being.	-0.07 (0.60)	15
Extraversion	Attention Seeking (Excitement Seeking)	Individuals scoring high on this facet are constantly in search of social stimulation; they are loud, loquacious, and entertaining.	-0.14 (0.79)	14
	Dominance	High scoring individuals are domineering, take charge and are often called by their peers as "natural leaders".	-0.15 (0.61)	17
Openness to Experience	Curiosity	High scoring individuals are inquisitive and perceptive; they are interested in learning new information and attend courses and workshops whenever they can.	-0.08 (0.79)	13
	Intellectual Efficiency	Individuals with high scores on this factor are able to process information quickly and would be described by others as knowledgeable, astute, and intellectual.	-0.19 (0.64)	14
	Tolerance	Individuals scoring high on Tolerance like to attend cultural events or meet and befriend people with different views. They also tend to better adapt to novel situations.	-0.42 (0.67)	13
Military Specific	Physical Conditioning	High scoring individuals routinely participate in vigorous sports or exercise and enjoy hard physical work.	0.12 (0.71)	17

Note. N = 4,763. Physical Conditioning corresponds to the Energy facet of Extraversion customized for the Army to enhance prediction of performance. Attention Seeking was previously called Excitement Seeking. Optimism was called Well-Being on TAPAS-95s and later renamed.

2.2.2 TAPAS 95s Construct Validity Results

Table 3 shows correlations between TAPAS-95s facets and relevant dimensions assessed by two other personality inventories included in the EEEM study: the AIM (White & Young, 1998) and the RBI (Kilcullen, Putka, McCloy, & Van Iddekinge, 2005). Also shown are correlations between TAPAS-95s facets and the AFQT.

As can be seen in Table 3, TAPAS-95s facets showed good construct validity. Intellectual Efficiency and Curiosity, for example, showed correlations of .38 and .24, respectively, with the AFQT. This was expected, given that both facets tap the intellectance aspects of Openness to Experience, which is known to correlate with cognitive ability (Woo, Chernyshenko, Stark, & Conz, 2014). The Intellectual Efficiency and Curiosity facets also correlated with the RBI Cognitive Flexibility scale (.33 and .41, respectively), which was also designed as a measure of Openness. The TAPAS Achievement facet correlated most strongly with AIM Work Orientation (.36), indicating that it measures similar behaviors. TAPAS Non-Delinquency correlated with AIM Dependability (.46) and Hostility to Authority (-.44); all these scales were intended to measure rule following and compliance with societal norms. As expected, TAPAS Dominance correlated .50 with the AIM and RBI Leadership scales, while showing much lower correlations with all other scales. Similarly, TAPAS Physical Conditioning correlated highly with AIM Physical Conditioning (.60) and RBI Fitness Motivation (.62) and much lower with everything else.

Table 3. Correlations between TAPAS-95s Facets and Selected Dimensions from the AIM, RBI, and ASVAB

		TAPAS-95s Facet											
	Facet	Ach.	Cur.	Non-Del.	Dom.	Eve. Tem.	Att. See.	Int. Eff.	Ord.	Phy. Con.	Tol.	Coo.	Opt.
AIM	Adjustment	.13	.20	.16	.05	.32	-.17	.13	.00	.09	.12	-.03	.39
	Agreeable	.09	.16	.26	-.04	.40	-.25	.05	.00	.05	.07	.07	.19
	Dependable	.16	.16	.46	.10	.15	-.31	.07	.11	.00	.06	-.02	.07
	Leadership	.19	.22	.03	.50	.02	.05	.23	.05	.10	.13	-.24	.06
	Physical Conditioning	.22	.10	.00	.04	.06	-.06	.02	.05	.60	.06	-.12	.05
	Work Orientation	.36	.23	.05	.22	.12	-.08	.17	.09	.30	.12	-.23	.08
	Lie Scale	.00	-.06	-.02	-.04	.01	-.02	-.03	-.02	.02	-.03	-.03	.02
RBI	Leadership	.15	.22	.03	.42	.04	.08	.23	.02	.13	.17	-.19	.06
	Cognitive Flexibility	.13	.41	.09	.17	.17	-.09	.33	-.03	.03	.25	-.09	.08
	Achievement	.23	.19	.19	.22	.01	-.06	.14	.11	.13	.13	-.13	-.02
	Fitness Motivation	.18	.06	-.09	.08	.04	.02	.06	-.02	.62	.02	-.18	.08
	Stress Tolerance	.16	.17	.06	.08	.26	-.12	.20	-.01	.14	.09	-.07	.31
	Hostility to Authority	-.14	-.18	-.44	-.06	-.19	.34	-.09	-.08	.05	-.08	-.06	-.10
	AFQT	-.06	.24	.06	.06	.14	-.07	.38	-.04	.00	.02	-.04	.18

Note. N = 2,422 – 3,362. Ach. = Achievement; Cur. = Curiosity; Non-Del. = Non-Delinquency; Dom. = Dominance; Eve. Tem. = Even-Tempered; Att. See. = Attention Seeking (previously called Excitement Seeking); Int. Eff. = Intellectual Efficiency; Ord. = Order; Ph. Con. = Physical Conditioning; Tol. = Tolerance; Coo. = Cooperation; Opt. = Optimism (called Well-Being on TAPAS-95s and later renamed). Scales assessing facets of the same Big Five dimensions are bolded. Correlations greater than .04 in absolute value are significant, $p < .05$.

2.2.3 TAPAS-95s Criterion Validity Results

Here, for criterion validity, we only report 6-month attrition, disciplinary incidents, Army Physical Fitness score, and self-reported Adjustment to Army Life scale scores from the ALQ for the combined TAPAS-95s sample, regardless of a soldier's MOS or education/AFQT category. Other results that are MOS or education/AFQT specific are in the technical report (Knapp & Heffner, 2010). Table 4 shows correlations between 12 TAPAS facets and the four outcomes. In addition, we show criterion validities for AFQT, multiple correlations (R s) when AFQT and TAPAS facets were combined (AFQT+TAPAS), and the incremental validity for TAPAS over AFQT. As can be seen in Table 4, for each criterion, there were 3-4 TAPAS facets with validities higher than those observed for the AFQT. And, because AFQT scores had only small to moderate correlations with TAPAS facets (see Table 3 above), considerable incremental validity was observed for the four criteria. For example, adding TAPAS facets to the AFQT increased the overall R by .198 for the 6-month attrition variable. This increase could be considered substantial given the multiplicity of reasons for attrition.

Table 4. Incremental Validity Results for TAPAS-95s Facets and Four Training Criteria

Big Five Factor	Predictor	6-month attrition	Disciplinary Incidents	Army Physical Fitness Test (APFT) Score	ALQ: Adjustment to Army Life
Agreeableness	Cooperation	-.015	-.031	-.111**	-.115**
	Achievement	-.036	-.021	.094*	.219**
Conscientiousness	Non-Delinquency	.008	-.090*	-.083*	.039
	Order	-.007	-.078**	.052	.031
Emotional Stability	Even Tempered	-.103**	-.104**	-.039	.076*
	Optimism	-.095**	.029	-.022	.123**
Extraversion	Attention Seeking	.037	.152**	.018	-.015
	Dominance	.031	.044	.017	.077*
Openness to Experiences	Curiosity	-.044	-.061	.022	.082**
	Intellectual Efficiency	-.027	.015	.022	.068
	Tolerance	-.023	-.037	-.004	.132**
Military Specific	Physical Conditioning	-.101**	.018	.293**	.199**
	AFQT	-.034	-.075*	.011	.111**
	Multiple <i>R</i> (AFQT + TAPAS)	.243**	.223**	.312**	.349**
	ΔR	.198**	.148**	.301**	.238**

Note: 6-month attrition, *N* = 1,696. Disciplinary incidents/APFT/Adjustment to Army Life, *N* = 719. ΔR = Increment in multiple correlation. Nagelkerke's *R* was used for the dichotomous 6-month attrition criterion variable. ** Indicates significant at the .01 level; * Indicates significant at the .05 level. Attention Seeking was previously called Excitement Seeking. Optimism was named Well-Being on TAPAS-95, but later renamed.

2.2.4 TAPAS-95s Adverse Impact Results

The final set of results for the initial field testing of TAPAS was concerned with race/ethnic and gender subgroup differences. Because this assessment system was intended mainly for use in personnel selection and classification contexts, the presence of marked differences in scale means across these groups (a.k.a. adverse impact) could limit test use. Table 5 shows TAPAS facets and AFQT score comparisons for females vs. males (F-M, Column 2), Blacks vs. Whites (B-W, Column 3), and Hispanics vs. White non-Hispanics (H-WNH, Column 4). In each comparison, negative values indicate lower means for protected groups. To facilitate interpretation, standardized group differences (Cohen's *d*) are reported (computed as the difference between respective facet score means divided by the majority group's SD).

As can be seen from Table 5, TAPAS scales showed predictable patterns of gender differences. Males had somewhat higher Physical Conditioning, Optimism, and Even-Tempered scores, while females had somewhat higher Non-Delinquency, Dominance, Order, and Tolerance scores. The magnitudes of the differences are small, with none exceeding .30 in either direction. For the race/ethnic differences, adverse impact results are even more encouraging. While the AFQT showed standardized group differences (*ds*) of -.63 and -.51, none of the TAPAS-95s scales

exhibited differences larger than .25 in any direction. In fact, the largest score differences in B-W comparisons were in favor of Blacks, who were more dominant, tolerant and orderly. Hispanics also were more tolerant than Whites, but their scores on Intellectual Efficiency and Non-Delinquency were somewhat lower.

Table 5. Subgroup Comparisons of AFQT and TAPAS-95s Scores

Predictor		Gender Differences	Race/Ethnic Differences	
		F-M <i>d</i>	B-W <i>d</i>	H-WNH <i>d</i>
	AFQT	-0.30	-0.63	-0.51
Agreeableness	Cooperation	0.09	-0.05	-0.01
	Achievement	0.12	-0.06	-0.07
Conscientiousness	Non-Delinquency	0.30	-0.01	-0.16
	Order	0.25	0.21	0.02
Emotional Stability	Even Tempered	-0.16	-0.01	-0.09
	Optimism (Well-Being)	-0.15	-0.04	0.07
Extraversion	Attention (Excitement) Seeking	0.00	-0.04	0.04
	Dominance	0.29	0.21	-0.02
Openness to Experience	Curiosity	0.09	0.08	0.01
	Intellectual Efficiency	-0.15	0.04	-0.17
	Tolerance	0.25	0.25	0.13
Military Specific	Physical Conditioning	-0.26	0.03	0.03

Note. N = 2,422. F-M = female vs. male; B-W = Black vs. White; H-WNH = Hispanic vs. White-non-Hispanic. Attention Seeking was previously called Excitement Seeking. Optimism was named Well-Being on TAPAS-95, but later renamed.

2.3 Summary and Conclusions from the Initial TAPAS Validation Efforts and Approval for Use at Military Entrance Processing Stations (MEPS)

The Army Class and EEEM research showed TAPAS to be a viable assessment tool with the potential to enhance new Soldier selection and classification decisions. Trait scores exhibited construct validity evidence with respect to other measures and criterion-related validity estimates were fairly high for outcomes not predicted well by AFQT. Criterion results clearly showed that including a personality inventory in the U.S. Army selection and classification test battery in addition to ASVAB would better identify applicants who are more motivated to finish their training and are capable of meeting the physical and emotional demands of military life. Demographic subgroup comparisons for TAPAS facets revealed little if any impact against members of protected groups. In fact, minorities and women earned higher TAPAS scores than members of comparison groups on several scales, meaning that if the TAPAS scores were used in conjunction with AFQT for selection decisions, the overall impact against protected groups might be reduced.

Based on the positive TAPAS-95s results from the EEEM research and taking into consideration the unique advantages of TAPAS (e.g., flexibility and resistance to faking), the U.S. Army

approved the initial operational testing and evaluation of the TAPAS for use with Army applicants at MEPS on April 03, 2009. Appendix A provides details of the TAPAS MEPS implementation memo. MEPS TAPAS testing was intended to be computerized and administered on the same platform as ASVAB. The U.S. Air Force and U.S. Navy also authorized TAPAS testing at MEPS for some of their applicants, but for research purposes only.

3.0 TAPAS MEPS TESTING

3.1 Chronological Order of TAPAS Version Releases

Since its initial deployment at MEPS in 2009, the TAPAS has undergone several revisions to meet the needs of test users. New facet scales were developed to assess promising constructs for predicting important military outcomes, and statement pools have been revised, updated, and expanded; details are in Section 4. Newer TAPAS versions have been progressively released with different mixes of facets, test specifications, and composites.

3.1.1 Initial Deployment: Army Phase 1

For the initial deployment of the TAPAS at the MEPS (Phase 1), a 13-facet, 104-item adaptive version (referred to as Version 2 or V2) was introduced at selected MEPS in May 2009. The V2 version removed Curiosity from the facets included in TAPAS-95s and added Sociability and Selflessness (Generosity). After a brief two-month testing period, the Army Research Institute (ARI) decided to add two more facets, Adjustment and Self-Control, and to increase test length to 120 items. Based on this decision, two new TAPAS versions were quickly released. In July 2009, a static 15-dimension (15-D), 120-item version replaced V2 at selected MEPS. Then, in September 2009, after TAPAS testing was expanded to all MEPS, an adaptive 15-D, 120-item version was deployed. Both versions, static Version 3 (V3) and adaptive Version 4 (V4), were administered until August 2011, with the majority of test takers completing the adaptive version. Table 6 summarizes facet changes. All items in Phase 1 were developed prior to TAPAS-95s (see Section 4.1).

Army and Air Force applicants started taking the TAPAS at MEPS from the beginning in May 2009. Navy applicants started taking the TAPAS in April 2011. Altogether, between May 2009 and August 2011, over 250,000 MEPS applicants had taken TAPAS. Of these 70 percent (%) were Army, 25% were Air Force, and the remaining 5% were Navy applicants.

Table 6. Army Phase 1 Facet Summary

22 TAPAS Facets in Original Taxonomy	TAPAS-95s	V2	V3/V4
Achievement, Attention Seeking, Cooperation, Dominance, Even Tempered, Intellectual Efficiency, Non-Delinquency, Optimism, Order, Physical Conditioning, and Tolerance	X	X	X
Curiosity	X		
Sociability and Selflessness		X	X
Adjustment and Self-Control			X
Ingenuity, Responsibility, Virtue, Consideration, Aesthetics, and Depth			

Note. Attention Seeking was previously called Excitement Seeking before TAPAS-95. Optimism was named Well-Being on TAPAS-95, but later renamed. Selflessness was previously called Generosity.

3.1.2 New Forms with MEPS-Only Items: Army Phase 2

In 2011, the most substantial revision of the TAPAS was completed. The purpose of this revision was to develop completely new item pools to be used exclusively at the MEPS (i.e., close-hold item pools not used for research or administration outside of the MEPS) and to expand research on the validity of the TAPAS scales by administering a small set of experimental scales along with a common core that had demonstrated validity in previous research. To this end, close-hold item pools were developed for the facets used in MEPS testing (see Table 7). These item pools have been used exclusively at MEPS where it is imperative that tests are secure.

Three new 15-D, 120-item TAPAS versions were created based on the new item pools in Phase 2. These versions were labeled Version 5 (V5), Version 7 (V7), and Version 8 (V8)². V5 contained the same TAPAS scales as V4 and was designed to collect additional data on existing TAPAS scales. V7 retained the 9 core scales from V4/V5 (see Table 7) that were used to calculate the operational composites (see Section 6), as well as Order, Cooperation, and Selflessness (Generosity), but also added three new facets: Situational Awareness, Commitment to Serve, and Adventure Seeking. V8 also retained the 9 core scales from V4/V5, Self-Control, Sociability, and Tolerance, which were included in V4/V5, Responsibility (from the original 22-facet taxonomy), and two other new facets: Courage and Team Orientation.

During Phase 2, there were a total of 23 facets that had DoD exclusive items, 18 from the original 22-facet taxonomy that had research item pools (V4/V5 facets plus Responsibility, Curiosity and Consideration), and 5 new facets without research item pools (Situational Awareness, Commitment to Serve, Adventure Seeking, Courage and Team Orientation). However, Curiosity and Consideration were not deployed during this phase. Non-Delinquency, Physical Conditioning, and Cooperation change themes slightly to not overlap with the 5 new facets. Also, during phase 2, there were a total of 4 facets (Ingenuity, Virtue, Aesthetics, and

² Although TAPAS V6 was developed, it was not implemented. As a result, it is not discussed in this report.

Depth) that had research item pools, but had no DoD exclusive items. Note that DoD exclusive Virtue items were later implemented in 2015. See Table 7 for a Phase 2 facet summary.

Table 7. Army Phase 2 Facet Summary

23 Facets with DoD Exclusive Items in Phase 2	V5	V7	V8	In Original 22-facet Taxonomy
Achievement, Adjustment, Attention Seeking, Dominance, Even Tempered, Intellectual Efficiency, Non-Delinquency ¹ , Optimism, and Physical Conditioning ¹	X	X	X	X
Order, Cooperation ¹ , and Selflessness	X	X		X
Self-Control, Sociability, and Tolerance	X		X	X
Responsibility			X	X
Situational Awareness, Commitment to Serve, and Adventure Seeking		X		
Courage and Team Orientation			X	
Curiosity and Consideration				X
<i>Note.</i> V5 facets are the same facets as V4 in Phase 1. Four facets from the 22-facet original taxonomy, Ingenuity, Virtue, Aesthetics, and Depth had non-exclusive research item pools only in 2011. Virtue was later added to the DoD exclusive research pool. ¹ Themes changed slightly in 2011 to not overlap with new facets.				

A new TAPAS CAT executable program to administer V5, V7, and V8 was developed in VB.NET and installed at the MEPS. This program was designed to randomly choose one of three TAPAS versions to administer to an examinee upon the initiation of a new testing session by a MEPS proctor. The versions varied in composition to ensure that each of the 21 facets would be administered to at least 40% of examinees overall. Army applicants remained the largest group taking the TAPAS (71%), followed by Navy (14.7%), and Air Force (14.3%). At the end of June 2013, the Navy discontinued collecting research data on TAPAS at MEPS.

3.1.3 Reduced Dimensionality: Army Phase 3

In August 2013, a decision was made by the ARI to reduce the number of administered facets from 15 to 13, while maintaining the test length at 120 items (this change was intended to improve reliability of TAPAS facet scores). Three new versions were implemented in September 2013 for Phase 3, Version 9 (V9), Version 10 (V10), and Version 11 (V11). There were 10 core facets across the three versions (see Table 8). Six core facets from the Phase 2 overlap with Phase 3 core facets. Out of the 23 facets with DoD Exclusive item pools, only Adjustment, Adventure Seeking, Self-Control, Curiosity and Consideration were not administered in one of the three versions (V9, V10, V11). See Table 8 for a summary of Phase 3 changes for the Army.

Table 8. Army Phase 3 Facet Summary

23 Facets with DoD Exclusive Item Pools in Phases 2 and 3	Phase 2 Core	Phase 3 Core	V9	V10	V11
Achievement, Dominance, Even Tempered, Intellectual Efficiency, Optimism, and Physical Conditioning ¹	X	X	X	X	X
Order, Selflessness, Sociability, and Tolerance		X	X	X	X
Adjustment	X				
Attention Seeking	X				X
Non-Delinquency ¹	X			X	
Commitment to Serve			X		X
Cooperation ¹ and Responsibility			X		
Courage and Situational Awareness				X	
Team Orientation					X
Adventure Seeking, Self-Control, Curiosity, and Consideration					

Note. Ingenuity, Virtue, Aesthetics, and Depth had non-exclusive research item pools only. Virtue was later added to the DoD exclusive research pool. ¹Themes changed slightly in 2011 to not overlap with new facets.

Several unmotivated respondent flags were also implemented (Stark et al., 2017; see Section 8) and a new executable program was installed at MEPS to randomly administer each version to eligible U.S. Army applicants.

Version 5 continued to be administered to Air Force applicants as part of their research effort. However, in August 2013 (just before V9, V10, & V11 were released), the Air Force increased the time limit from 30 minutes to 45 minutes, and kept the same facets from V5. This version is called AF V5.1. In October 2014, the Air Force implemented separate PSMs for Special Operations specialties (then called Battlefield Airman) and related specialties (Rose, Manley, & Weissmuller, 2013). The PSMs composites included ASVAB subtests and TAPAS facets, and were designed to decrease training attrition. During Phase 3, some PSMs were dropped or changed, and others were added for additional specialties. PSMs typically include ASVAB subtests and TAPAS facets, and sometimes include other special tests.

At the end of 2014, the United States Marine Corps (USMC) expressed interest in having a dedicated TAPAS version to be administered to their applicants for research purposes, and, as a result, a 15-D, 120 item USMC version (USMC V1.1) was implemented at MEPS in August 2015. Two facets (Ingenuity and Virtue) used research item pools instead of DoD exclusive item pools.

During Phase 3 (the 2013-2019 time period), nearly 1.2 million Army, Air Force, and USMC applicants completed TAPAS assessments at MEPS.

3.1.4 New Versions: Army Phase 4

In October 2019 (Phase 4), several new TAPAS MEPS versions were implemented. The change was driven by various factors. The Army had wanted to focus on a common set of 13 TAPAS facets which would be used to compute various personnel selection and screening composites, while, at the same time, to continue to administer a handful of new facets for research purposes. To meet these requirements, three new TAPAS versions, Version 12 (V12), Version 13 (V13), and Version 14 (V14), were created for Army applicant MEPS testing. The first part of the assessment for all three versions was identical: a 13-D, 133 item adaptive test that focused on core TAPAS facets. The second part of the assessment, which started from item 134 and finished at item 176, was different for each version and included 4 additional “research” TAPAS facets (Commitment to Serve is a research facet across all three versions). Accordingly, test results were saved twice for each examinee: 1) after item 133 (Part 1) to be used for operational purposes, and 2) after item 176 (Part 2) to be used for research purposes. During this Phase, 6 new facets (Army Self-Efficacy, Humility, Persistence, Self-Efficacy, Machiavellianism, and Virtue) were added to the DoD exclusive pool taxonomy, but one facet (Machiavellianism) was not deployed in Versions 12, 13, and 14. Table 9 summaries the core facets across phases and for Versions 12, 13, and 14.

Table 9. Army Phase 4 Facet Summary

29 Facets with DoD Exclusive Item Pools	Phase 2 Core	Phase 3 Core	Phase 4 Core	V12	V13	V14
Achievement, Dominance, Even Tempered, Intellectual Efficiency, Optimism, and Physical Conditioning ¹	X	X	X	X	X	X
Order, Sociability, and Tolerance		X	X	X	X	X
Selflessness		X				
Team Orientation			X	X	X	X
Adjustment, Attention Seeking, Non-Delinquency ¹	X		X	X	X	X
Commitment to Serve ²				X	X	X
Army Self-Efficacy ³					X	X
Humility ³				X		X
Persistence ³				X	X	
Self-Efficacy ³				X		
Virtue ³					X	X
Machiavellianism ^{3,4} , Adventure Seeking, Cooperation ¹ , Curiosity, Consideration, Courage, Responsibility, Self-Control, and Situational Awareness						

Note. Three facets from the research pool (Ingenuity, Aesthetics, and Depth) do not have DoD exclusive items.

¹Themes changed slightly in 2011 to not overlap with new facets. ²On all three Phase 4 versions, but not a core facet. ³New facets in Phase 4, all other facet DoD exclusive item pools added in 2011. ⁴DoD exclusive items created in phase 4, but currently not on any version.

The Air Force implemented a 124-item, 15D adaptive test (AF V5.2) that replaced the previously administered Order and Intellectual Efficiency facets with Responsibility and Situational Awareness facets. Intellectual Efficiency was removed because of its somewhat strong correlation with the ASVAB, and researchers wanted to test Situational Awareness. Order was removed because at that time, out of the current PSMs, only one had Order, and Air Force researchers wanted to try out a different Conscientiousness facet (which tend to have strong correlations with training performance). During this phase, new PSMs were continually added or edited based on requests from Career Field Managers. As of October 2022, there were 12 operational PSMs for 13 AFSCs; (1A8X1 and 1N3X1 share a PSM). The focus of the PSMs is to reduce adverse impact for all gender and racial/ethnic groups, in addition to reducing training attrition. Most AFSCs with PSMs also have either a Mechanical, Administrative, General or Electronic (MAGE) ASVAB composite (must meet MAGE and PSM cutoffs), only Explosive Ordnance Disposable and the linguists (1A8X1 and 1N3X1) do not (but their PSMs have ASVAB subtests). However, in the future, it is likely that recruits can qualify by meeting a PSM cutoff or MAGE cutoff. Currently all PSMs have ASVAB subtests, and/or TAPAS facets, but no special tests.

The USMC is not currently using TAPAS for enlistment eligibility decisions. With an automatic waiver in place, the scores of an applicant's TAPAS composites are currently not factored into their accession package. The USMC has a TAPAS composite that was created to operationalize TAPAS based on success in boot camp using a logistic regression model. Some of the TAPAS facets that were once significant predictors of success in bootcamp have shifted from that initial analysis. The USMC is considering other uses for TAPAS.

Table 10 below shows the chronological order of the deployment of each TAPAS version at the MEPS. The table shows the dates of each version's administration, the number of facets assessed, the number of items administered, and, when available, the number of examinees with valid data who completed each version.

Table 10. TAPAS MEPS Versions by Service and Administration Date

Army Phase	Notes	Version	Services	Dates Administered	# of Facets	# of Items	Sample Size
1	V3/V4 – same facets (adds adjustment & self-control to V2 facets).	V2	Army & AF	May 2009 – July 2009 ¹	13	104	2,290
		V3(s)	Army, AF, & Navy	July 2009 – Aug 2011 ²	15	120	26,074
		V4	Army, AF, & Navy	Sept 2009 – Aug 2011 ²	15	120	228,154 ³
2	DoD exclusive pool created for 23 facets ⁴ . V4/V5 – facets same but diff. items. Navy stops.	V5	Army	Aug 2011 – Sept 2013 (20%)	15	120	73,555
		V7	Army	Aug 2011 – Sept 2013 (40%)	15	120	147,578
		V8	Army	Aug 2011 – Sept 2013 (40%)	15	120	147,796
2/3	V5 w/ 45 min time limit starts Aug 2013. USMC starts. IER flags implemented	AF V5/V5.1 ⁵	AF	Aug 2011 – Oct 2019	15	120	175,178
		USMC 1.1	USMC	Aug 2015 – Oct 2019	15	120	66,570
3	IER flags implemented. Reduced to 13 facets for increased reliability.	V9	Army	Sept 2013 – Oct 2019 (33.3%)	13	120	293,839
		V10	Army	Sept 2013 – Oct 2019 (33.3%)	13	120	293,513
		V11	Army	Sept 2013 – Oct 2019 (33.3%)	13	120	295,011
4	AF replaces order & intellectual efficiency. USMC new version. IER items added.	AF V5.2	AF	Oct 2019 – present	15	124	61,447
		USMC V2.1	USMC	Oct 2019 – present	12	132	74,174
4	New DoD exclusive item pools made for 6 facets ⁶ . Core 13 facets for first part of test. New IER items added.	V12	Army	Oct 2019 – present (33.3%)	13/17 ⁷	133/176 ⁷	60,704
		V13	Army	Oct 2019 – present (33.3%)	13/17 ⁷	133/176 ⁷	60,519
		V14	Army	Oct 2019 – present (33.3%)	13/17 ⁷	133/176 ⁷	60,607

Note. ¹In Sept 2009 testing was expanded to all MEPS. ²For Navy: April 2011- Aug 2011. ³Invalid V4 data (N = 23,803) are excluded; this was a result of a software problem. ⁴For the DoD exclusive item pool 5 facets were added, 4 were removed, and 3 changed content slightly from the 22-facet taxonomy for the original research statement pool. ⁵AF V5.1 is the same assessment as AF V5 but has a 45-minute time limit. ⁶5 new facets and virtue (was in research pool, but not DoD exclusive pool). ⁷133 items for 13 core facets (presented first). Grand total of 32 facets (22 in research pool, 29 in DoD exclusive pool). Except V3(s) and 95(s), all TAPAS versions are adaptive. S = static. DoD = Department of Defense. Diff. = different. w/ = with. IER = Insufficient Effort Responding. N/A = not available.

3.2 TAPAS MEPS Version Creation

The first step in creating a TAPAS version is to decide which personality facets should be measured. The TAPAS facet library currently has 32 facets to choose from and new facets are added periodically in response to test user requests (e.g., the Persistence facet was added in 2019 to cover behavioral patterns related to grit and resilience). Most TAPAS versions were designed to measure 15 personality facets, but this can vary between 12 and 17 (see Table 10 above). The item pools for each facet in the database contain 40-50 statements, which have been previously pretested on large samples of military recruits. Each statement has three IRT parameters (alpha, delta, tau) and one social desirability parameter; these are used by TAPAS test administration and scoring algorithms.

Table 11 shows the facets assessed by each MEPS TAPAS version. As can be seen from the table, 27 facets from the TAPAS facet library (including Ingenuity in the research pool, and not including Consideration, Aesthetics, Curiosity, Depth, and Machiavellianism) have been used in one or more MEPS TAPAS versions. Several facets (e.g., Achievement, Dominance, Optimism, Physical Condition) have been used in nearly all TAPAS versions; these facets are often referred to as “core” TAPAS facets. A more detailed description of the facets can be found in Table 1 and Section 4 of this report.

Table 11. TAPAS Facets Assessed at MEPS by Version

	Phase 1: Initial May 09- July11		Phase 2: MEPS-Only Items Aug 11-Aug 13			Phase 3: Reduced Dimensionality Sep 13-Aug 19					Phase 4: New Versions Sep 19 - Present				
TAPAS Facets	13D V2	15D Static V3s & V4	15D V5	15D V7	15D V8	13D V9	13D V10	13D V11	15D AF V5.1	15D USMC V1.1	17D V12	17D V13	17D V14	12D USMC V2.1	15D AF V5.2
Agreeableness															
Consideration ^{1,B}	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Cooperation ^B	X	X	X	X	--	X	--	--	X	--	--	--	--	--	X
Selflessness ^B	X	X	X	X	--	X	X	X	X	X	--	--	--	X	X
Humility ^E											X	--	X	--	--
Conscientiousness															
Achievement ^B		X	X	X	X	X	X	X	X	X	X	X	X	X	X
Non-Delinquency ^{2,B}	X	X	X	X	X	--	X	--	X	--	X	X	X	--	X
Order ^B	X	X	X	X	--	X	X	X	X	--	X	X	X	X	--
Persistence ^E											X	X	--	X	--
Responsibility ^B	--	--	--	--	X	X	--	--	--	X	--	--	--	--	X
Self-control ^B	--	X	X	--	X	--	--	--	X	--	--	--	--	--	X
Virtue ^B	--	--	--	--	--	--	--	--	--	X	--	X	X	--	--
Emotional Stability															
Adjustment ^B	--	X	X	X	X	--	--	--	X	X	X	X	X	--	X
Even Tempered ^B	X	X	X	X	X	X	X	X	X	X	X	X	X	--	X
Optimism ^B	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Extraversion															
Attention-Seeking ^B	X	X	X	X	X	--	--	X	X	--	X	X	X	--	X
Dominance ^B	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Sociability ^B	X	X	X	--	X	X	X	X	X	X	X	X	X	X	X
Openness															
Aesthetics ^R	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Curiosity	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Depth ^R	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Ingenuity ^R	--	--	--	--	--	--	--	--	--	X ³	--	--	--	--	--
Intellectual Eff. ^B	X	X	X	X	X	X	X	X	X	--	X	X	X	--	--
Tolerance ^B	X	X	X	--	X	X	X	X	X	X	X	X	X	X	X

Table 11 (continued)

	Phase 1: Initial May 09- July 11		Phase 2: MEPS-Only Items Aug 11-Aug 13			Phase 3: Reduced Dimensionality Sep 13-Aug 19					Phase 4: New Versions Sep 19 – Present				
TAPAS Facets	13D V2	15D Static V3s & V4	15D V5	15D V7	15D V8	13D V9	13D V10	13D V11	15D AF V5.1	15D USMC V1.1	17D V12	17D V13	17D V14	12D USMC V2.1	15D AF V5.2
Military Specific															
Adventure Seeking ^E			--	X	--	--	--	--	--	--	--	--	--	--	--
Army Self-effic ^E											--	X	X	--	--
Commit to Serve ^E			--	X	--	X	--	X	--	X	X	X	X	X	--
Courage ^E			--	--	X	--	X	--	--	X	--	--	--	X	--
Physical Cond ^B	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Machiavellianism											--	--	--	--	--
Self-efficacy ^E											X	--	--	--	--
Situational Aware ^E			--	X	--	--	X	--	--	--	--	--	--	--	X
Team Orientation ^E			--	--	X	--	--	X	--	--	X	X	X	--	--

Note: Xs indicate that the facet is administered on the version. Dashes indicate that the facet was available, but not administered on that version. Blanks indicate that the facet items were not yet created for the respective pool. In Phases with multiple versions, form administration was random: Phase 2, 20% on V5, 40% on V7, 40% on V8. In the Army for Phases 3 and 4, each version was administered to 1/3 of the examinees. Phase 4 facets highlighted in red are only administered in Part 2 of the test (after item 133). Facets not administered in the MEPS: Consideration, Aesthetics, Curiosity, Depth, Machiavellianism. For each facet, R = research item pool, E = DoD exclusive item pool, B = Both research and DoD exclusive item pool. ¹Not administered at MEPS. ²Item themes changed after Phase 1. ³Research item pool administered at MEPS after Phase 1.

The second step in TAPAS version creation is to decide on the number of items (pairs of statements) to administer and to create a test blueprint that specifies all permissible multidimensional and unidimensional statement combinations. For example, a 7-item, 3-dimensional test may have an underlying test blueprint of five multidimensional combinations (1-2, 2-3, 1-3, 1-2, 1-3) and of two unidimensional combinations (2-2, 3-3). Decisions about which dimensional combinations to include in the test blueprint are driven by both psychometric and substantive reasons. Each test should contain a reasonable number of statements per facet, so measurement precision is maintained. At the same time, repeating the same facet combinations too many times, or pairing some highly correlated facets, or having only multidimensional combinations may negatively affect the test taker experience. For these reasons, TAPAS test version blueprints typically have had close to equal number of statements per facet (e.g., a 120-item, 15-facet test would have 16 statements per facet distributed across various dimension combinations), with one or two unidimensional combination per facet (e.g., a 15-D test blueprint would have 15 or 30 unidimensional combinations). In some earlier TAPAS versions, no multidimensional combinations were allowed for Commitment to Serve and Courage facets (e.g., a test version with 16-statements per facet would have 8 unidimensional combinations specified for Commitment to Serve). Appendix B shows the test blueprint for TAPAS version V9.

The third step is to decide on the test administration mode (static or adaptive) and to specify item construction and administration constraints. All but one TAPAS version at MEPS have been adaptive, meaning that items are constructed and administered in real time as test takers progress through the test. In a nutshell, the TAPAS adaptive test administration algorithm proceeds as follows:

- a. Begin with the facet combination from the test blueprint for the first item (e.g., 1-2).
- b. Construct all permissible items for that combination by pairing all available statements for facet 1 with all available statements for facet 2, subject to statement pairing constraints (e.g., a statement may only be used twice, pairs of statements must have social desirability parameters and extremity parameters that are not too dissimilar).
- c. Calculate item information values at the examinee's current estimated facet scores (these indicate measurement precision) for each permissible item (i.e., pair of statements), and pick one item randomly among the most informative to administer.
- d. After the item has been administered, update current facet scores and proceed to the next facet combination in the test blueprint; repeat steps b, c, and d.
- e. Terminate the test after the last facet combination is administered, update final scores, and save results into a designated database file.

Section 7 describes in detail specifics of the CAT algorithm test construction and administration process.

In the case of a static TAPAS test (V3), the test blueprint was used to construct items beforehand, and these were administered to all examinees in the same order. Scoring was only done at the end of the test and all test statistics were saved into a designated database similar to what is done with adaptive tests.

3.3 TAPAS MEPS Test Administration Procedures

TAPAS test administration procedures at MEPS for all TAPAS versions have remained nearly the same since the inception. Each testing session is initiated by a test administrator who enters the examinee's social security number. Next, each examinee is asked to read information related to the purpose of the assessment. For example, for Version 4, applicants were apprised that the scores might be used to determine eligibility for enlistment or for research. Then an instruction page appears providing detailed information about answering TAPAS items and showing some examples. Examinees are told to consider how they typically think, feel, and act, and to indicate which statement in each pair (i.e., each pairwise preference item) is "more like me." They are informed that some pairs may be difficult to answer and, in such cases, they should consider both options carefully and indicate the one that describes them, perhaps just slightly better than the other. After making their choice by clicking on the appropriate statement, they should affirm their response and continue with the assessment by clicking the "Next Item" button.

Testing proceeds in this manner until all items are completed or the 45-minute time limit elapses. In the event of a test interruption, the administrator can restart the testing session at a later time from the point of interruption as testing progress is saved in a temporary file. Detailed results for each testing session are saved and transferred to a central database upon test completion. These results include item responses, facet and composite scores, as well as various test diagnostics designed to detect unmotivated responding.

4.0 TAPAS FACETS

This section describes narrow trait domains (facets) assessed by the TAPAS. The initial 22-facet TAPAS taxonomy was developed to closely map onto the well-known Big Five personality framework (Digman, 1990; Goldberg, 1990; Norman, 1963). Subsequently, facets were added based on literature reviews suggesting potential for incremental validity.

In addition to the facets, this section describes statement pools that were developed for their assessment. The initial TAPAS MEPS deployment (V2, V3, and V4) was based mainly on a statement pool developed by the DCG to support a broad spectrum of research activities involving forced-choice personality tests in military contexts; this statement pool is referred to in this report as the "research pool." Following the successful MEPS implementation, additional statement pools were developed for facets intended to be administered at MEPS in the long run; these statement pools are referred to in this report as "DoD exclusive pools" and their contents are close-hold.

4.1 Initial 22-facet TAPAS Taxonomy

In the last 30 years, personality researchers have reached a general consensus that the Big Five broad personality factors, Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience, are sufficient to adequately describe normal personality (see Borkenau & Ostendorf, 1990; Costa & McCrae, 1988; Digman, 1990; Goldberg, 1990, 1993; Hogan, 1991; Saucier, 1992). Importantly, numerous studies have been conducted to map existing inventories onto the Big Five structure (e.g., Chernyshenko, Stark, & Chan, 2001;

McCrae, Costa, & Piedmont, 1993), which further facilitated the integration of the vast empirical literature. Consequently, most current personality test manuals include a section detailing how their scales and/or scale composites relate to the Big Five (e.g., Conn & Reike, 1994).

Not surprisingly, the initial TAPAS facet taxonomy focused on identifying a comprehensive set of non-redundant narrow traits, which, if desired, could be combined to form the Big Five. The taxonomy was empirically based and combined the results of Saucier and Ostendorf's (1999) analyses of 312 vectors of responses to 500 personality adjectives with a series of factor analyses of questionnaire data from seven major personality inventories: the revised NEO Personality Inventory (Costa & McCrae, 1992; Costa, McCrae, & Dye, 1991), the Sixteen Personality Factor Questionnaire (Conn & Rieke, 1994), California Personality Inventory (Gough, 1987), the Multidimensional Personality Questionnaire (Tellegen, 1982), the Jackson Personality Inventory – Revised (Jackson, 1994), the Hogan Personality Inventory (Hogan & Hogan, 1992), and the Abridged Big Five-Dimensional Circumplex scales from the International Personality Item Pool (Goldberg, 1997). The main assumption of this empirical approach to taxonomy creation was that all important personality traits had been already encoded in the human lexicon or in existing personality scales, so studying the covariation among them would thus lead to identification of all important facets.

Analyses of lexical and questionnaire-based studies yielded a 22-facet taxonomy distributed across the Big Five (see Drasgow et al. [2012] for detailed discussion of results). Specifically, Conscientiousness and Openness to Experience had 6 facets each, Extraversion had 4 facets, while Emotional Stability and Agreeableness each had 3 facets. Table 1 shows current and initial names for the 22 facets, lists examples of existing adjective markers, and provides brief descriptions of a high scoring individual. All early TAPAS tests administered in the 2006-2011 time period were created by selecting the most relevant facets from this initial facet set.

4.2 2011 Revision of TAPAS Facet Taxonomy

In 2011, the most substantial revision of TAPAS was completed. One of the main goals of that revision was to update the TAPAS facet taxonomy. Also, during this time DoD exclusive items were created. Five new facets were added to the TAPAS taxonomy to enhance the capabilities of the assessment system. The five facets tapped into behavioral domains directly relevant to military life and, thus, had the potential to increase the validity of TAPAS composites for selection and classification decisions.

The first new facet was Adventure Seeking, which focuses primarily on high intensity, high risk outdoor activities. It was expected that this facet would be particularly relevant for predicting performance and retention of Soldiers requiring long periods of outdoor activity, such as Infantry and Special Forces MOS.

The second new facet was Commitment to Serve, which assesses one's level of identification with the military and commitment to a military lifestyle. High scoring individuals respect the military and take pride in being able to serve their country. This facet was included in the RBI (Kilcullen et al., 2005) and has shown promise in predicting Soldier retention.

The third new facet was Courage, which assesses how brave and daring applicants are when faced with adversity. High scoring individuals stand up to challenges and indicate a willingness to operate in dangerous situations. This facet was expected to predict combat performance, retention, and re-enlistment intentions.

Team Orientation was the fourth new TAPAS facet. Behaviors associated with this facet deal with the desire to work in a team environment. This facet was expected to predict peer and supervisory performance ratings of teamwork, especially in jobs requiring extensive group activities.

The final new facet was Situational Awareness. Scores reflect how vigilant and attentive Soldiers are to their external environments. Individuals scoring high on this facet pay attention to their surroundings and rarely get lost or surprised. This facet was deemed particularly relevant for predicting performance in combat and guard-related duties.

Because five new facets were added to the TAPAS taxonomy and some of the original 22 facets were not being considered as candidates for MEPS testing, a decision was made by ARI to not create DoD exclusive items for four facets, Aesthetics, Depth, Ingenuity, and Virtue (created later), from the original taxonomy, leaving 23 facets in the updated facet taxonomy of DoD exclusive item pools. Minor adjustments were made to facet definitions to accommodate the new facets. Teamwork themes previously found in the Cooperation facet were migrated to the Team Orientation facet, while outdoor activities that were part of the Physical Conditioning facet migrated to Adventure Seeking. Also, the Non-Delinquency facet had traditionalism vs. liberalism themes removed and its revision focused exclusively on rule-following. Table 12 below lists the 2011 changes made to the TAPAS facet taxonomy.

Table 12. 2011 Updates to the TAPAS Facet Taxonomy (DoD Exclusive only)

Action	Affected Facets
Added	Adventure Seeking, Commitment to Serve, Team Orientation, Courage, Situational Awareness
Removed	Aesthetics, Depth, Ingenuity, Virtue ¹
Updated	Non-Delinquency, Physical Conditioning, Cooperation

Note. ¹Re-added to DoD exclusive taxonomy later.

4.3 Addition of New Experimental TAPAS Facets to Support Enhanced Suitability Screening

During the 2015-2017 time period, six additional facets were added to the DoD exclusive TAPAS facet item pool taxonomy to support the Enhanced Suitability Screening of positions of significant trust and authority; five were new and one, Virtue, was part of the original TAPAS taxonomy. Details of this work are provided in Nye et al. (2018a; 2018b; 2020a). The new facets were initially developed to support validation studies with Recruiters and Drill Sergeants

(DS(s)), but later five of these facets were added as experimental facets in Phase 4 of TAPAS MEPS testing (Versions 12, 13, and 14).

An initial literature search identified 11 potential facets that could be developed to supplement the existing TAPAS facets. Several of these facets (e.g., Narcissism, Machiavellianism) were related to the dark side of personality (Paulhus & Williams, 2002). Other facets had notable criterion-related validities with performance or counterproductive behaviors. For example, past research found that Persistence has a strong positive relationship with performance ratings (.39; Tsai, Chen, & Liu, 2007) and Humility has a negative relationship with delinquency (ranging from -.34 to -.55; Lee, Ashton, & de Vries, 2005).

Based on the results of a literature review, six facets were added to the existing 23 TAPAS facets. These facets included Machiavellianism (not yet administered at MEPS), Army Self-Efficacy, Self-Efficacy, Persistence, Humility, and Virtue. These facets were selected because they had limited overlap with existing TAPAS scales and had demonstrated validity in past research. The one exception was the Virtue facet which had been a part of the initial TAPAS taxonomy but was dropped from DoD exclusive taxonomy in 2011. Table 13 lists the names of these six experimental facets added to the TAPAS taxonomy along with known adjective markers and descriptions of high scoring individuals.

Table 13. Description of Six New Experimental TAPAS Facets

Facet Name	Key Adjectives	Brief Description
Army Self-Efficacy	N/A	High scoring individuals are confident in their ability to accomplish any task that they encounter in the military.
Humility	Humble, modest, pretentious, boastful	High scoring individuals think of others before themselves and are not preoccupied with being recognized for their accomplishments.
Machiavellianism	Cunning, scheming, unscrupulous, ethical, just	High scoring individuals tend to manipulate or exploit others to get what they want.
Self-Efficacy	Confident, self-assured, self-doubtful	High scoring individuals believe they can effectively deal with most difficult situations.
Persistence	Persistent, determined, relentless, gritty, unrelenting, laid back	High scoring individual persist in the face of obstacles and see projects through until completion.
Virtue	honest, frank, misleading	High scoring individuals strive to adhere to standards of honesty, morality, and “good Samaritan” behavior.

Note. N/A = not applicable because single word adjectives are not available for this complex facet (i.e., self-efficacy in the Army). Virtue and Persistence are facets underlying the Big 5 dimension of Conscientiousness; the other facets are not part of the Big 5.

4.4 Summary

TAPAS was designed to be a flexible platform where new facets could be added, old facets could be deleted, and various sets of facets could be researched. The goal, of course, was to identify a set of facets with strong predictive power that could provide incremental validity over the ASVAB. Since its original implementation for enlistment screening, a variety of studies have been conducted to examine the validity and incremental validity of the facets.

5.0 STATEMENT POOLS

A personality assessment measure is only as good as its items. For TAPAS, each item consists of two statements. Thus, development of statements is a critical step.

5.1 General Steps for Statement Pool Development

For TAPAS adaptive tests to be viable, each TAPAS facet needs a pool of 40 to 60 statements for which IRT and social desirability parameters have been estimated. The process of TAPAS statement pool development has typically proceeded as follows:

1. Content domains relevant to each facet are identified by examining the relevant psychological literature and 70 to 80 statements per facet are written by subject matter experts. These statements are written to span the respective trait continua, varying in extremity from low to moderate to high. This not only broadens the variety of statements that can be presented to examinees, but also balances measurement precision along the entire trait continuum, which is particularly helpful in a computerized adaptive testing environment. An example of a statement from the Order facet that reflects a high standing on this trait would be “I keep detailed notes of important meetings and lectures.” All statements are then reviewed for length, clarity, redundancy, and sensitivity. After deleting redundant statements, the statements are carefully edited. Ultimately, 50 to 60 statements remain. They are then submitted for approval by the ARI Institutional Review Board (IRB).
2. After IRB approval, statements are assembled into pretest forms and administered to large representative samples of recent Army recruits who are very similar to the military applicant population. Each data collection would typically pretest 150-250 statements representing multiple TAPAS facets (usually 6 to 10 at a time). To estimate IRT statement parameters (discrimination, extremity, and threshold), samples of 350-500 respondents are asked, under “honest” testing conditions, to indicate their level of agreement with each personality statement using a 4-point scale, where 1 = strongly disagree, 2 = disagree, 3 = agree, and 4 = strongly agree. To estimate social desirability parameters, samples of 30-50 recruits are asked, under “fake good” conditions, to respond to each statement using the same 4-point scale. In the “fake good” section, respondents are told to pretend they are not yet in the Army, but very much want to be, and scores on this assessment will be used to make enlistment decisions. Thus, they should answer in a way that makes them look like “good Army material.” The directions for the Pretest Questionnaire for “honest” and “fake good” conditions and two sample items are shown in Appendices C and D.
3. Next, pretest data from honest conditions are cleaned (various response checks are used to identify and remove unmotivated respondents), responses are dichotomized, and statement discrimination, extremity, and threshold parameters, denoted α_i , δ_i , and τ_{il} respectively, for the Generalized Graded Unfolding Model (GGUM) are then estimated (Roberts, Fang, Cui, & Wang, 2006). The dichotomous case of the GGUM may be written as follows:

$$P[U_i = 1 | \theta_j] = \frac{\exp(\alpha_i[(\theta_j - \delta_i) - \tau_{il}]) + \exp(\alpha_i[2(\theta_j - \delta_i) - \tau_{il}])}{1 + \exp(\alpha_i[3(\theta_j - \delta_i)]) + \exp(\alpha_i[(\theta_j - \delta_i) - \tau_{il}]) + \exp(\alpha_i[2(\theta_j - \delta_i) - \tau_{il}])},$$

where θ_j is the location of respondent j on the continuum underlying responses, α_i is the

discrimination parameter for statement i , δ_i is the location of statement i on the continuum underlying responses, and τ_{il} is the location of the subjective response category threshold on the latent continuum.

Like many other IRT models, GGUM parameters for a given sample are estimated under the assumption that the distribution of person parameters (trait scores) is normal with a mean of 0 and an SD of one. Because statements from the same TAPAS facet are often calibrated using data collected from different samples of recruits, which could differ somewhat in their trait distributions, statement parameters from different data collections are put on a common metric through a procedure known as linking (see Segall, 1983, for an introduction to linking). To do so, mean location and mean discrimination parameters for statements appearing in common across forms are used to calculate linking constants and place all pretested statement parameters onto a common scale; this is often termed mean/mean linking in the IRT literature (Kim & Lee, 2004).

4. In the final step of statement pool development, the polytomous (4-point scale) data from the “fake good” conditions are used to estimate one social desirability parameter per statement by averaging the endorsed response codes over examinees. The social desirability parameters are then added to the GGUM parameters to complete parameter estimation for each statement set. Statement parameters are then closely examined and those having low discrimination parameters (e.g., below .40) or uninterpretable location or social desirability parameters are excluded from the final statement pool.

5.2 Development of the Initial 22-facet TAPAS Research Statement Pool

Over 1,200 statements measuring 22 TAPAS facets were initially developed. These statements were written to reflect low, medium, and high locations on each trait continuum. Pretesting of this TAPAS Research statement pool began in November of 2005 and ended in April of 2008. Recruit volunteers were obtained at Fort Jackson, Fort Leonard Wood, and Fort Benning; all data collections complied with American Psychological Association (APA) ethical guidelines for research with human subjects. The breakdown of various samples and the number of statements pretested are shown in Table 14.

Table 14. Breakdown of Samples Used to Estimate GGUM Parameters for TAPAS Statements

Date	Number of Recruits	Pretest Site	Number of TAPAS Statements Pretested
November 2005	270	Fort Leonard Wood	225
February 2006	272	Fort Leonard Wood	150
March 2006	525	Fort Jackson	150
June 2006	588	Fort Jackson	225
August 2006	532	Fort Jackson	225
January 2007	456	Fort Jackson	221
January 2007	456	Fort Jackson	221
February 2007 Part 1	319	Fort Leonard Wood	221
February 2007 Part 2	385	Fort Leonard Wood	208
May 2007	429	Fort Jackson	200
June 2007	585	Fort Benning	210
February 2008	452	Fort Benning	320

GGUM and social desirability parameters were then estimated for each statement in the pool and poorly performing statement were excluded. In total, 985 usable statements for the TAPAS Research pool were retained; the detailed breakdown of the number of statements per TAPAS facet is shown in Table 15. Two example statements are also shown for each TAPAS facet - one statement with a positive location parameter and the other with a negative location parameter.

Concurrent with the TAPAS research statement pool development, ARI researchers wrote new statements to possibly augment the AIM inventory (see White & Young, 1998, for a description of the AIM). Several dozen statements were pretested at Fort Jackson and Fort Leonard Wood using the same samples of recruits used in TAPAS pool development. Because AIM statements could be straightforwardly mapped onto the TAPAS facets, a decision was made in 2008 to augment the TAPAS statement pool with the ARI statements. Altogether, 149 ARI statements measuring 9 facets were added to form the initial TAPAS statement pool. A breakdown of the resulting statement pool for each TAPAS facet is presented in Table 15.

Table 15. Number of Statements Available for Each of the 22 TAPAS Facets

TAPAS Facet	TAPAS Pool	ARI Pool	Total Available	Examples of Statements with Positive and Negative Locations
Agreeableness				
Consideration	48		48	<i>Most people would say that I am a loving and forgiving person.</i> <i>I make people feel at ease so that they think they can tell me anything.</i>
Cooperation	45	17	62	<i>I am a really easy person to live with.</i> <i>I have often been critical of others.</i> <i>I contribute to charity regularly.</i>
Selflessness	43		43	<i>I only help people when I know I will get something in return.</i>
Conscientiousness				
Achievement	53	22	75	<i>I try to be the best at anything I do.</i> <i>I finish tasks at my convenience.</i>
Non-Delinquency	34	17	51	<i>I support long-established rules and traditions.</i> <i>When I was in school, I used to break rules quite regularly.</i>
Order	41		41	<i>I am definitely more organized than most people.</i> <i>Others always tell me to clean up my work area.</i> <i>I have made great personal sacrifices to do what I have promised.</i>
Responsibility	54		54	<i>When things go wrong, I'd rather blame it on bad luck than admit that I may have been at fault.</i>
Self-Control	56		56	<i>I am really good at tasks that require a careful and cautious approach.</i> <i>I often rush into action without thinking about the consequences.</i>
Virtue	40	8	48	<i>I firmly believe that under no circumstances is it okay to lie.</i> <i>I try to do the right thing, but sometimes it is necessary to cut some corners.</i>

Table 15 (continued)

TAPAS Facet	TAPAS Pool	ARI Pool	Total Available	Examples of Statements with Positive and Negative Locations
Emot. Stability				
Adjustment	41	14	55	<i>Even if I've had a really stressful day at work, I fall asleep easily.</i>
				<i>Because I constantly worry about things, it is hard for me to relax.</i>
Even Tempered	38	14	52	<i>Even during a particularly heated argument, I keep my emotions under control.</i>
				<i>People who know me well would say that I am moody.</i>
Optimism	39	12	51	<i>I never get depressed.</i>
				<i>I have a hard time finding positive things to say about myself.</i>
Extraversion				
Attention Seeking	49		49	<i>I like to be the center of attention.</i>
				<i>I don't like to be noticed.</i>
Dominance	42	24	66	<i>After joining a group, I usually end up becoming the leader.</i>
				<i>I've been told that I need to be more assertive.</i>
Sociability	40		40	<i>I'll talk to anyone.</i>
				<i>It takes a while to get to know me.</i>
Openness				
Aesthetics	43		43	<i>I appreciate the paintings of well-known artists.</i>
				<i>I think viewing art is a waste of time</i>
				<i>I like to analyze things instead of taking them at face value.</i>
Curiosity	43		43	<i>Even when I am interested in something I'll rarely look into it.</i>
				<i>I try not to think too deeply about the future.</i>
Depth	50		50	<i>One of the main goals in life should be understanding its meaning.</i>
				<i>Generating new ideas is effortless for me.</i>
Ingenuity	45		45	<i>I rarely take an idea and apply it in a new way.</i>
Intellectual Efficiency	40		40	<i>I am very quick at processing information.</i>
				<i>I usually struggle to solve complex problems;</i>
				<i>I feel that an opportunity to learn about the culture of others is something to be treasured.</i>
Tolerance	37		37	<i>I like visiting familiar places and avoid trips outside my country as best I can.</i>
Military Specific				
Physical Conditioning	64	21	85	<i>I like to exercise.</i>
				<i>I don't consider myself to be an athletic person.</i>
Total	985	149	1,134	

Note: Items for Non-Delinquency, Cooperation, and Physical Conditioning illustrate themes in place after 2011.

5.3 Development of the 23-facet DoD Exclusive Statement Pool

After the decision was made to revise the TAPAS facet taxonomy to have 23 facets (18 original and 5 new), another statement pool for exclusive use by the DoD was developed. The primary reason was to enhance test security as subsets of statements in the original research pool had been used in other military applications and it was expected that these statements would continue to be used for research and in other non-secure settings. Starting with TAPAS V5, all subsequent MEPS TAPAS versions were implemented using the DoD exclusive statement pool (besides Ingenuity on USMC V1.1).

The development of the second statement pool proceeded in a manner similar to the research pool development. Nearly 1,300 new statements were written for the 23 facets and reviewed for length, clarity, redundancy, sensitivity, and the degree of content overlap with the research pool. Some statements were modified to improve readability, and some were flagged for removal due to high similarity with statements in the research pool. Statements that passed this review were assembled into pretest forms and administered in 2009-2010 to several samples of Army recruits as shown in Table 16. All data collections complied with the Army's and the APA's ethical guidelines for research with human subjects.

Table 16. Samples Used to Estimate GGUM Parameters for the Second TAPAS Statement Pool

Location	Date	Sample Size
Fort Leonard Wood	10-Aug-09	528
Fort Leonard Wood	18-Aug-09	462
Fort Jackson	16-Oct-09	524
Fort Benning	9-Jul-10	837
Fort Leonard Wood	1-Aug-10	789
Fort Sill	15-Aug-10	1,302
Fort Leonard Wood	22-Aug-10	778
Total		5,220

For each testing session, multiple forms of pretest questionnaires were developed. Multiple forms were needed to efficiently collect the data required for estimating GGUM and social desirability statement parameters. Common subsets of 5 to 7 statements per facet were included in questionnaire forms administered within and across testing sessions so that parameter estimates could be placed on a common scale; statements from the original 18 TAPAS facets were placed on the research pool metric using mean/mean linking, thus facilitating comparisons to past TAPAS versions.

As before, each form of a questionnaire contained two sections. The first section asked examinees to respond honestly. The second section asked examinees to fake good. In both

sections, data were collected using a 4-point response format, where 1 = strongly disagree, 2 = disagree, 3 = agree, and 4 = strongly agree. The honest section always preceded the faking section in the questionnaire because it was believed that it would be easier for examinees to shift from an honest to a faking mindset than the reverse. The honest section of each questionnaire form typically contained 180-220 statements measuring six to ten facets of personality, while the faking section of each questionnaire form contained 40 to 50 statements reflecting varying numbers of facets.

Data from the recruit samples were processed and cleaned to remove invalid entries. Data from the honest conditions were dichotomized and analyzed for each facet separately, using the GGUM2004 computer program (Roberts et al., 2006). Three GGUM parameters were estimated for each statement: discrimination α_i , location δ_i , and threshold τ_{il} ; once parameters were estimated, mean-mean linking was done to place statements from different forms on the common metric. The polytomous data from the faking conditions were then used to estimate one social desirability parameter per statement by averaging the endorsed response codes over examinees. Poorly performing statements (e.g., low discriminations, uninterpretable locations) were excluded yielding a total of 1,052 new statements measuring the 23 facets in the updated TAPAS taxonomy. ARI statements, which had been used in the 2009 TAPAS MEPS testing, were moved into the DoD exclusive pool, because they had not been exposed in TAPAS-related testing outside of the MEPS. In total, this effort produced 1,142 usable statements for the DoD exclusive TAPAS statement pool; the final numbers for each facet shown in Table 17.

Table 17. Numbers of Statements Representing each of the 23 Facets in the Second TAPAS Statement Pool

Facet Name	Number of ARI Statements	Number of New Statements	Total
<i>Original Facets</i>			
Achievement	11	46	57
Adjustment	9	44	53
Attention Seeking		47	47
Consideration		56	56
Cooperation	8	38	46
Curiosity		46	46
Dominance	14	37	51
Even Tempered	9	44	53
Intellectual Efficiency		44	44
Non-Delinquency	12	34	46
Optimism	8	39	47
Order		50	50
Physical Conditioning	19	35	54
Responsibility		42	42
Self-Control		42	42
Selflessness		56	56
Sociability		48	48
Tolerance		44	44
<i>New Facets</i>			
Adventure Seeking		51	51
Commitment to Serve		52	52
Courage		56	56
Situational Awareness		48	48
Team Orientation		53	53
Total	90	1,052	1,142

5.4 Developing Statement Pools for Five New Experimental Facets and Virtue

In total, 310 statements for 6 new TAPAS facets were developed following the same process described above. These newly developed statements were then administered to large representative samples of Soldiers in the Regular Army, Army National Guard, and Army Reserve components. Pretesting began in September 2015 and ended in April 2016. Over 2,200 recruit volunteers from Fort Campbell, Fort Carson, Fort Knox, Fort Leonard Wood, Fort Meade and a number of other installations participated in the pretesting. Approximately 73% of the sample were men and 50.4 % were Caucasian.

Multiple forms of pretest questionnaires were used to efficiently collect the data required for estimating GGUM statement parameters and social desirability parameters. As before, each questionnaire contained two main sections. The first section contained up to 160 statements and asked examinees to respond honestly; the second section contained up to 80 statements asking examinees to fake good. In addition, each section contained up to 4 statements designed to flag unmotivated examinees by asking respondents to select a particular option (e.g., strongly agree) on the form.

Data from the pretest samples were then processed and cleaned to remove unmotivated examinees who provided invalid entries for response check statements. Data from the honest conditions were dichotomized and GGUM statement parameters were estimated for each new trait separately, using the GGUM2004 software (Roberts et al., 2006). The polytomous data from the faking conditions were used to estimate one social desirability parameter per statement by averaging the endorsed responses over examinees.

Several statements had to be dropped during parameter estimation to facilitate GGUM2004 program convergence. Statements having GGUM discrimination parameters below .40 were also excluded, because they would have been very unlikely candidates for inclusion in an adaptive test administration. Table 18 shows the breakdown of statements for the six new TAPAS item pools. Specifically, for each facet, we show the number of pretested statements, the number of final statements after problematic statements were dropped, and an example of a statement reflecting a high level of the trait. In total, this effort produced 278 usable statements, with all traits having at least 45 statements.

Table 18. Numbers of Statements Representing each of the Six New TAPAS Item Pools

Trait Name	# of Statement s Pretested	Final # of Statement s	Example Statement
Army	48	45	<i>I think that military training will be easy for me.</i>
Self- Efficacy	54	46	<i>I don't think that I'm better than other people.</i>
Humility	53	46	<i>I have been accused of "playing games" to get what I want.</i>
Machiavellianism	50	45	<i>I hate leaving things incomplete or unfinished.</i>
Persistence	49	46	<i>I expect to master new skills faster than most others.</i>
Self-Efficacy	56	50	<i>I have a reputation for being honest and ethical.</i>
Total	310	278	

5.5 Additional TAPAS Forms to Support DoD Personality Research Projects

This section of the report describes additional TAPAS forms that were requested and used during 2009-2021 to support various personality research projects involving DoD personnel. All of these TAPAS forms were static (non-adaptive), and administered in paper-and-pencil or computerized formats as part of various field studies. In total, 11 such custom TAPAS forms of varying dimensionality and test length were created, and many of these are still being used in ongoing research investigations. Table 19 indicates each form name, the number and the names of facets assessed, and the intended study population.

Table 19. Additional TAPAS Forms and Facets Assessed

Count	TAPAS Form Name	Intended Population	Year Created	# of Items (2 statements each)	# of Dims	Facet Names
1	DS_TAPAS_18s	Drill Sergeants	2009	143	18	Achievement, Adjustment, Attention Seeking, Consideration, Dominance, Even Tempered, Ingenuity, Intellectual Efficiency, Non-Delinquency, Optimism, Order, Physical Conditioning, Responsibility, Self-Control, Selflessness, Sociability, Tolerance, Virtue
2	ROTS_TAPAS_12s	ROTC Cadets	2010	95	12	Achievement, Adjustment, Cooperation, Curiosity, Dominance, Even Tempered, Intellectual Efficiency, Non-Delinquency, Optimism, Physical Conditioning, Responsibility, Tolerance
3	ARSOF_TAPAS_15s	Special Forces Applicants	2012	120	15	Achievement, Adjustment, Adventure Seeking, Attention Seeking, Courage, Dominance, Even Tempered, Intellectual Efficiency, Non-Delinquency, Optimism, Physical Conditioning, Responsibility, Situational Awareness, Team Orientation ¹ , Tolerance
4	35Q_TAPAS_15s_v1	Cryptologic Network Warfare	2014	120	15	Achievement, Adjustment, Attention Seeking, Cooperation, Curiosity, Dominance, Even Tempered, Ingenuity, Intellectual Efficiency, Non-Delinquency, Optimism, Order, Physical Conditioning, Responsibility, Self-Control
5	35Q_TAPAS_15s_v2	Cryptologic Network Warfare	2015	120	15	Achievement, Attention Seeking, Commitment to Serve, Dominance, Even Tempered, Ingenuity, Intellectual Efficiency, Non-Delinquency, Optimism, Order, Physical Conditioning, Responsibility, Selflessness, Sociability, Virtue

Table 19 (Continued)

Count	TAPAS Form Name	Intended Population	Year Created	# of Items (2 statements each)	# of Dims	Facet Names
6	USMC_OCS_TAPAS_18s	USMC Officer Cadets	2016	141/119	18	Achievement, Adjustment, Commitment to Serve, Courage, Dominance, Even Tempered, Ingenuity, Intellectual Efficiency, Non-Delinquency, Optimism, Order, Physical Conditioning, Responsibility, Selflessness, Sociability, Team Orientation ¹ , Tolerance, Virtue
7	USMC_OCS_TAPAS_15s	USMC Officer Cadets	2017/18	123	15	Achievement, Adaptability, Competence, Decisiveness, Dominance, Even Tempered, Initiative, Intellectual Efficiency, Optimism, Physical Condition, Resilience ¹ , Responsibility, Sense of Purpose ¹ , Sociability, Team Orientation ¹
8	NSMRL_TAPAS_11s_A	Navy Submarines	2018	91	11	Achievement, Dominance, Even Tempered, Humility, Intellectual Efficiency (instead of Curiosity or Depth), Non-Delinquency, Optimism, Order, Physical Conditioning, Selflessness, Sociability
9	NSMRL_TAPAS_11s_B	Navy Submarines	2018	91	11	Adaptability, Adjustment, Attention Seeking, Initiative ¹ , Persistence, Resilience ¹ , Responsibility, Self-Efficacy, Team Orientation ¹ , Tolerance, Virtue
10	AF_TAPAS_15s	Air Force Personnel	2018	120	15	Achievement, Adjustment, Attention Seeking, Cooperation, Dominance, Even Tempered, Intellectual Efficiency, Non-Delinquency, Optimism, Order, Physical Conditioning, Self-Control, Selflessness, Sociability, Tolerance
11	AF_Dark_TAPAS_8s	Air Force Personnel	2019	58	8	Psychopathy ¹ , Sadism ¹ , Narcissism ¹ , Machiavellianism ¹ , Achievement, Even Tempered, Selflessness, Virtue
Note: ¹ Not part of the 32 facets developed for the Army (the Air Force Machiavellianism item pool is different from the Army Machiavellianism item pool)						

In 2009, ARI's field unit at Ft. Benning initiated a study of non-cognitive predictors of performance of DS. Because critical criteria of interest for the study were mentoring, teaching, leading, counseling and indiscipline, personality facets were expected to be important. After reviewing past DS studies, an 18-facet, 143 item static TAPAS form known as DS_TAPAS_18s was developed. It was administered in 2010 to several hundred DS personnel together with a number of other predictors; self- and observer-criteria rating were also collected. This initial work ultimately led to the development of the Noncommissioned Officer Special Assignment Battery (NSAB; Nye et al., 2018a). Further information about the NSAB is provided in Horgen et al. (2013) and Nye et al. (2020a).

In 2010, ARI and the Human Resources Research Organization conducted a large-scale study of Reserve Officer Training Corps (ROTC) cadets to identify potential predictors of leadership performance. Over 1,500 cadets participated in the study during their leader develop and assessment course. A 12-facet, 95 item static TAPAS form (ROTC_TAPAS_12s) was created and administered together with several other predictors (e.g., biodata, values). Results of the study are available in Legree et al. (2014).

In 2012, a study was conducted to explore the usefulness of the TAPAS for identifying Soldiers who might be selected for Army Special Operations Forces (ARSOF) training. After several discussions about characteristics of candidates who are successful, a 15-facet, 120-item TAPAS version (ARSOF_TAPAS_12s) was created and administered to 1,216 special forces candidates prior to their ARSOF assessment and selection course at Fort Bragg. The criterion for this research was the Special Forces Assessment and Selection outcome for each candidate in the course and several TAPAS facets were found to have significant correlations with that outcome (see Nye et al., 2014). Overall, it was found that a TAPAS composite could substantially improve prediction of Soldiers who would be successful in the ARSOF assessment and selection course.

In August 2014, a static version of TAPAS was requested to support research with 35Q MOS soldiers (Cryptologic Network Warfare Specialist³). The research was initiated to identify non-cognitive predictors of the 6-month Joint Cyber Analysis Course completion. Following discussions with ARI researchers, a 120-item, 15-facet static test was created and administered to multiple cohorts of 35Q trainees in November 2014, January 2015, and April 2015. In July 2015, the 35Q test was revised and the second static version was created to support course completion research. The test length and dimensionality remained the same, but four facets from the first version (Adjustment, Cooperation, Curiosity, and Self-Control) were replaced with four new facets (Commitment to Serve, Selflessness, Sociability, and Virtue). The list of facets implemented in the two 35Q static versions is shown in Table 19.

In August 2016, the USMC initiated a series of personality assessment projects involving Officer Candidate School (OCS) and United States Naval Academy cadets. Following discussions with USMC researchers, an 18-facet, 141 item static TAPAS form (USMC_OCS_TAPAS_18s) was created and administered to two cohorts of cadets in September 2016 and January 2017 (N =

³ To the best of our knowledge, a report on this work has not been published.

584). In March 2017, the test was revised and a second static version was created to support the ongoing USMC research efforts. The test dimensionality remained the same, but the test was shortened to 119 items. From May 2017, all subsequent cohorts of USMC cadets have completed the USMC_OCS_TAPAS_18s version and DCG continues to support test scoring and database management. To date, over 7,000 cadets have taken the test.

In 2017, USMC requested assistance in the development of another TAPAS research form for officer cadets. It included 6 new research facets (Adaptability, Resilience, Sense of Purpose, Decisiveness, Competence, and Initiative) for which statement pools were developed and pretested as well as 9 existing facets for which research statement pools were already available. The resulting 15-facet, 123-item static TAPAS form (USMC_OCS_TAPAS_15s) was to be used as an alternative to the USMC_OCS_TAPAS_118s form described above. Results of the USMC research are provided by Harvey et al. (2018).

In 2018, the Naval Submarine Medical Research Lab (NSMRL) initiated a study of personality characteristics that may be useful for predicting the performance of Navy Submariners. Based on previous work by NSMRL, a total of 22 personality facets that are relevant to Navy Submariners were identified. Of these, 17 facets were already part of the TAPAS taxonomy and had associated research statement pools available. The remaining 5 facets were either new or did not have research statement pools available, so 18-25 statements for each facet were subsequently developed and pretested to create two TAPAS forms having 11 non-overlapping facets and 91 items each. The two forms are named NSMRL_TAPAS_11s_A and NSMRL_TAPAS_11s_B, and facets composing each form are listed in Table 19. Data collections are currently ongoing with NSMRL periodically sending de-identified response patterns to DCG for scoring.

In 2018-2019, the Air Force Research Laboratory commissioned three static TAPAS forms to support ongoing research evaluating the viability of personality testing in applicant settings. All three forms assessed the same 15 personality facets currently administered by the Air Force at the MEPS; the facets are shown in Table 19. The first form consisted of 90 items utilizing a traditional 4-point Likert response scale (strongly disagree, disagree, agree, strongly agree). The second form consisted of 120 unidimensional pairwise preference items; for each item, both statements assessed the same underlying facet, and respondents were asked to choose the statement in each pair that is “more like me.” The third form consisted of 120 multidimensional pairwise preference items and was labeled the Air Force TAPAS static research form (AF_TAPAS_15s). A sample of about 350 Basic Recruits was first asked to complete all three forms honestly for research purposes. The cross-form correlations of facets (i.e., convergent validities) were reasonably large. The sample was then asked to “convince the Air Force that you would make a good Airman” (i.e., fake good). The multidimensional pairwise preference form appeared most resistant to faking (i.e., had the least score inflation), but none of the forms was effective in predicting criterion variables in the faking condition. Details of the study can be found in Chernyshenko et al. (2019).

In a second Air Force study, two forms were created to assess four Dark side personality traits (Psychopathy, Sadism, Narcissism, and Machiavellianism). The forms also included four normal personality facets from the TAPAS taxonomy (Achievement, Even Tempered, Selflessness, and Virtue). Thus, each form was an 8-facet, 58-item static assessment. One form used a traditional

5-point Likert scale (1 = strongly disagree; 2 = disagree; 3 = neither disagree nor agree; 4 = agree; 5 = strongly agree). The other form (AF_TAPAS_Dark_8s) used a pairwise preference format. On-line crowd sourcing platforms were used to collect data in honest and fake good conditions. After cleaning, there were 504 cases for the honest condition and 478 cases for the fake good condition. Reasonable large convergent validities were found for the normal personality facets in the honest condition, but the convergent validities were at best modest in this condition for the Dark side traits. Convergent validities were lower in the fake good condition. The Likert form showed higher correlations with a wide range of criterion variables in the honest condition. In the faking condition, neither form had substantial correlations. Details of the study can be found in Dragow et al. (2020).

5.6 Summary

As can be seen from this section, it is a bit misleading to talk about “the TAPAS” because a large (and growing) number of forms have been assembled. Statement pools are continuously being developed as more facets are added to the library. Moreover, the personality characteristics needed for effective performance will certainly vary across military occupations: The “right stuff” to be a first-term enlisted Soldier in Infantry is almost certainly different than what is needed for more senior Army occupations, Air Force occupations, and so forth. Thus, it seems likely that TAPAS forms will continue to be added to meet the needs of the Services.

6.0 ARMY TAPAS COMPOSITE SCORES

The TAPAS computer program provides facet raw scores (thetas), facet scale scores (i.e., normed Z-scores), and SEs. However, it is not recommended to make decisions based on the facet raw or scale scores. Instead, TAPAS also calculates several composites that are intended to be considered during enlistment screening. These composites are weighted combinations of the normed Z-scores for several TAPAS facets and, thus, are approximately normally distributed. To facilitate decision-making, all Army composites (e.g., Can-Do) are scaled to have a mean of 100 and an SD of 20. Hence, a composite value of 100 would correspond to the 50th percentile, while a composite value of 70 would correspond to the 10th percentile. The specific facet comprising TAPAS composites are not detailed in this report as this is close-hold information.

The current TAPAS composites have their roots in Phase 1 of the Army TAPAS testing described in Section 3.1.1. Two unit-weighted composites were initially implemented: Can-Do and Will-Do (see Allen et al., 2010). The TAPAS Can-Do composite was designed to predict Soldiers’ training performance, while the Will-Do composite was designed to predict Soldiers’ motivation. The TAPAS Can-Do and Will-Do composites were later refined in subsequent research and an additional composite, known at various times as either the TAPAS Attrition, Persistence, or Adaptation composite, was added to predict attrition (Nye et al., 2012b, 2013). Finally, a fourth composite was added in Phase 4 of TAPAS testing to predict attrition due to misconduct (i.e., the TAPAS Conduct composite). Note that because Z-scores are used to calculate the composites, the most recent composites can be calculated retrospectively for archival TAPAS data as long as the facets comprising the newer composite were administered.

Consistent with the criterion-focused validation approach (Sackett et al., 2012), the composites of TAPAS facets have been developed to predict Can-Do performance, Will-Do performance, and Attrition/Adaptation (both terms have been used). The Can-Do and Will-Do performance criteria are composites of several outcomes and were developed in light of previous validation research. For example, in Project A (Job Performance Measurement Project after ASVAB misnorming), two composites labeled Can-Do and Will-Do performance were created for examining Soldier performance (Campbell & Knapp, 2001). Similar Can-Do and Will-Do criteria were also examined in the EEEM project (Allen et al., 2010). Therefore, these same criteria were the focus of TAPAS composite development. In addition, because attrition represents a substantial cost for the Army, this outcome was also used as a criterion.

More specifically, the Can-Do and Will-Do criteria for the TAPAS composite development were based on factor analyses of various criterion measures and in consultation with subject matter experts (e.g., Army Non-Commissioned Officers, ARI psychologists) to develop a conceptual model of Soldier performance (Nye et al., 2012a). When creating these criterion composites, more emphasis was placed on creating a manageable number of outcomes for prediction rather than a unidimensional combination of dependent variables. Therefore, Can-Do performance was comprised of scores on the Army-wide and MOS-specific job knowledge tests. Will-Do performance consisted of scales on the ALQ (e.g., adjustment, commitment, reenlistment intentions), APFT scores, and disciplinary incidents. For validation studies in initial military training, training achievement and training failure were also included in the Will-Do criterion composite. Peer and supervisor ratings of performance have also been included in the Will-Do composite in some projects (Nye et al., 2012a), but were excluded from other projects due to the relatively small number of performance ratings available for analyses (e.g., Nye et al., 2020b). For both Can-Do and Will-Do performance, scores for each criterion were first standardized to account for differences in their SDs and then summed to create overall scores for the composites. Given their importance to the Army, APFT scores and disciplinary incidents were double-weighted in the Will-Do criterion composite whereas the other components were unit weighted.

Attrition has been examined as a separate outcome. Depending on the data available, the attrition variable that was used for TAPAS composite development generally reflected first-term attrition that occurred between 6 and 36 months in the Army. Although early research on the TAPAS predictors of attrition examined attrition in its original form (e.g., Nye et al., 2012a), subsequent research has reverse coded this variable (1 = Did Not Attrit, 0 = Attrit) and relabeled it Adaptation so that all the focal criteria were scaled in the same direction (e.g., Nye et al., 2020b).

Composites of the TAPAS facets were created to predict the Can-Do, Will-Do, and Adaptation criteria. These *predictor* composites have been labeled the TAPAS Can-Do, Will-Do, and Adaptation composites. It is important to note that the TAPAS Can-Do, Will-Do, and Adaptation composites were not designed to measure specific constructs but instead predict outcomes important to the Services. As noted in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014), validity evidence should provide support for the intended interpretations and uses of a test. The TAPAS composites were specifically designed to predict Soldiers' attitudes and behavior on the job and, therefore, to identify the applicants with the highest potential for success in the Army.

Nye et al. (2012b) summarized the development of the initial TAPAS composites and found R s for predicting the Can-Do, Will-Do, and Attrition criteria of .29, .31, and .11, respectively. In addition, this work also demonstrated that high scores on the TAPAS composites could compensate for low scores on the AFQT. These results are illustrated in Figure 1. Graphs are shown for the average APFT and ALQ Army-life adjustment scores as well as the percentages of disciplinary incidents and 6-month attrition. The horizontal axes for these figures reflect the AFQT categories. To create these figures, Nye et al. (2012b) used the Will-Do and Attrition composites as equally weighted multiple hurdles to hypothetically select out the bottom 10% of individuals in AFQT categories IIIB and IV (i.e., the lowest scoring categories on the AFQT). Those in the bottom 10% were considered to have failed the TAPAS composite screen while individuals scoring in the top 90% on the TAPAS screen were considered to have passed.

As shown in Figure 1, individuals that passed the TAPAS composite screen generally performed substantially better than those with predicted scores below the cutoff. The average APFT scores for the passing group were 15 points higher than the average score in the failing group. In addition, individuals that passed the TAPAS screens were nearly 40% less likely to leave the Army before their term of enlistment was completed. Overall, performance for individuals in AFQT Categories IIIB or IV that passed the TAPAS screen was comparable to individuals in Categories II or IIIA. Thus, the TAPAS composites appear useful for identifying and selecting high potential Soldiers who score in AFQT Categories IIIB and IV.

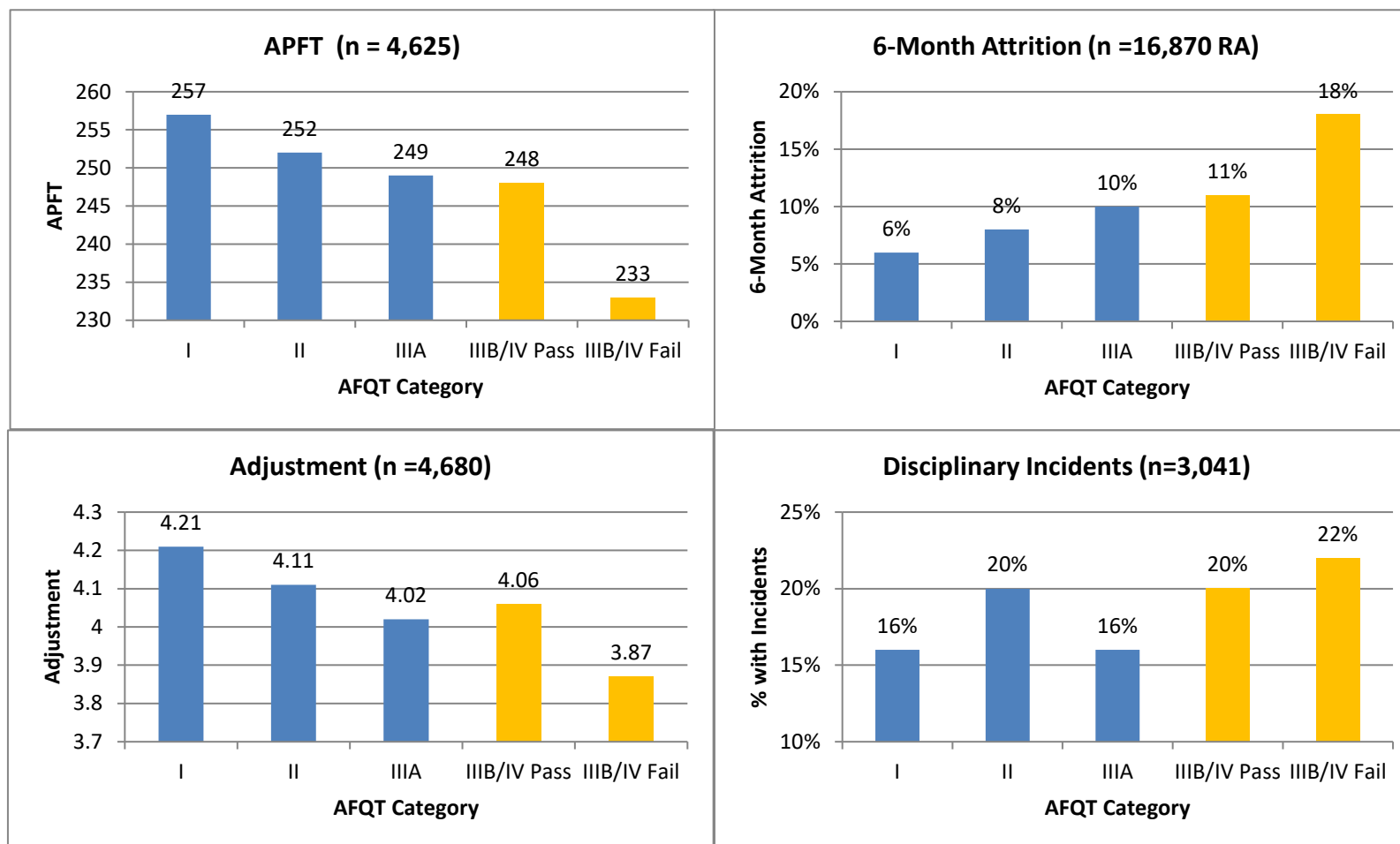


Figure 1. APFT scores, 6-Month Attrition, Army Life Adjustment, and Disciplinary Incidents by AFQT category (from Nye et al., 2012b)

Subsequent research revised the TAPAS Can-Do, Will-Do, and Attrition composites using larger sample sizes to obtain more stable results (Nye, White, et al., 2020c). These revised composites had similar relationships with their corresponding criteria (i.e., compared with the original composites), with R s of .26, .31, and .12 for predicting Can-Do, Will-Do, and Adaptation outcomes, respectively.

More recently, the TAPAS composites were updated and expanded again using data collected by ARI as part of the Validation of Accession Screening Tools project. First, the scales composing the TAPAS composites were updated using larger samples of TAPAS and criterion scores. These larger sample sizes provide more accurate and consistent estimates of the composite weights. Second, this work helped to identify any potential changes in the predictive validity of the specific TAPAS facets making up each composite that might have occurred over time. Finally, a fourth TAPAS composite was added to predict misconduct attrition. Again, given the cost of attrition, the Army is particularly interested in predicting this outcome. In addition, analyses indicated that the predictors of attrition varied by the reason for leaving. In other words, different TAPAS facets predicted attrition due to misconduct and attrition for other reasons. Therefore, examining the different types of attrition helped to improve the prediction of these outcomes. The four revised composites continued to show validity for predicting their corresponding outcomes with R s of .24 (Can-Do), .28 (Will-Do), .13 (36-month Adaptation), and .18 (Misconduct Attrition). Nevertheless, these new composites were only developed recently so more research is needed to evaluate their validity in operational conditions. Consistent with the Adaptation composite, the TAPAS composite for predicting misconduct attrition is reverse-coded (1 = Did Not Attrit due to misconduct, 0 = Attrit due to misconduct) so that all TAPAS composites are scaled in the same direction. The composite for predicting this reverse-scored outcome is labeled the TAPAS Conduct composite.

Importantly, there is evidence that customizing TAPAS composites for specific MOS improves prediction. Nye et al. (2012a) examined MOS-specific TAPAS Can-Do, Will-Do, and Attrition composites in four large MOS including Infantry (11B), Military Police (31B), Combat Medics (68W), and Motor Transport Operators (88M). They found customizing composites (computed from the facet scores) for each MOS to yield improved, and substantial, validities. For example, developing separate equations (i.e., composites of the facets) for the Can-Do, Will-Do, and Attrition criteria in Infantry yielded R s of .28, .33, and .22, respectively. This work also indicated that these relationships could have substantial practical importance. For example, in Infantry, which was the largest MOS in the sample, results showed that individuals scoring highest on the MOS-specific TAPAS Attrition composite were 78% less likely to leave the Army during their first 6-months of service than those with the lowest scores. Similar results were obtained for the other MOS examined in that research.

Nye et al. (2020b) later replicated this variation across MOS with larger sample sizes and an additional MOS (MOS 91B; Wheeled Vehicle Mechanics) and found comparable results. For example, the R s, using facet scores as predictors, for Infantry were .25, .30, and .18 for Can-Do, Will-Do and Adaptation criteria, respectively. Evidence for the important differences in the facets with nonzero regression weights and in the size of their weights across MOS is provided by the finding of Nye et al. (2020b) that using the MOS-specific composites for MOS

classification could improve predicted performance by more than .50 SDs for at least 40% of the sample.

Since first administered in the MEPS in 2009, the Army has used TAPAS composites in several ways. For Tier 1 applicants, the Army first began using TAPAS operationally in 2009 to screen out low-motivated AFQT Category IV applicants who scored below the 10th percentile on TAPAS. In 2011, AFQT Category IIIB applicants were added to this Tier 1 screen. The Tier 1 operational screen was suspended in 2015 in order to meet the accessions mission. For Tier 2 candidates, the Army began using TAPAS operationally in 2014 to screen out AFQT Category I-III A low-motivated applicants who scored below the 30th percentile on TAPAS. The Tier 2 operational screen was suspended in 2017. Throughout this time, ARI has actively conducted research on their predictive validity. See, for example, Kirkendall et al. (2020) and Nye et al. (2020c; 2020d).

7.0 DETAILED DESCRIPTIONS OF TAPAS VERSION CONFIGURABLE SPECIFICATIONS AND COMPUTER ADAPTIVE TESTING ALGORITHM

TAPAS testing at MEPS is ordinarily computer adaptive, which means that each examinee sees a unique sequence of items tailored to maximize the precision of his/her facet trait scores. The only exception was Version 3 in which every examinee saw the same sequence of items; it was administered briefly in 2009. As the TAPAS testing program evolved, the Army and other Services requested periodic updates for TAPAS versions by changing facet configurations, statement pool compositions, test lengths, and blueprints, as well as composite calculations. Also, as a result of additional research (Stark et al., 2017), several diagnostic flags and response checks were added to identify those who were potentially unmotivated. Altogether, there were over a dozen of TAPAS versions implemented at MEPS from 2009 to 2019.

This section will describe in detail how TAPAS CAT versions are created and configured. Specifically, it discusses:

- a) how TAPAS test blueprints and statement pools are specified, and
- b) the logic and mathematics behind the adaptive item selection algorithm and the relevant configurable test specification variables related to 2AFC item creation (e.g., repetition of statements, item location and social desirability pairing constraints),

To illustrate each of these discussion points, we will use one of the most recent Army MEPS testing versions, Version 12 or V12, because it contains the most up-to-date TAPAS configuration. A brief and non-technical description of TAPAS MEPS testing can be found in Section 3 of this report.

7.1 Specifying Facets, Statement Pools, and Test Blueprints

The first step in creating a TAPAS version is to decide which personality facets to measure and which of the available statement pools to use. The current MEPS TAPAS library contains 32

facets and the majority of these have both research (i.e., used in a wide variety of military and civilian studies) and operational (i.e., DoD use only) statement pools available (see Sections 4 and 5 for more detail). V12, which was implemented at MEPS in October 2019, was configured to measure 17 facets. Each facet statement pool has 40-50 statements. Each statement has a unique global identifier (Id), Name (measured facet and statement number), Content (what the statement says), three IRT parameters ($\alpha_i, \delta_i, \tau_i$ from calibrating data collected in “honest” responding conditions during pretesting), and a social desirability parameter (desirability rating based on data collected in “fake-good” conditions during pretesting) that are used during test administration procedures. Example specifications for a statement measuring the Tolerance facet are shown in Figure 2 below.

```
"Name": "Tolerance 34",
"Content": " My close friends come from a diverse range of backgrounds.",
"Alpha": 0.87,
"Delta": 1.39,
"Tau": -3.23,
"SocialDesirability": 3.58,
"Id": 72523.
```

Figure 2. Example Specifications for a Tolerance Statement

After facets and statement pools are chosen by the test designer, decisions are made about the number of substantive 2AFC items (pairs) to administer for scoring/assessment purposes, and the number of response check items to add, if any, to readily identify careless or aberrant responders. The substantive 2AFC items can be unidimensional (both statements measure the same facet) or multidimensional (the statements measure different facets). The response check items use the same 2AFC format, but statements indicate that the respondent should select a particular alternative (e.g., "For data quality check, please select this option.").

V12 was designated to have 170 substantive 2AFC items and 6 response check items (i.e., 176 items in total). The test has two parts. Part 1 measures 13 facets and consists of 129 substantive items and 4 response check items. Part 2 assesses 4 additional facets using 41 substantive items and 2 response check items. Test results for Part 1 (the first 133 items) and the whole test (all 176 items) are saved in separate output files. Versions 13 and 14 (V13 and V14) have identical configurations to Version 12 for Part 1 (i.e., same 13 facets are measured), but they measure different facets in Part 2. Note that all earlier TAPAS MEPS versions had much simpler test configurations; they had only one part instead of two, they were shorter (typically 120 items), and most did not contain any response check items, as the latter were not implemented until October 2019.

Once these basic configuration elements are decided, the next step is to create a test blueprint, or table of content specifications (constraints), that identifies permissible multidimensional and unidimensional “item types” (i.e., combinations of statements representing various facets or

dimensions; D). For example, with a test measuring 3 facets (3-D), there are three possible multidimensional item types (1-2, 1-3, 2-3) and 3 possible unidimensional item types (1-1, 2-2, 3-3). A test designer may allow all possible combinations, or disallow some combinations based on expert judgment. The test designer must also choose the initial ordering of combinations. For example, a hypothetical 3-D, 7-item test could be composed of items having the following content specifications: 1-2, 1-3, 1-1, 2-3, 1-3, 3-3, 1-2. The order of the facets within these combinations (e.g., 1-2 vs. 2-1) is irrelevant, because it is randomized during item selection. Also, in the rare event that an item satisfying a content specification cannot be found due to an exceedingly small pool or too restrictive matching constraints (discussed later), the adaptive item selection algorithm may substitute another permissible item type.

Decisions about overall test length and which facet combinations to include in the test blueprint are driven by both psychometric and substantive considerations. Simulation research with the IRT model selected for TAPAS testing (Stark et al., 2012) confirmed that adaptive tests can achieve measurement precision similar to nonadaptive tests (with randomly selected items) that are twice as long and, even with adaptive testing, it is desirable to have 10 or more 2AFC items for every facet measured. In psychometric articles, this has been referred to as “items per facet” to provide a way of comparing tests having different dimensionalities (numbers of facets) and different overall lengths. For example, a 5-D test having 10 items per facet would consist of 50 substantive 2AFC items. A 10-D test of 5 items per facet would also have 50 2AFC items, but much lower measurement precision because each facet is represented fewer times. To achieve comparable measurement precision, in theory, one would need to increase the overall test length to 100 2AFC items but doing so could lead to examinee fatigue and careless responding that offsets anticipated reliability gains.

Beyond test length considerations, care must be taken when choosing the combinations of facets to include in a test. Repeating the same combinations too many times, pairing highly correlated facets, or having only multidimensional combinations could negatively affect the test taker experience. For these reasons, TAPAS test blueprints have typically specified one or two unidimensional items per facet and approximately equal numbers of the permitted multidimensional combinations. However, in some earlier TAPAS versions, no multidimensional combinations were allowed for the Commitment to Serve and Courage facets, and Versions 4 and 5 did not have any unidimensional pairs for some facets (e.g., Achievement, Physical Conditioning, and Self-Control). Ultimately, decisions about how many and which uni- and multidimensional combinations to include in a particular version's blueprint have been made by ARI personnel. Additional constraints are discussed in Section 7.2.2.

Finally, to allow estimation of trait scores as soon as possible during a test, facet combinations in the test blueprint are grouped into “linking” and “main” subsets. The *linking subset* contains item types based on a circular linking design (e.g., for a 5-D test, the linking subtest may contain 1-2, 2-3, 3-4, 4-5, 5-1 multidimensional combinations and one 2-2 unidimensional combination). The *main subset* contains all remaining uni- and multidimensional item types and is administered after the linking subset. For TAPAS versions containing Part 2, an *additional subset* was specified containing facet combinations with one or both statements representing research facets.

Table 20 presents blueprint specifications for V12. The first 13 rows show information for 13 facets assessed in Part 1, while the last 4 rows show information for the 4 additional facets assessed in Part 2. Table columns show the number of statements appearing in either “uni” or “multi” -dimensional combinations in the linking, main, and additional subsets; the last column shows the total number of statements to be administered. For example, for the Adjustment facet, 2 statements are designated to appear in multidimensional combinations in the *linking* subset. For the *main* subset, 4 statements are designated as uni and, thus will form 2 unidimensional 2AFC items, while 13 statements are designated for multidimensional 2AFC items. Finally, 2 more statements are designated for multidimensional combinations in the *additional* subset. In total, each Version 12 TAPAS test will have, for example, 21 Adjustment statements appearing in 2 unidimensional and 17 multidimensional 2AFC items.

Table 20. Statement-Level Test Blueprint Specifications for Version 12

Count	TAPAS Facet	Blueprint Subtest						Total
		Linking		Main		Additional		
		uni	multi	uni	multi	uni	multi	
1	Achievement	-	2	6	14	-	2	24
2	Adjustment	-	2	4	13	-	2	21
3	Attention-Seeking	-	2	4	14	-	2	22
4	Dominance	-	2	4	13	-	2	21
5	Even Tempered	-	2	4	13	-	2	21
6	Intellectual Efficiency	-	2	4	13	-	2	21
7	Non-Delinquency	-	2	4	17	-	2	25
8	Optimism	-	2	4	13	-	2	21
9	Order	-	2	4	13	-	2	21
10	Physical Conditioning	-	2	4	13	-	2	21
11	Sociability	2	2	2	13	-	2	21
12	Team Orientation	-	2	4	16	-	2	24
13	Tolerance	-	2	4	13	-	2	21
14	Commitment to Serve	-	-	-	-	4	10	14
15	Humility	-	-	-	-	2	12	14
16	Persistence	-	-	-	-	2	12	14
17	Self-Efficacy	-	-	-	-	2	12	14

Figure 3 below shows an example of the linking subset for the V12 test blueprint specifications. It contains 13 multidimensional item types organized in a circular linking pattern and one unidimensional item type (Sociability-Sociability).

```

"InitialLinkedItemTypes": [{
    "Index": 1,
    "Dimension1": "Intellectual Efficiency",
    "Dimension2": "Attention Seeking",
    "Id": 15179},
    {"Index": 2,
    "Dimension1": "Non-Delinquency",
    "Dimension2": "Sociability",
    "Id": 15180},
    {"Index": 3,
    "Dimension1": "Optimism",
    "Dimension2": "Order",
    "Id": 15181},
    {"Index": 4,
    "Dimension1": "Physical Condition",
    "Dimension2": "Adjustment",
    "Id": 15182},
    {"Index": 5,
    "Dimension1": "Team Orientation",
    "Dimension2": "Achievement",
    "Id": 15183},
    {"Index": 6,
    "Dimension1": "Dominance",
    "Dimension2": "Tolerance",
    "Id": 15184},
    {"Index": 7,
    "Dimension1": "Tolerance",
    "Dimension2": "Optimism",
    "Id": 15185},
    {"Index": 8,
    "Dimension1": "Attention Seeking",
    "Dimension2": "Even Tempered",
    "Id": 15186},
    {"Index": 9,
    "Dimension1": "Sociability",
    "Dimension2": "Sociability",
    "Id": 15187},
    {"Index": 10,
    "Dimension1": "Non-Delinquency",
    "Dimension2": "Dominance",
    "Id": 15188},
    {"Index": 11,
    "Dimension1": "Sociability",
    "Dimension2": "Intellectual Efficiency",
    "Id": 15189},
    {"Index": 12,
    "Dimension1": "Adjustment",
    "Dimension2": "Team Orientation",
    "Id": 15190},
    {"Index": 13,
    "Dimension1": "Physical Condition",
    "Dimension2": "Even Tempered",
    "Id": 15191},
    {"Index": 14,
    "Dimension1": "Order",
    "Dimension2": "Achievement",
    "Id": 15192}],

```

Figure 3. Example of V12 Test Blueprint Specifications for the Linking Subtest

7.2 TAPAS CAT Administration Algorithm and Configurable Test Specification Variables

In this section, we describe how the TAPAS CAT administration actually works. We begin with a step by step, nontechnical description of the item administration process. We then highlight important configurable test specification variables that govern item creation and selection. Finally, we provide mathematical details for readers interested in knowing how item response and information functions are computed within the program. Details of TAPAS test scoring are presented in next section of this report.

7.2.1 Logic of the TAPAS CAT Algorithm

Every TAPAS testing session starts with the computer program reading in the test blueprint (i.e., linking, main, and additional subsets). To enhance test security, item types are then randomly ordered within each subset, so an examinee will see a different sequence of facet combinations compared to other examinees. So, for an examinee who is taking V12, the first item could be any of the 14 facet combinations from the linking subset shown above, while the 15th item could be any combination from the main subset, and so on. Moreover, for each 2AFC item, the positions of the first and second facet displayed to an examinee are randomized. Hence, it is unlikely that examinees will receive the same sequence of item types with facets displayed in the same positions. And, when considering that statements fulfilling these content specifications are chosen adaptively, duplicate tests are highly unlikely.

Once the facet combination for the first item is determined, the algorithm loads statement pools for the relevant facet(s) and constructs a set of potential 2AFC items that meet several item construction specifications (a.k.a., constraints). In a nutshell, there are constraints to control faking and exposure (see details in the next subsection), so not every statement from the first facet is allowed to be paired with every statement from the second facet. Nevertheless, provided that the relevant statement pools have 40-50 statements each and item construction specifications are not too strict, there will be dozens if not hundreds of available 2AFC items at any point in the test.

Next, for each available 2AFC item, the algorithm uses IRT statement parameters and trait scores to calculate the expected amount of measurement information provided by each item (mathematical details of these calculations are provided below). At the start of the test, trait scores are not yet known, and are therefore assumed to be close to the mean of the prior distribution (i.e., set to zero plus/minus a small random number). However, as the test progresses, the actual estimated trait scores are used to calculate the respective item information values.

Next, the available item providing the maximum information is identified. Although it would be best to administer this item from a measurement efficiency standpoint, it would inevitably be overexposed, because all examinees have similar trait scores at the start of a test and those encountering the same facet combination would receive that same item. To avoid that, the algorithm establishes an information criterion (initially, 70% of maximum information), identifies a group of available items meeting that criterion, and randomly selects an item from that group to be administered. (As the test progresses, this information criterion may be relaxed

automatically by the adaptive algorithm, in increments of 5% - 10%, so that at least one and typically 3 or more additional items are available to choose from.)

After that item is administered and a response is recorded, examinee trait scores are estimated and updated in the test database. The 2AFC item is marked as "used" in the database and cannot be administered again. However, statements composing that item may appear in other 2AFC items subject to statement exposure constraints discussed below.

The test will then continue to the next facet combination, following exactly the same process of identifying an available item set, calculating each available item's information, identifying the maximum information, establishing an information criterion, finding a group of items meeting that criterion, and picking an item randomly from that group to administer. Further, the test continues until the last designated facet combination is administered (e.g., Item 176 in V12 or Item 120 in V5). After the last item has been administered, the final trait scores and SE estimates are computed and saved in the designated output file, together with other relevant information for that version (e.g., item responses, latencies, composite scores, aberrance and response check diagnostics). Depending on the version, the program also saves test results periodically in a temporary file (e.g., after every 5 items), in case the test is interrupted and needs to be resumed later.

7.2.2 Configurable Specifications for Item Construction and Selection Constraints

Figure 4 below shows an example of eight configurable Item Construction and Selection Constraints implemented in V12. The same variables were used in other TAPAS versions, though the values may differ somewhat. Details about these constraints are provided below.

```
"MaximumTimesStimulusCanBeUsed": 2,  
"StimuliNotUsedInLastXItems": 75,  
"MaxMultiSocialDiff": 2,  
"MaxMultiDeltaDiff": 3.5,  
"MinUniDeltaDiff": 1,  
"MaxUniDeltaDiff": 3.5,  
"MaxUniSocialDiff": 2,  
"PctMax": 0.7,
```

Figure 4. Item Construction and Selection Constraints for V12

MaximumTimesStimulusCanBeUsed is an item construction constraint that manages the exposure of individual statements. In all TAPAS versions to date, the value for this constraint has been set to "2", meaning that statements in the pool can appear in 2AFC items only twice.

StimuliNotUsedInLastXItems is an item construction constraint that also focuses on statement exposure. It specifies how many other 2AFC items need to be administered before a previously used statement can be reused in another item. Statements with very high IRT discrimination

parameters (e.g., alpha parameters larger than 2.0) tend to be used more frequently by the algorithm, because they carry a lot of measurement information regardless of the other statements they are paired with. Increasing the value for this constraint prevents highly discriminating statements from being reused too soon, so memory and other response sets are mitigated. In V12 the value for this variable is set to 75, meaning that a statement cannot appear again until another 75 test items have been administered. In other versions that have shorter test length, the value of this constraint is smaller.

MaxMultiSocialDiff is an item construction constraint for multidimensional 2AFC items that caps the allowed difference between the social desirability parameters of the statements composing the items. To mitigate faking, it is desirable for the two statements composing a 2AFC item to have very similar social desirability parameters. However, setting this value too small can severely limit the number of available items during item selection unless there are very large statement pools with social desirability parameters distributed well across the trait continua.

MaxMultiDeltaDiff is an item construction constraint for multidimensional 2AFC items that caps the difference between the extremity/location (delta parameters) of the statements composing items. To mitigate faking, it is desirable for the two statements composing a multidimensional 2AFC item to have similar extremity parameters. However, setting this value too small can severely limit the number of available items during item selection unless there are very large statement pools with extremity parameters distributed across the trait continua. Also, note that statement extremity and social desirability parameters tend to correlate.

MinUniDeltaDiff is an item construction constraint for unidimensional 2AFC items that sets the minimum allowed difference between the extremity/location (delta parameters) of the statements composing items. To mitigate faking, it is desirable for the statements composing unidimensional 2AFC items to have similar extremity parameters, but not the same. However, there is a tradeoff in terms of psychometric information; pairing statements that are located further from each other on the trait continuum (larger delta differences) tends to provide more information than pairing statements located near each other.

MaxUniDeltaDiff is an item construction constraint for unidimensional 2AFC items that caps the difference between the extremity/location (delta parameters) of the statements composing items. To mitigate faking, it is desirable for the two statements composing a unidimensional 2AFC item to have similar extremity parameters. However, setting this value too small can severely limit the number of available items during item selection unless there are very large statement pools with extremity parameters distributed across the trait continua. This value must be larger than MinUniDeltaDiff.

MaxUniSocialDiff is an item construction constraint for unidimensional 2AFC items that caps the allowed difference between the social desirability parameters of the statements composing the items. To mitigate faking, it is desirable for the two statements composing a 2AFC item to have very similar social desirability parameters. However, setting this value too small can severely limit the number of available items during item selection unless there are very large statement pools with social desirability parameters distributed well across the trait continua. This

value can be the same or different from MaxMultiSocialDiff, but it is important to remember that social desirability and extremity tend to correlate and pairing statements located further apart on the trait continuum tends to provide more information.

PctMax is an item selection constraint used to establish an information criterion for choosing an item to administer. It is the percentage of the maximum information provided by any available item at an examinee's trait scores. Recall that during item selection, items satisfying the information criterion are marked as candidates for administration and an item is selected randomly from that group. If too few items are found, PctMax is decreased automatically in increments of 5% - 10% until more than one item is identified or the percentage falls below a low threshold (e.g., 30% of maximum).

In summary, there are many constraints governing the construction and selection of items in TAPAS tests. Content constraints set the desired facet combinations. The settings of variables above regulate the exposure of statements, matching on extremity and social desirability during 2AFC item composition, and the prioritization of available items for selection based on information. The settings commonly used for TAPAS tests were determined through experimentation and evaluated in simulations to ensure predictable functioning with pools of appropriate size and breadth. Importantly, the algorithm also contains many safeguards aimed at preventing premature termination in unusual scenarios typically involving small pools and tight constraints. The algorithm can automatically relax some constraints by progressively adjusting initial settings, or even substituting a facet combination, to find available items to administer. The exact sequence of decisions and how they interrelate is illustrated in detail in the close-hold TAPAS computer code that was delivered to the ARI.

Figure 5 below shows an example of an item presented to an examinee taking V12 and how it was recorded internally in the output file. This item appeared on the test as the 9th item, so it came from the linking subtest blue-print specification (Achievement-Team Orientation pairing). The first statement in the pair presented was Achievement 125, while the second statement was Team Orientation 38. It can be noted that both statements had similar and positive location parameters (deltas) and similar social desirability parameters; both met pairing constraints for multidimensional 2AFC items described above. The examinee took 4.25 seconds to answer, and the second statement was selected in that pair by the examinee.

```

"Number": 9,
"TimeElapsed": 4.25,
"FirstItem":
  {"Alpha": 1.92,
   "Delta": 2.88,
   "Tau": -3.99,
   "SocialDesirability": 3.57,
   "Name": "Achievement 125",
   "Selected": true},
"SecondItem": {
  "Alpha": 2.31,
  "Delta": 4.39,
  "Tau": -5.49,
  "SocialDesirability": 3.63,
  "Name": "Team Orientation 38",
  "Selected": false}

```

Figure 5. Example of an Item Administered on Version 12

7.3 Mathematical Details for Calculating Item and Test Information Values During CAT Administration

To implement adaptive testing with 2AFC items, TAPAS uses the MUPP IRT model proposed by Stark (2002) and described in Stark et al. (2005). The model assumes that when a respondent is presented with a pair of stimuli (e.g., personality statements), denoted as stimuli s and t , and is asked to indicate a preference, the respondent evaluates each stimulus separately and makes *independent* decisions about stimulus endorsement. If a respondent's endorsement propensity is equal for both stimuli, the individual must reevaluate the stimuli independently until a preference is reached. (Note that this assumption is similar to that of Andrich's (1995) hyperbolic cosine model for unidimensional pairwise preferences.) Thus, the probability of endorsing a stimulus s over a stimulus t can be formally written as

$$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) = \frac{P_{st}\{1,0\}}{P_{st}\{1,0\} + P_{st}\{0,1\}} \approx \frac{P_s\{1\}P_t\{0\}}{P_s\{1\}P_t\{0\} + P_s\{0\}P_t\{1\}}, \quad (1)$$

where:

i = index for items, consisting of pairs of stimuli, where $i = 1$ to I ,

d = index for dimensions (i.e., facets), where $d = 1, \dots, D$,

s, t = indices for the first and second stimuli, respectively, in an item,

$\theta_{d_s}, \theta_{d_t}$ = latent trait values for a respondent on facets d_s and d_t respectively,

$P_s\{1\}, P_s\{0\}$ = probability of endorsing/not endorsing stimulus s at θ_{d_s} ,

$P_t\{1\}, P_t\{0\}$ = probability of endorsing/not endorsing stimulus t at θ_{d_t} ,

$P_{st}\{1,0\}$ = joint probability of endorsing stimulus s , and not endorsing stimulus t at $(\theta_{d_s}, \theta_{d_t})$,

$P_{st}\{0,1\}$ = joint probability of not endorsing stimulus s , and endorsing stimulus t at $(\theta_{d_s}, \theta_{d_t})$, and

$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t})$ = probability of respondent j preferring stimulus s to stimulus t in pairwise preference item i .

A preference is represented by the joint outcome {Agree (1), Disagree (0)} or {Disagree (0), Agree (1)}. An outcome of {1,0} indicates that stimulus s was preferred to stimulus t , and is considered a positive response; an outcome of {0,1} indicates that stimulus t was preferred to s (a negative response). Thus, the response data for this model are dichotomous. The probability of endorsing a stimulus in a pairwise preference item depends on θ_{d_s} and θ_{d_t} as well as the model for single-stimulus responding. TAPAS uses the GGUM (Roberts, Donoghue, & Laughlin, 2000), which was found to fit single-stimulus personality data in early research (Chernyshenko et al., 2007; Stark et al., 2006).

The GGUM assumes an ideal point process underlies single-stimulus responding; i.e., as the distance between a respondent's location on a trait continuum (called the ideal point) and the location of a statement increases, the probability of endorsing that statement *decreases*. This assumption implies a single-peaked, bell-shaped response function. The general form of the GGUM is derived in Roberts et al. (2000). However, for this application, only the special cases for binary, Disagree ($Z = 0$) and Agree ($Z = 1$), responses are needed. The specialized equations are shown below:

$$P[Z_s = 0 | \theta_{d_s}] = \frac{1 + \exp\left(\alpha_s \left[3(\theta_{d_s} - \delta_s)\right]\right)}{\gamma_s}, \text{ and} \quad (2a)$$

$$P[Z_s = 1 | \theta_{d_s}] = \frac{\exp\left(\alpha_s \left[(\theta_{d_s} - \delta_s) - \tau_{s1}\right]\right) + \exp\left(\alpha_s \left[2(\theta_{d_s} - \delta_s) - \tau_{s1}\right]\right)}{\gamma_s}, \quad (2b)$$

where

Z_s = an observable response to stimulus s , where 0 indicates disagreement and 1 indicates agreement,

θ_{d_s} = the location of a respondent (trait score) on the facet represented by stimulus s ,

δ_s = the location parameter of stimulus s on the latent continuum,

α_s = the discrimination parameter for stimulus s ,

τ_{sk} = the threshold parameter indicating the location of the k^{th} subjective response category threshold on the latent continuum, where $k=1$ for binary responses,

γ_s = a normalizing factor that is required to make the observable response probabilities, summed over response options, add to 1. Thus, $\gamma_s = P[Z_s = 0 | \theta_{d_s}] + P[Z_s = 1 | \theta_{d_s}]$, or explicitly:

$$\gamma_s = 1 + \exp\left(\alpha_s \left[3(\theta_{d_s} - \delta_s)\right]\right) + \exp\left(\alpha_s \left[(\theta_{d_s} - \delta_s) - \tau_{s1}\right]\right) + \exp\left(\alpha_s \left[2(\theta_{d_s} - \delta_s) - \tau_{s1}\right]\right). \quad (2c)$$

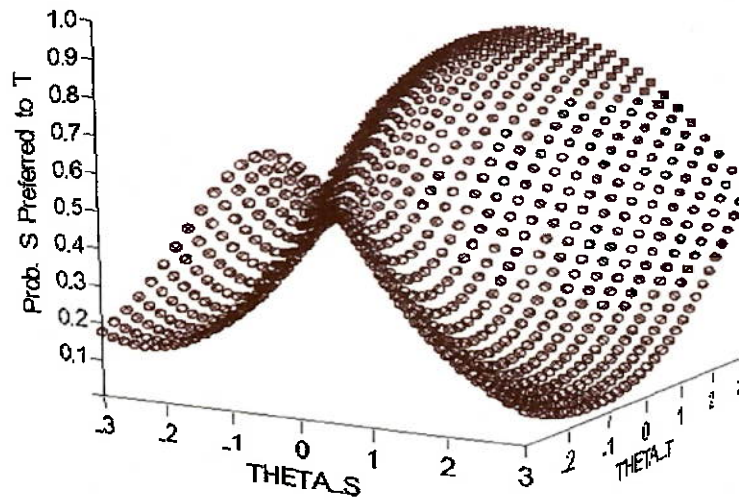
Analogous equations hold for stimulus t .

After the response probabilities for individual statements are computed using Equation 2, they can be substituted into the general equation for the MUPP model:

$$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) = \frac{P\{Z_s = 1 | \theta_{d_s}\} P\{Z_t = 0 | \theta_{d_t}\}}{P\{Z_s = 1 | \theta_{d_s}\} P\{Z_t = 0 | \theta_{d_t}\} + P\{Z_s = 0 | \theta_{d_s}\} P\{Z_t = 1 | \theta_{d_t}\}}. \quad (3)$$

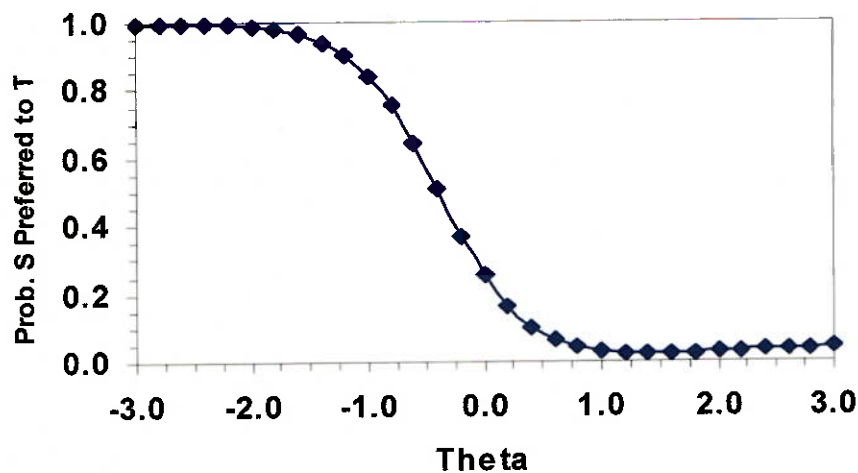
Equation 3 can then be used to compute the MUPP 2AFC item response functions (IRFs). An example IRF for a multidimensional 2AFC item is shown in Figure 6a and an example IRF for a unidimensional 2AFC item is shown in Figure 6b. Note that the probability of preferring a particular stimulus depends on the respective distances of the stimuli from the respondent. Assuming equal discrimination and threshold parameters, a respondent is more likely to choose the proximal stimulus as “more like me” in a pairwise preference task. For details on MUPP IRFs, see Stark et al. (2005) and Stark et al. (2012). The fit of the MUPP model was examined by Drasgow et al. (in press). Using a new method to assess fit, which is described in their report, Drasgow et al. found that the MUPP model provided a fair, but not excellent, description of the data.

IRF Involving Stimuli on Different Dimensions



$$(\alpha_S = 0.8, \delta_S = 1.2, \tau_S = -0.3) (\alpha_T = 2.0, \delta_T = 0.3, \tau_T = 0.1)$$

IRF Involving Stimuli on Same Dimension



$$(\alpha_S = 1.1, \delta_S = -2.3, \tau_S = -3.1) (\alpha_T = 1.4, \delta_T = 1.1, \tau_T = -2.4)$$

Figure 6. MUPP Item Response Functions (IRFs) for an Item Composed of Stimuli Representing Different Facets and an Item Composed of Stimuli Representing the Same Facet

Because an individual 2AFC item, i , can involve at most two facets, a general expression for item information $I_i(\tilde{\theta})$ is:

$$I_i(\theta_{d_s}, \theta_{d_t}) = \frac{[P'_{(s>t)_i}]^2}{[P_{(s>t)_i}][Q_{(s>t)_i}]}, \quad (4)$$

where $P_{(s>t)_i}$ is shorthand notation for $P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t})$, given by Equation 3, and

$$P'_{(s>t)_i} = \begin{cases} \frac{\partial P_{(s>t)_i}}{\partial \theta_{d_s}} \\ \frac{\partial P_{(s>t)_i}}{\partial \theta_{d_t}} \end{cases} \text{ if stimuli } s \text{ and } t \text{ lie on different facets (i.e., } d_s \neq d_t \text{),}$$

or alternately

$$P'_{(s>t)_i} = \left[\frac{\partial P_{(s>t)_i}}{\partial \theta_d} \right] \text{ if stimuli } s \text{ and } t \text{ lie on the same facet (i.e., } d_s = d_t = d \text{),}$$

and the numerator of Equation 4 can be computed by taking the inner product of the vector expressions, as shown:

$$\left[P'_{(s>t)_i} \right]^2 = \left(\frac{\partial P_{(s>t)_i}}{\partial \theta_{d_s}} \right)^2 + \left(\frac{\partial P_{(s>t)_i}}{\partial \theta_{d_t}} \right)^2 \text{ if stimuli } s \text{ and } t \text{ lie on different facets,} \quad (6a)$$

and

$$\left[P'_{(s>t)_i} \right]^2 = \left(\frac{\partial P_{(s>t)_i}}{\partial \theta_d} \right)^2 \text{ if stimuli } s \text{ and } t \text{ lie on the same facet, } d. \quad (6b)$$

To use these general equations for TAPAS, the first partial derivatives for the GGUM-based MUPP formulation are needed. Readers interested in those details may refer to Stark et al. (2005), Joo, Lee, and Stark (2018), and Lee, Joo, Stark, and Chernyshenko (2019).

Figure 7 presents MUPP item information functions (IIFs) corresponding to the IRFs shown in Figure 6 above. In general, items composed of statements having larger discrimination parameters provide more information, and information is highest where the change in IRF slope is greatest. Where IRFs are relatively flat, item information approaches zero.

IIF Involving Stimuli on Different Dimensions

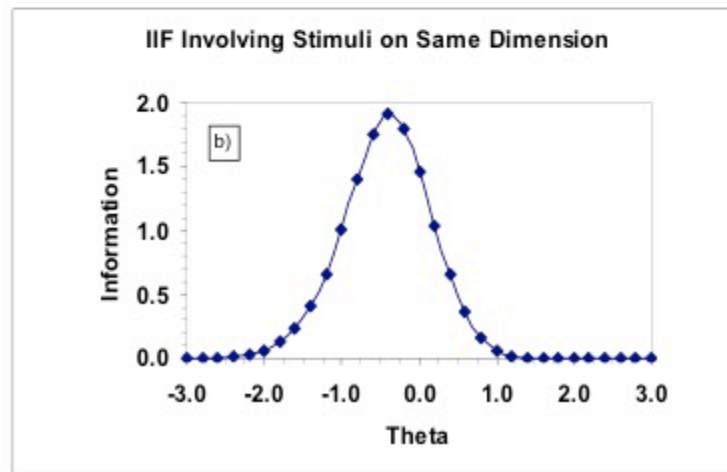
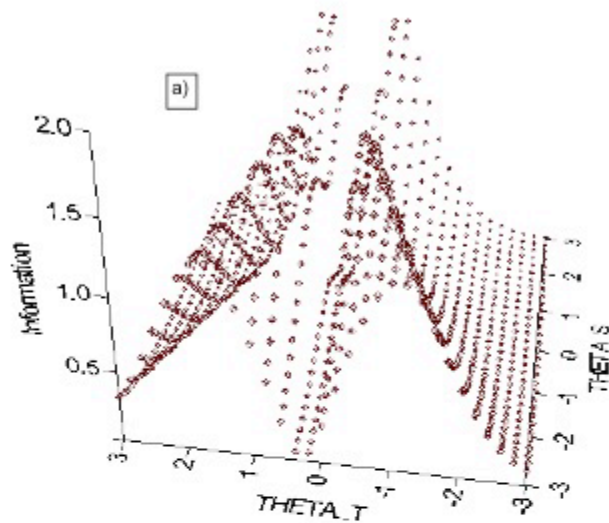


Figure 7. MUPP Item Information Functions (IIFs) for the Items Having IRFs Shown in Figure 6

In summary, designing TAPAS tests involves decisions about dimensionality (number of facets to measure), content (how statements measuring different facets can be combined to form 2AFC items), test length (total number of substantive and response check items), and constraints/initial settings for item construction and selection. These constraints are used to identify 2AFC items that are “available” for administration at each point in a test, Equations 1-4 are used to identify a subset of best items from an information standpoint, and an item is then chosen randomly from that subset for administration. This and other stochastic features of the TAPAS algorithm (e.g., random ordering of item types within linking, main, and additional subtests, random ordering of statement positions within items, small variations in initially assumed trait scores) make it unlikely that examinees will receive duplicate tests.

Although it is true that not administering the most informative item at every measurement opportunity is suboptimal, and even low-level variations in test composition may stretch assumptions of parallel measurement (across examinees and upon retesting), we believe the benefits for test security outweigh the potential disadvantages. In practice, examinee trait scores change considerably for most facets over the course of a test, so items identified as informative at the beginning may contribute little to observed information in the end. This was one of the considerations in early discussions about two-part testing. In addition to exploring new facets beyond the “core” group used to calculate screening composites, two-part testing offers an opportunity to compare the SEs of core facet scores and dynamically (re-) configure part 2 test specifications to improve their measurement precision. In the future, two-part testing capabilities can also be applied to retesting, and potentially with a combination of 2AFC and three-alternative forced choice formats.

8.0 TAPAS SCORING, NORMING, AND REPORTING TEST RESULTS AND DIAGNOSTICS

This section focuses on various test results and diagnostics that are typically included in an output file after completion of each TAPAS test. In particular, we describe in some detail how 1) TAPAS raw scores (i.e., untransformed latent trait estimates) and their SEs are computed, 2) how raw scores are normed and are transformed into corresponding Z-, T-, and percentile scores, 3) how Z-scores are used to produce TAPAS composites, and 4) aberrance indicators that may be used to identify potentially unmotivated examinees. As before, we illustrate each of these discussion points with an example output from one of the most recent Army MEPS testing versions (Version 12).

8.1 TAPAS Trait and Standard Error Estimation

TAPAS scores are based on a GGUM formulation of the MUPP IRT model (Stark, 2002) and they are estimated using Bayes modal estimation (see Drasgow et al., 2012; Stark et al., 2005). Conceptually, the goal of the scoring procedure is to find a set of trait scores that makes a respondent’s pattern of 2AFC item responses most likely, based on “known” (i.e., previously estimated) statement parameters and assumed prior distributions for the measured latent traits. For TAPAS scoring, the latent traits are assumed to have independent standard normal prior

distributions and, as always, the effects of the chosen priors (e.g., regression of estimated scores toward the prior means) diminish as test length increases.

Formally, the likelihood of an examinee's response pattern for an n -item TAPAS test can be written as:

$$L(\tilde{\mathbf{u}}, \tilde{\boldsymbol{\theta}}) = \left\{ \prod_{i=1}^n [P_{(s>t)_i}]^{u_i} [1 - P_{(s>t)_i}]^{1-u_i} \right\} * f(\tilde{\boldsymbol{\theta}}), \quad (5)$$

where $\tilde{\boldsymbol{\theta}} = (\theta_{d'=1}, \theta_{d'=2}, \dots, \theta_{d'=D})$ is a vector of latent trait scores for D -dimensions, $\tilde{\mathbf{u}}$ represents an examinee's response pattern, u_i is the dichotomous response to item i , $P_{(s>t)_i}$ is the probability of preferring statement s to statement t in item i , and $f(\tilde{\boldsymbol{\theta}})$ is a D -dimensional prior density function, which is assumed to be the product of independent standard normal distributions,

$$\prod_{d'=1}^D \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\theta_{d'}^2}{2\sigma^2}\right). \quad (6)$$

Taking the natural log, $\ln$, of the combined equations gives the log posterior,

$$\ln L(\tilde{\mathbf{u}}, \tilde{\boldsymbol{\theta}}) = \sum_{i=1}^n [u_i \ln P_{(s>t)_i} + (1-u_i) \ln (1 - P_{(s>t)_i})] + \sum_{d'=1}^D \left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{\theta_{d'}^2}{2\sigma^2} \right], \quad (7)$$

leaving the following set of equations to be solved:

$$\frac{\partial \ln L}{\partial \tilde{\boldsymbol{\theta}}} = \begin{bmatrix} \frac{\partial \ln L}{\partial \theta_{d'=1}} \\ \frac{\partial \ln L}{\partial \theta_{d'=2}} \\ \dots \\ \frac{\partial \ln L}{\partial \theta_{d'=D}} \end{bmatrix} = 0. \quad (8)$$

These equations are solved numerically to obtain a vector of latent trait scores for each respondent using the DFPMIN subroutine (Press, Flannery, Teukolsky, & Vetterling, 1990) in conjunction with supplied functions for computing the posterior and its first derivatives. DFPMIN performs a D -dimensional maximization (minimizes the negative log likelihood) using

a Broyden-Fletcher-Goldfarb-Shanno algorithm and, in the process, provides an approximation of the inverse Hessian matrix used to estimate SEs. As was mentioned previously, TAPAS trait scores are continuously updated during the test, because they are used by the CAT algorithm to identify potentially informative items to administer.

Simulation research indicated that the SEs based on the inverse Hessian approximation were typically larger than empirical SDs, (Stark & Drasgow, 2002; Stark et al., 2005). Consequently, Stark et al. (2012) developed an alternative replication method that has been used since to estimate TAPAS SEs. Upon completion of a test, 50 response patterns are simulated using an examinee's final trait scores and the statement parameters for administered items. The simulated response patterns are scored, and the SDs of the scores for each facet, across replications, are reported as "replication SEs". As shown by Stark et al. (2012), this replication method provides SE estimates that are much closer to the empirical (true) SDs than previously used approaches (i.e., based on the approximated inverse Hessian matrix or a jack-knife procedure).

8.2 TAPAS Norming and Raw Score Transformation

As discussed in Section 2 of this report, TAPAS testing at the MEPS has undergone several revisions. New pools of statements have been developed to assess promising constructs, original statement pools have been revised and expanded, and several flags have been implemented to detect unmotivated responding. Altogether, more than a dozen TAPAS versions have been administered in MEPS since 2009. Some of these versions have similar test specifications (i.e., they assess 9 to 13 overlapping (common) facets with the same overall dimensionality and length), while other versions differ considerably in content, dimensionality, length, and even the number of parts. Consequently, although IRT raw scores should, in theory, be comparable for examinees taking the same version of a test, the IRT scores for *different* versions have been normed and transformed to facilitate comparisons *across* versions. This is akin to developing concordance tables for relating scores on different achievement tests in academic admissions contexts.

In 1980 and again in 1997, the U.S. DoD, jointly with the U.S. Department of Labor, conducted large-scale norming studies for the ASVAB (Martin, 1998, 1999; U.S. DoD, 1982a, 1982b). Great efforts were made to secure probability samples that could be weighted to represent the American youth population. A representative form of ASVAB was administered to these samples and subsequent forms of ASVAB have been equated to this reference form. A scale score of X on, say, the ASVAB Arithmetic Reasoning subtest might correspond to the 65th percentile of the youth population, regardless of the specific ASVAB form.

TAPAS has not been administered to a representative sample of American youth. Consequently, it is not possible to equate or link new TAPAS forms back to a reference form. Instead, each TAPAS form is normed (by the process described below) when a sufficient sample of current respondents is available.

Facet scale scores (i.e., transformed facet raw scores) are created by the following process when a sufficient number of cases is available. First, a norm table is constructed that specifies the range of raw scores corresponding to each percentile point for that TAPAS form. Then each

examinee's raw score (the latent trait estimate $\hat{\theta}$ computed by the above process) for each facet is converted into a percentile rank score using the norm table specific to that version. The examinees' percentile scores are then converted into Z-scores using an inverse cumulative normal probability table, so that the resulting Z-scores for the norming sample will have a normal distribution with a mean of zero and a variance of 1. The Z-scores are then converted to T-scores using a linear transformation, $T=10Z+50$ (mean=50, SD=10).

As noted in Table 11, forms administered concurrently within a Phase (e.g., V9, V10, and V11) are randomly assigned to examinees. Consequently, assigning equal scale scores to equal percentile ranks across concurrent forms is equipercentile equating. To the extent that there is not drift in trait distributions across Phases, it is also equipercentile equating over Phases.

As shown in Figure 8 below, TAPAS forms report five primary scoring-related results for each facet: *raw score* (Theta), *replication standard error* (SETheta), *percentile score* (Percentile), *Z-score* (ZScore), and *T-score* (TScore). Note that when conducting validation studies with multiple TAPAS versions, the normed Z-scores or T-scores should be used instead of raw scores because the $\hat{\theta}$ metrics can vary across forms (e.g., due to differences in test length).

```
"Name": "Achievement",
  "Theta": 0.4905,
  "SETheta": 0.2699,
  "TScore": 57.1,
  "ZScore": 0.71,
  "Percentile": 76
```

Figure 8. Example of TAPAS Score-Related Output for the Achievement Facet

When a new TAPAS version is fielded, no norming group exists to relate raw scores to percentiles (and thus Z- and T- scores) because it has not been possible to conduct norming studies prior to deployment. Some way of obtaining scale scores for new TAPAS versions is needed for several reasons. First, some facets may not have been administered previously; note that ARI researchers have taken pains to explore a wide range of personality characteristic in an attempt to improve prediction of various aspects of performance. Second, changes in test design (e.g., different numbers of facets, and numbers of statements measuring each facet) can change the distributions of trait estimates, rendering comparisons of raw scores from different forms invalid. Moreover, it is possible that the distributions of scores for previously administered facets may change over time; note however that our analyses have shown little evidence of score drift.

To address the need for scale scores as soon as a new form is fielded the following process has been used. Data from previous TAPAS versions are used when possible. Simulated test administrations, using the item parameters for new facets, are used to generate a large number of

cases. Raw score distributions from previous versions and the simulated respondents are then used to create “provisional norms” that can be used as soon as a form is fielded. When an adequately large number of examinees (e.g., 10,000) has been tested with the new version, “operational norms” are determined. Revised TAPAS software with the new norm table is then submitted for operational use.

For example, Table 21 below shows the first five percentile ranges for the Achievement facet of TAPAS V4 based on a sample of 45,527 applicants for Army enlistment. As shown, the raw $\hat{\theta}$ values from -5 (corresponding to the lower bound) to -.925 (corresponding to the midpoint between -1.0 and -.85) are assigned a value of $Z = -2.33$, values from -.9249 to -.8050 are assigned a value of $Z = -2.05$, and so on. Note that starting with TAPAS V9, these ranges are computed based on the actual percentile values and not adjacent midpoint values. An alternative way to compute Z-scores would have been to use the norm group’s mean and SD for the Achievement facet to convert scores (e.g., .16 and .484 in the table below). Potentially, however, raw trait values that are outliers could result in very large Z-scores, which is undesirable when Z-scores are used to calculate subsequent composite scores. Using percentile conversions effectively limits the minimum Z-score for a facet to -2.33 and the maximum score to +2.33.

Table 21. Example Norm Table for TAPAS Raw Score Conversions

Facet		ACHIEVEMENT			
Valid		45,527			
Missing		0			
Mean		.16			
SD		.484			
Percentiles	Bounds	Min	Max	Z-score	
	-5				
1	-1.00	-5.0000	-0.9250	-2.33	
2	-.85	-0.9249	-0.8050	-2.05	
3	-.76	-0.8049	-0.7250	-1.88	
4	-.69	-0.7249	-0.6650	-1.75	
5	-.64	-0.6649	-0.6150	-1.64	

Detailed information about the norm groups that were used for each TAPAS version is presented below.

V4 - the norm group comprised 45,527 Army examinees who completed TAPAS V4 between May 2009 and May 2010.

V5 - the norm group comprised 21,989 Army examinees who completed TAPAS V5 between August 2011 and July 2012.

V7 - the norm group comprised 43,997 Army examinees who completed TAPAS V7 between August 2011 and July 2012.

V8 - the norm group comprised 44,110 Army examinees who completed TAPAS V8 between August 2011 and July 2012.

V9 - the norm group comprised 3,657 Army examinees who completed TAPAS V9 in September 2013.

V10 - the norm group comprised 3,609 Army examinees who completed TAPAS V10 in September 2013.

V11 - the norm group comprised 3,617 Army examinees who completed TAPAS V11 in September 2013.

V12, V13, V14 - the norm group comprised 54,617 Army examinees who completed Part 1 of V12, V13, and V14 between October 2019 and May 2020. Note that norming was done only for Part 1 of these versions as only Part 1 was to be used for decision making.

Provisional norms in use by the Air Force and Marines are based on Army data. To the best of our knowledge, Service specific norms for the Air Force and Marines have not been developed.

Note that, in the past, program updates to incorporate new norming tables were rare due to the length of the approval process and the difficulties involved with tracking and replacing all TAPAS executable files. However, with TAPAS testing applications now running on the Cloud platform, updates can be deployed much faster.

8.3 TAPAS Composite Scores

As described in a section 6.0, the TAPAS computer program also calculates several composites intended to be used for screening purposes. These composites are weighted combinations of the normalized Z-scores for several TAPAS facets and, thus, are normally distributed. To facilitate decision-making, all composites are scaled to have a mean of 100 and a SD of 20. Hence, a composite value of 100 would correspond to the 50th percentile, while a composite value of 70 would correspond to the 10th percentile.

Figure 9 shows an example of four composite scores currently computed for V12. Calculations are based on Z-scores as described above.

```

"Composites": [{ "Name": "Adaptation",
                  "Value": 101.29},
                { "Name": "Can Do",
                  "Value": 73.68},
                { "Name": "Conduct",
                  "Value": 84.53},
                { "Name": "Will Do",
                  "Value": 104.22}],

```

Figure 9. Example of Composite Scores Produced by TAPAS Version 12

8.4 TAPAS Diagnostic Flags for Unmotivated Responding

Unmotivated examinees are those individuals who do not put effort into responding accurately to an assessment and provide poor quality response data. Sometimes called insufficient effort responding (Bowling et al., 2016), scores for unmotivated examinees are unlikely to predict future performance. Moreover, having too many unmotivated examinees can also result in biased sample-level statistics (e.g., norm group means and SDs, validity coefficients, regression weights, etc.). Therefore, it is recommended that researchers remove unmotivated examinees from the data prior to conducting analyses. To help identify unmotivated responding, four response flags have been progressively implemented for TAPAS testing. These are a) rapid responding flags, b) patterned responding flags, c) an unusual response flag (ℓ_z), and d) a random/inattention flag. An output example of diagnostics from V12 is shown in Figure 10.

"NumberOfItemsCompleted": 176,
"PercentComplete": "100",
"TimeElapsed": 0.0080168,
"Status": "Complete",
"EarlyTerminationReason": "",
"LZFlagTriggered": true,
"MarkovFlagTriggered": true,
"TotalResponseChecks": 6,
"CorrectResponseChecks": 3,
"TimeFlagTriggered": true,
"LzObs": -2.3621499524470813,
"LzCrit01": -0.78695341224321169,
"LzCrit05": -0.57251848506877,
"Markov": 6.93142857142857,

Figure 10. TAPAS Diagnostic Flags

The paragraphs below provide details about these diagnostic flags.

Rapid Responding. Rapid response flags use item response times to identify potentially unmotivated examinees. Historically, two flags have been used to identify rapid responding. The first flag, TimeElapsed, is based on the total testing time, and individuals who complete the TAPAS in less than 688 seconds are flagged as potentially unmotivated. This cutoff was chosen by examining the typical testing time for a 120-item TAPAS test (i.e., V4); the distribution of total testing time was bimodal and 688 appeared to be a reasonable cut score to separate the two modes. The total testing time flag is computed after the test is complete, and it is recommended that researchers exclude anyone with a total testing time of less than 688 seconds.

The second rapid response flag, TimeFlag, focuses on item-level data. In Phase 1 and Phase 2 TAPAS testing (V4, V5, V7, and V8), individuals were flagged for rapid responding if they responded to 21 or more items in less than 2 seconds per item. However, based on subsequent research (Stark et al., 2017), a decision was made to change this standard and include this response flag in the actual TAPAS software. Since making this change, the TimeFlag is triggered

whenever a respondent answers more than 12 items in less than 2 seconds each. All Phase 3 and Phase 4 TAPAS versions use this flag, which is output by the TAPAS software after an examinee finishes the test. Again, individuals who are flagged by the TimeFlag should be removed from the dataset prior to conducting analyses.

Patterned Responding. The second approach to identifying unmotivated responding examines observed response patterns to identify individuals who provide patterned responses (e.g., ABABAB or AAAAAA). Initially, the TAPAS tests incorporated a response flag based on the number of times an individual selected response option A during a 120-item test. The idea behind this flag was that, because the order of two response options is randomized prior to the presentation of each item, selecting option A too many or too few times would indicate patterned responding. A limitation of this rather simple approach is that it could only flag AAAAAA or BBBBBB patterns, but was insensitive to an alternating response pattern (e.g., ABABAB).

To develop an index that would be sensitive to a wider variety of patterned responding, the Markov flag was created and implemented. This flag uses the Markov chain transition matrix for each examinee as shown in Figure 11 below to compare observed vs. expected responses. The values in the cells of the Markov matrix indicate the number of times two particular response options are observed on successive trials. For example, if an examinee completed a test with six items by selecting options ABBAAA, there would be five response patterns: AB, BB, BA, AA, and AA. The Markov values in the 2x2 table would be AA=2, AB=1, BA=1, and BB=1. The Markov matrix shown in Figure 11 illustrates an example of the AA, BB, BA, and AB counts for a 120-item test. Due to the randomized ordering of response options in the TAPAS, the expected counts for each Markov value should be equal to 29.75 or $(\# \text{ items} - 1)/4$. An overall Markov value can be computed as the sum of $(\text{Observed} - \text{Expected})^2 / \text{Expected}$ values across all four cells. The larger the overall Markov value, the higher the likelihood of patterned responding. As can be seen in the figure, the observed counts for the four Markov cells are not too far from the expected value of 29.75. The overall Markov value for this table is 1.44.

The MarkovFlag variable has been implemented in all Phase 3 and Phase 4 TAPAS versions and is triggered whenever the Markov value exceeds 31.06. This cut off value was chosen based on the results of a simulation study as well as analysis of TAPAS V5 data (both had 120 items; for details of this study see Stark et al., 2017). As with the random response flags, individuals who are flagged by the MarkovFlag should be removed from the dataset prior to analyses.

Response to Item i	Response to Item i+1		
		A	B
	A	26	29
	B	29	35

Figure 11. Example of a Markov Chain Transition Matrix

Unusual Responding. The third approach to identifying unmotivated examinees utilizes a well-known IRT fit statistic, ℓ_z , which was originally introduced by Drasgow et al. (1985) as the approximately standardized log likelihood of a response pattern. This statistic is sensitive to unusual response patterns and may be triggered when applicants respond randomly or are trying to fake good.

The log likelihood for a 2AFC test involving D dimensions and I items can be written as

$$\ell_0(\hat{\theta}) = \sum_{i=1}^I u_i \log P_{(s>t)_i} + (1 - u_i) \log(1 - P_{(s>t)_i}), \quad (9)$$

where $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_D)'$ is a vector of latent trait estimates, u_i is a dichotomously scored response to the i^{th} pairwise preference item and equals 1 if statement s is selected, and $P_{(s>t)_i}$ is the probability that a respondent prefers statement s to statement t in item i given the respondent's trait scores $(\hat{\theta}_{d_s}, \hat{\theta}_{d_t})$ on the facets assessed by the item. Accordingly, the expectation and variance of ℓ_0 can be written as

$$E(\ell_0) \approx \sum_{i=1}^I P_{(s>t)_i} \log P_{(s>t)_i} + (1 - P_{(s>t)_i}) \log(1 - P_{(s>t)_i}) \quad (10)$$

and

$$Var(\ell_0) \approx \sum_{i=1}^I P_{(s>t)_i} (1 - P_{(s>t)_i}) \left\{ \log \frac{P_{(s>t)_i}}{(1 - P_{(s>t)_i})} \right\}^2. \quad (11)$$

The multidimensional ℓ_0 can then be standardized to compute the ℓ_z value as shown below.

$$\ell_z = \frac{\ell_0 - E(\ell_0)}{\sqrt{Var(\ell_0)}}. \quad (12)$$

The LzFlag is implemented in all Phase 3 and Phase 4 TAPAS software versions. To the extent that ℓ_z follows a standard normal distribution, values less than -1.64 would trigger the LzFlag. However, its distribution is based on asymptotic theory and our experience suggests that the TAPAS facet assessments are too short for this distribution theory to hold. At this time, more research is needed to better understand the performance of the LzFlag with TAPAS facet scores, and researchers are advised to use their judgment whether to remove or keep those flagged using it.

Random/Inattention Flag. This flag is the most recent flag aimed at detecting unmotivated responding and has been implemented only in the Phase 4 TAPAS versions (V12, V13, and V14). The idea is to periodically present an examinee with a “response check” item with instructions to select a particular response option (e.g., *Select option “A” for this item*). Attentive examinees should have no difficulty following instructions and, thus, should answer all response check items correctly. On the other hand, examinees who are inattentive or responding randomly have a 50% probability of selecting the “wrong” response option. A total of 6 response check items appears in each of the TAPAS versions V12, V13, and V14; response checks are also included in the most recent versions of the AF and USMC TAPAS forms (AF V5.2 and USMC V2.1). After the test is completed, the TAPAS software reports a variable labeled CorrectResponseChecks, which provides a count of the number of times the response check items were answered correctly. At this point, more research is needed to better understand the performance of the CorrectResponseChecks flag, so researchers are advised to use their judgment whether to remove individuals who respond incorrectly to one or more of these items. In our research with civilian samples, we routinely remove respondents who miss two or more attention check items. An example is shown below in Figure 12.

```

"ResponseChecks":
  {
    "Position": 30,
    "TimeElapsed": 0.0009937,
    "Correct": true,
    "FirstItem": {
      "Name": "Random Response Question 30-A",
      "Content": "For data quality check, please select this option.",
      "Selected": true,
      "Correct": true
    },
    "SecondItem": {
      "Name": "Random Response Question 30-B",
      "Content": "Do not select option B.",
      "Selected": false,
      "Correct": false
    }
  }

```

Figure 12. Example of Response Check Item

8.5 Summary of TAPAS Test Results and Guidance on TAPAS Score Interpretation

Care is needed when interpreting TAPAS facet scores. First, note that 13 to 17 facets are assessed in about 20 minutes. This imposes limits on the reliability of the facet scores. Unfortunately, increasing test length to increase reliability may be problematic: There may be a sharp limit on how many ways we can ask about Selflessness or Team Orientation (for example) before the statements become overly redundant. Therefore, we recommend using composites of several facet scores, which will be more reliable, to inform decisions, rather than a single facet score. This is analogous to the use of subtest scores from the ASVAB, where composites are used to determine enlistment eligibility.

Another reason for caution is that TAPAS norms are based on applicant samples, not random or representative samples of American youth. It is known that the distribution of cognitive ability in military applicant samples differs from the distribution in the American youth population (e.g., applicant samples have fewer individuals with very low cognitive ability). Differences between

the two groups in terms of their personalities are not known. Thus, interpretations of Z-scores or T-scores need to be in the context of the applicants who complete the assessment.

The TAPAS raw scores (i.e., latent trait estimates) should also be interpreted carefully. Although software for IRT trait estimation often assumes that traits follow a standard normal distribution, the distributions of estimates from TAPAS often do not have a mean of zero and a SD of one. For example, the mean and SD for Achievement are .16 and .48, respectively. We expect the SDs to be less than one because Bayesian estimation is used. Bayesian estimation requires the use of a prior distribution (e.g., the standard normal distribution is used for TAPAS trait estimates) and trait estimates tend to be pulled toward the mean of this distribution. Therefore, whereas maximum likelihood estimates of traits would have SDs greater than one, the TAPAS's Bayesian estimates invariably have SDs less than one.

In sum, the meaning of Z-scores, after the appropriate norming sample has been collected and norms have been developed, reflects an individual's relative standing with respect to the norm group – applicants to the Army. Here a Z-score of -2.33 is at the 1st percentile, +2.33 is at the 99th percentile, and 0.0 is the 50th percentile.

9.0 RELIABILITY OF TAPAS SCALE SCORES ACROSS TAPAS MEPS VERSIONS

Estimating reliability for an adaptive assessment is difficult. The most common method for estimating reliability, coefficient alpha, cannot be used because different people answer different items. Test-retest reliability is a viable approach, provided that respondents are equally motivated to respond on both occasions. Moreover, for a personality assessment, the respondents need to be motivated in the same way on both occasions: They cannot be answering honestly on one occasion and faking good on the other. Moreover, if the adaptive algorithm produces forms that are not parallel in the sense of CTT, the test-retest correlation will be lower than the true reliability of either form. Thus, the test-retest correlations provided in a subsequent section of this report should be viewed as lower bounds to the facets' reliabilities.

Turning to more theoretical approaches, there are multiple ways to estimate reliability of a multidimensional adaptive assessment. For example, a simulation study can use the item parameters in the item bank to simulate responses and then the true trait values can be correlated with the estimated trait values. This approach assumes that item parameters are known and examinees are responding according to the psychometric model used in the simulation, both of which are unlikely to be true. Thus, the simulation approach in all likelihood overestimates reliability.

In the results described below, we use what is known as IRT marginal reliability: one minus the ratio of the average squared SE of $\hat{\theta}$ to the observed variance of $\hat{\theta}$. This mimics the CTT formula for reliability ρ ,

$$\rho = 1 - \frac{\sigma_E^2}{\sigma_X^2},$$

where σ_E^2 is the error variance and σ_X^2 is the observed score variance. The challenge for marginal reliability is to accurately estimate the SE of $\hat{\theta}$. As noted in Section 8.1, we found that using the inverse Hessian approximation produced SEs that were too large. Consequently, we use the replication method developed by Stark et al. (2012) that is described in Section 8.1. It should be emphasized that further research is needed on estimating the SE of $\hat{\theta}$ values.

In summary, the reliability of TAPAS facet scores obtained from MEPS administrations were estimated to provide information about the precision of scores. IRT reliability was computed for each facet using the examinees' trait scores and SEs as follows (Raju & Oshima, 2005):

1. Square the SE of each examinee's trait score.
2. Average the squared SEs to get the "error variance."
3. Compute the variance of the examinees' trait scores; this is the "observed score variance."
4. Apply the CTT definition of reliability shown above.

Table 22 shows the reliability estimates for various TAPAS versions based on this approach.

Table 22. IRT Reliability Estimates for TAPAS MEPS Versions

TAPAS Facet	V5	V7	V8	V9	V10	V11	V5,7,8 Average	V9,10,11 Average	Change
Achievement	.56	.55	.54	.62	.61	.64	.55	.63	.08
Adjustment	.60	.57	.51				.56		
Adventure Seeking		.72					.72		
Attention Seeking	.66	.66	.64			.67	.65	.67	.02
Commitment to Serve		.48		.65		.71	.48	.68	.20
Cooperation	.65	.65		.71			.65	.71	.06
Courage			.47		.56		.47	.56	.09
Dominance	.68	.66	.64	.76	.71	.73	.66	.73	.07
Even Tempered	.54	.57	.60	.66	.68	.70	.57	.68	.11
Intellectual Efficiency	.62	.60	.61	.69	.67	.67	.61	.68	.07
Non-Delinquency	.57	.60	.59		.64		.59	.64	.05
Optimism	.47	.44	.45	.47	.47	.51	.45	.48	.03
Order	.67	.65		.69	.68	.70	.66	.69	.03
Physical Conditioning	.69	.70	.70	.76	.75	.75	.70	.75	.06
Responsibility			.59	.67			.59	.67	.08
Self-Control	.49		.49				.49		
Selflessness	.64	.62		.68	.70	.69	.63	.69	.06
Situational Awareness		.44			.47		.44	.47	.03
Sociability	.73		.70	.79	.80	.77	.72	.79	.07
Team Orientation			.60			.67	.60	.67	.07
Tolerance	.62		.60	.69	.70	.69	.61	.69	.08

It is important to note that 13 to 15 scale scores are produced by TAPAS in a median response time of just over 20 minutes. If testing time for TAPAS was increased to, say, 3 minutes per scale score, many more items could be administered, and we would expect the reliability estimates to substantially increase. This is evident in the higher reliabilities of Versions 9, 10, and 11, which measured 13 facets using 120 items, as compared to Versions 5, 7, and 8 that measured 15 facets using 120 items.

Note also that TAPAS marginal reliabilities for Versions 9-11 are comparable to or slightly lower than coefficient alpha reliabilities of other Army personality measures using single-statements (Likert-type) response formats. For example, the reliabilities of the Rational Biodata Inventory (RBI; Kilcullen et al. 2005), another instrument of substantial interest to the Army, has reliabilities that range from .42 to .76, with an average of .67. It is possible, however, that those alphas are substantially inflated by single-subject response-consistency bias and other types of correlated error. For example, a three item Extraversion scale with the items “I like to go to parties,” “I like to go out with friends,” and “I often go to parties with friends” would likely have a coefficient alpha exceeding .9. However, it is questionable whether 90% of the variance of

observed scores can be attributable to true variance in Extraversion or how well such a narrow scale would measure a latent construct of Extraversion. An example of one of the most reliable forced-choice instruments is the AIM (White & Young, 1998). It provides four statements and respondents must select the statement that is “most like me” and the statement that is “least like me”. It has reliabilities that range from .70 to .78. It is important to note that the AIM went through a very rigorous development process.

Tables 23 through 30 below give conditional SEs, based on the replication method, as a function of the latent trait for various TAPAS forms. Here $\hat{\theta}$ values were first sorted from low to high for each facet. Then the average of SE estimates was computed for trait estimates in the bottom five percent, for the 6th through the 10th percentile, etc. For many static tests, a plot of the conditional SEs is U-shaped, with less precision for low and high trait values and more precision in the center of the distribution. From the tables below, it appears that TAPAS’s adaptive algorithm is functioning well (and the item pools have sufficient items for low, intermediate, and high trait levels): plots of the conditional SEs would be quite flat across trait levels.

Table 23. Marginal Reliabilities and Conditional Standard Errors for Version 4

Facet Name	Marginal Reliability	Average SE	Average Standard Errors at Each Percentile Range																			
			1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Achievement	.58	.31	.31	.30	.30	.30	.30	.30	.30	.30	.30	.30	.31	.31	.31	.31	.31	.32	.32	.32	.33	.33
Adjustment	.56	.37	.36	.36	.36	.37	.36	.37	.37	.37	.37	.37	.37	.37	.37	.38	.38	.38	.38	.38	.40	.42
Attention Seeking	.67	.30	.31	.32	.31	.30	.30	.29	.29	.29	.28	.28	.28	.28	.29	.29	.29	.30	.31	.32	.33	.35
Cooperation	.38	.29	.30	.29	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.29	.29	.30	.30	.31	.33
Dominance	.76	.29	.31	.30	.29	.29	.28	.28	.27	.27	.27	.27	.27	.27	.28	.28	.28	.29	.29	.30	.31	.30
Even Tempered	.59	.30	.32	.30	.30	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.30	.30	.31	.31	.31	.32
Intellectual Efficiency	.68	.33	.34	.33	.33	.33	.33	.33	.32	.32	.32	.32	.32	.32	.32	.33	.33	.33	.33	.34	.34	.34
Non-Delinquency	.52	.31	.32	.32	.32	.31	.31	.31	.31	.31	.31	.31	.31	.31	.31	.31	.31	.31	.32	.32	.32	.33
Optimism	.50	.32	.32	.32	.31	.31	.31	.31	.31	.31	.31	.32	.32	.32	.32	.32	.33	.33	.33	.34	.34	.35
Order	.69	.30	.31	.33	.33	.32	.32	.31	.31	.31	.30	.30	.29	.29	.28	.28	.28	.28	.27	.28	.28	.30
Physical Conditioning	.78	.29	.32	.31	.30	.30	.29	.28	.28	.28	.27	.27	.27	.28	.28	.28	.29	.29	.30	.31	.32	.30
Self-Control	.47	.38	.38	.38	.39	.38	.38	.38	.38	.38	.38	.38	.38	.38	.38	.38	.38	.38	.38	.38	.38	.37
Selflessness	.51	.30	.31	.31	.30	.30	.30	.30	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.30	.30	.32
Sociability	.64	.35	.37	.37	.37	.36	.36	.35	.35	.35	.35	.35	.35	.35	.35	.34	.34	.34	.34	.34	.35	.35
Tolerance	.59	.36	.37	.37	.37	.36	.36	.36	.36	.36	.36	.36	.36	.36	.36	.36	.36	.36	.36	.36	.36	.36

Note: N = 201,224.

Table 24. Marginal Reliabilities and Conditional Standard Errors for Version 5

Facet Name	Marginal Reliability	Average SE	Average Standard Errors at Each Percentile Range																			
			1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Achievement	.58	.32	.33	.32	.32	.32	.31	.32	.31	.32	.31	.31	.32	.32	.32	.32	.32	.33	.33	.33	.33	.33
Adjustment	.62	.24	.27	.25	.24	.23	.23	.23	.23	.22	.22	.23	.23	.23	.23	.23	.23	.24	.25	.26	.27	.29
Attention Seeking	.66	.34	.33	.34	.35	.34	.34	.34	.34	.34	.34	.33	.33	.33	.34	.34	.34	.34	.34	.35	.35	.35
Cooperation	.65	.31	.31	.30	.29	.29	.29	.29	.29	.30	.30	.30	.30	.30	.30	.31	.31	.31	.31	.32	.34	.37
Dominance	.69	.28	.32	.29	.28	.28	.27	.27	.26	.26	.26	.26	.26	.27	.27	.28	.28	.29	.29	.30	.30	.28
Even Tempered	.53	.32	.34	.33	.32	.32	.32	.31	.31	.31	.31	.31	.32	.32	.31	.32	.33	.32	.33	.33	.34	.35
Intellectual Efficiency	.64	.32	.33	.33	.32	.31	.32	.31	.31	.31	.31	.31	.31	.31	.31	.32	.31	.32	.32	.33	.33	.33
Non-Delinquency	.56	.34	.35	.35	.35	.35	.34	.34	.34	.34	.34	.33	.33	.34	.33	.34	.34	.34	.34	.34	.34	.33
Optimism	.44	.33	.33	.33	.32	.33	.32	.32	.32	.32	.32	.32	.32	.32	.33	.33	.32	.32	.33	.33	.34	.34
Order	.65	.32	.33	.32	.33	.33	.34	.33	.33	.32	.32	.32	.31	.31	.30	.30	.30	.30	.30	.30	.30	.32
Physical Conditioning	.72	.30	.34	.32	.32	.30	.30	.30	.28	.28	.28	.28	.28	.28	.28	.28	.29	.30	.30	.31	.32	.30
Self-Control	.48	.34	.33	.34	.35	.35	.34	.34	.34	.34	.34	.34	.34	.34	.34	.34	.34	.33	.33	.34	.34	.35
Selflessness	.65	.25	.30	.28	.26	.26	.26	.25	.25	.25	.24	.24	.23	.24	.24	.23	.24	.24	.25	.25	.26	.28
Sociability	.74	.29	.30	.32	.31	.30	.30	.30	.29	.29	.28	.28	.28	.28	.27	.27	.27	.27	.27	.28	.29	.30
Tolerance	.64	.31	.34	.33	.33	.32	.32	.32	.31	.31	.30	.30	.30	.32	.30	.30	.30	.30	.31	.31	.32	.32

Note: N = 5,967

Table 25. Marginal Reliabilities and Conditional Standard Errors for Version 7

Facet Name	Marginal Reliability	Average SE	Average Standard Errors at Each Percentile Range																			
			1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Achievement	.55	.32	.33	.32	.32	.31	.31	.31	.31	.31	.32	.31	.31	.31	.32	.32	.32	.32	.32	.33	.33	.33
Adjustment	.55	.25	.27	.25	.25	.24	.24	.24	.24	.23	.23	.24	.23	.24	.25	.25	.25	.26	.27	.28	.29	.30
Attention Seeking	.66	.34	.33	.34	.35	.35	.34	.34	.34	.34	.34	.34	.33	.33	.34	.33	.33	.34	.34	.35	.35	.35
Cooperation	.64	.31	.31	.30	.29	.29	.29	.29	.29	.30	.30	.30	.31	.30	.30	.31	.31	.31	.31	.32	.34	.38
Dominance	.68	.28	.32	.30	.29	.28	.28	.27	.26	.26	.26	.26	.26	.26	.27	.27	.28	.28	.29	.29	.30	.29
Even Tempered	.55	.32	.33	.32	.31	.31	.30	.30	.30	.30	.31	.30	.30	.31	.31	.31	.31	.32	.33	.32	.33	.35
Intellectual Efficiency	.63	.32	.34	.33	.32	.32	.32	.31	.31	.31	.31	.31	.31	.31	.31	.31	.31	.32	.32	.32	.32	.32
Non-Delinquency	.60	.33	.35	.34	.34	.33	.33	.33	.32	.32	.32	.32	.33	.32	.33	.32	.33	.33	.33	.33	.33	.33
Optimism	.43	.32	.34	.33	.33	.32	.32	.32	.32	.32	.31	.32	.31	.32	.32	.32	.32	.32	.33	.33	.33	.34
Order	.66	.32	.33	.33	.33	.34	.33	.33	.33	.32	.32	.31	.31	.31	.30	.30	.30	.30	.29	.30	.31	.32
Physical Conditioning	.72	.29	.33	.32	.32	.31	.30	.29	.28	.28	.28	.27	.27	.28	.28	.28	.28	.28	.29	.30	.31	.31
Selflessness	.62	.26	.31	.30	.28	.27	.27	.26	.26	.26	.25	.25	.25	.25	.25	.25	.25	.25	.26	.26	.27	.29
Adventure Seeking	.73	.31	.31	.32	.32	.32	.31	.31	.31	.30	.30	.29	.30	.30	.29	.29	.29	.30	.30	.31	.32	.33
Commitment to Serve	.50	.36	.36	.37	.36	.36	.36	.37	.36	.36	.36	.37	.36	.36	.37	.37	.37	.37	.36	.36	.35	.30
Situational Awareness	.43	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37

Note: N = 11,836.

Table 26. Marginal Reliabilities and Conditional Standard Errors for Version 8

Facet Name	Marginal Reliability	Average SE	Average Standard Errors at Each Percentile Range																			
			1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Achievement	.55	.32	.33	.32	.32	.32	.32	.32	.32	.31	.32	.32	.32	.32	.32	.32	.32	.32	.33	.33	.33	.34
Adjustment	.54	.25	.29	.26	.26	.25	.24	.24	.24	.23	.24	.24	.24	.24	.24	.24	.25	.25	.26	.26	.28	.30
Attention Seeking	.65	.34	.33	.34	.34	.34	.35	.34	.34	.33	.33	.34	.33	.33	.34	.34	.34	.34	.34	.35	.35	.35
Dominance	.66	.28	.32	.30	.29	.28	.27	.26	.26	.26	.26	.26	.26	.26	.26	.27	.27	.28	.29	.30	.30	.29
Even Tempered	.57	.32	.34	.33	.32	.31	.31	.31	.31	.31	.31	.31	.31	.31	.31	.32	.32	.32	.33	.33	.34	.36
Intellectual Efficiency	.64	.32	.34	.33	.32	.32	.32	.31	.31	.31	.31	.31	.31	.31	.31	.31	.31	.31	.32	.32	.33	.33
Non-Delinquency	.59	.33	.35	.34	.34	.34	.34	.34	.34	.33	.33	.33	.33	.33	.33	.33	.33	.33	.33	.33	.33	.33
Optimism	.42	.32	.34	.33	.32	.33	.32	.32	.32	.32	.31	.32	.32	.32	.32	.32	.32	.32	.33	.33	.33	.34
Physical Conditioning	.71	.30	.34	.33	.32	.31	.31	.30	.29	.29	.28	.28	.28	.28	.28	.28	.29	.28	.29	.30	.31	.31
Self-Control	.48	.32	.33	.33	.32	.33	.33	.32	.33	.32	.33	.32	.32	.32	.32	.32	.32	.32	.32	.32	.33	.34
Sociability	.72	.29	.30	.32	.32	.31	.30	.30	.29	.29	.29	.28	.28	.28	.28	.27	.27	.27	.28	.28	.29	.30
Tolerance	.61	.31	.33	.34	.33	.32	.32	.32	.31	.31	.31	.31	.31	.31	.31	.30	.31	.31	.31	.31	.32	.33
Courage	.49	.39	.39	.40	.40	.40	.40	.41	.40	.41	.41	.40	.40	.40	.40	.40	.40	.39	.38	.37	.33	.29
Responsibility	.58	.29	.27	.26	.27	.27	.27	.28	.28	.28	.28	.29	.29	.30	.30	.30	.30	.30	.31	.31	.33	.39
Team Orientation	.61	.30	.32	.32	.31	.30	.30	.29	.29	.29	.29	.29	.29	.28	.28	.29	.29	.29	.29	.30	.30	.32

Note: N = 11,741.

Table 27. Marginal Reliabilities and Conditional Standard Errors for Version 9

Facet Name	Marginal Reliability	Average SE	Average Standard Errors at Each Percentile Range																			
			1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Achievement	.62	.30	.31	.30	.30	.30	.29	.30	.30	.30	.30	.30	.30	.30	.30	.31	.31	.31	.31	.32	.32	.32
Commitment to Serve	.66	.32	.29	.33	.33	.33	.32	.32	.31	.31	.31	.31	.32	.32	.32	.33	.33	.33	.33	.33	.32	.26
Cooperation	.70	.30	.30	.28	.28	.28	.28	.28	.29	.29	.29	.30	.30	.30	.30	.31	.31	.31	.32	.33	.34	.37
Dominance	.73	.26	.31	.28	.27	.26	.25	.24	.24	.24	.24	.24	.24	.25	.25	.26	.26	.27	.28	.29	.29	.29
Even Tempered	.64	.30	.31	.30	.29	.29	.29	.29	.29	.29	.29	.30	.30	.30	.30	.31	.31	.32	.32	.32	.33	.35
Intellectual Efficiency	.67	.30	.32	.31	.31	.30	.30	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.30	.30	.31	.32
Optimism	.47	.30	.31	.30	.30	.29	.30	.29	.30	.30	.30	.30	.30	.30	.30	.31	.31	.31	.31	.32	.32	.33
Order	.69	.30	.34	.32	.32	.32	.32	.31	.31	.30	.30	.29	.29	.28	.28	.28	.28	.28	.28	.28	.29	.30
Physical Conditioning	.74	.28	.31	.30	.28	.27	.27	.26	.26	.26	.26	.26	.26	.26	.26	.27	.27	.28	.28	.29	.30	.30
Responsibility	.65	.29	.26	.26	.26	.26	.27	.27	.28	.28	.28	.29	.29	.29	.30	.30	.30	.30	.31	.32	.33	.38
Selflessness	.66	.24	.29	.27	.26	.25	.24	.24	.23	.23	.23	.23	.22	.22	.22	.23	.23	.23	.24	.24	.25	.27
Sociability	.78	.27	.29	.31	.31	.30	.29	.28	.27	.27	.26	.26	.26	.26	.25	.25	.25	.25	.26	.26	.27	.29
Tolerance	.67	.29	.33	.32	.31	.30	.29	.29	.29	.28	.28	.28	.28	.28	.28	.28	.28	.28	.29	.29	.30	.31

Note: N = 53,579.

Table 28. Marginal Reliabilities and Conditional Standard Errors for Version 10

Facet Name	Marginal Reliability	Average SE	Average Standard Errors at Each Percentile Range																			
			1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Achievement	.60	.30	.30	.29	.29	.29	.29	.29	.29	.29	.29	.30	.30	.30	.30	.30	.31	.31	.31	.32	.32	.32
Courage	.56	.34	.33	.34	.34	.35	.35	.36	.37	.37	.37	.37	.37	.37	.37	.36	.36	.34	.30	.28	.29	.29
Dominance	.72	.26	.31	.28	.27	.26	.25	.24	.24	.24	.24	.24	.25	.25	.25	.26	.27	.27	.28	.29	.29	.29
Even Tempered	.67	.31	.31	.29	.29	.28	.28	.29	.29	.29	.30	.30	.30	.31	.31	.31	.32	.32	.32	.33	.33	.36
Intellectual Efficiency	.66	.29	.32	.31	.30	.29	.29	.29	.29	.28	.28	.28	.28	.28	.29	.29	.29	.29	.30	.30	.31	.32
Non-Delinquency	.62	.32	.34	.33	.33	.32	.32	.32	.32	.32	.32	.32	.32	.32	.32	.32	.32	.32	.32	.32	.32	.32
Optimism	.47	.30	.31	.30	.29	.29	.29	.29	.29	.30	.30	.30	.30	.30	.30	.30	.31	.31	.31	.32	.32	.33
Order	.69	.30	.33	.32	.32	.32	.32	.31	.30	.30	.30	.29	.29	.28	.28	.28	.28	.28	.28	.28	.29	.30
Physical Conditioning	.75	.28	.31	.29	.28	.27	.27	.26	.26	.26	.26	.25	.26	.26	.26	.27	.27	.28	.29	.30	.31	.30
Selflessness	.66	.24	.30	.27	.26	.25	.24	.24	.23	.23	.23	.23	.22	.22	.22	.23	.23	.23	.24	.25	.26	.27
Sociability	.78	.27	.29	.30	.30	.29	.29	.28	.27	.27	.26	.26	.25	.25	.25	.25	.25	.25	.26	.26	.27	.29
Situational Awareness	.49	.35	.34	.35	.35	.34	.34	.35	.35	.35	.35	.35	.35	.35	.35	.35	.35	.36	.36	.36	.36	.36
Tolerance	.67	.29	.32	.32	.31	.30	.29	.29	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.29	.29	.30	.31

Note: N = 53,285.

Table 29. Marginal Reliabilities and Conditional Standard Errors for Version 11

Facet Name	Marginal Reliability	Average SE	Average Standard Errors at Each Percentile Range																			
			1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Achievement	.62	.30	.30	.30	.29	.29	.29	.29	.29	.29	.30	.30	.30	.30	.30	.30	.31	.31	.31	.32	.32	.33
Attention Seeking	.63	.34	.33	.34	.34	.34	.34	.34	.34	.34	.33	.33	.33	.33	.33	.33	.33	.33	.33	.34	.35	.35
Commitment to Serve	.72	.33	.31	.34	.34	.34	.34	.33	.33	.33	.33	.33	.33	.34	.34	.35	.34	.34	.34	.30	.25	.26
Dominance	.71	.26	.31	.28	.27	.26	.25	.25	.24	.24	.24	.24	.24	.25	.25	.26	.26	.27	.28	.29	.29	.29
Even Tempered	.68	.31	.32	.29	.29	.28	.29	.29	.29	.29	.30	.30	.30	.30	.31	.31	.32	.32	.32	.33	.33	.36
Intellectual Efficiency	.65	.30	.32	.32	.31	.30	.30	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.30	.30	.31	.32
Optimism	.52	.30	.31	.30	.29	.29	.29	.29	.30	.29	.30	.30	.30	.30	.30	.30	.31	.31	.31	.32	.32	.33
Order	.69	.30	.34	.32	.32	.32	.32	.31	.31	.30	.29	.29	.29	.28	.28	.28	.28	.28	.28	.28	.29	.30
Physical Conditioning	.75	.28	.31	.29	.28	.27	.27	.26	.26	.26	.26	.26	.26	.26	.26	.27	.27	.28	.29	.29	.30	.30
Selflessness	.67	.24	.29	.27	.26	.25	.24	.24	.23	.23	.23	.22	.22	.22	.23	.23	.23	.23	.24	.25	.26	.28
Sociability	.76	.27	.29	.31	.30	.29	.29	.28	.27	.27	.26	.26	.26	.25	.25	.25	.25	.25	.26	.26	.27	.29
Team Orientation	.64	.29	.30	.29	.28	.28	.28	.27	.27	.27	.27	.28	.28	.28	.28	.29	.29	.30	.30	.31	.31	.32
Tolerance	.67	.29	.33	.32	.31	.30	.29	.29	.28	.28	.28	.28	.28	.28	.28	.28	.28	.29	.29	.29	.30	.31

Note: N = 54,003.

Table 30. Marginal Reliabilities and Conditional Standard Errors for Versions 12, 13, and 14 (Part 1, 133 Items)

Facet Name	Marginal Reliability	Average SE	Average Standard Errors at Each Percentile Range																			
			1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Achievement	.59	.30	.31	.30	.30	.30	.29	.29	.29	.30	.30	.30	.30	.30	.30	.30	.30	.31	.31	.31	.32	.33
Adjustment	.64	.25	.27	.25	.24	.23	.23	.23	.23	.23	.23	.23	.23	.24	.24	.24	.25	.25	.26	.27	.29	.31
Attention Seeking	.61	.33	.33	.34	.34	.34	.34	.33	.33	.33	.33	.33	.33	.33	.32	.33	.33	.33	.33	.33	.34	.35
Dominance	.68	.27	.33	.29	.29	.28	.27	.26	.26	.26	.25	.25	.25	.25	.26	.26	.26	.27	.28	.29	.29	.30
Even Tempered	.63	.30	.32	.30	.29	.29	.29	.29	.29	.29	.29	.29	.30	.30	.30	.31	.31	.31	.32	.32	.33	.36
Intellectual Efficiency	.65	.30	.33	.32	.31	.30	.30	.30	.29	.29	.29	.29	.29	.29	.29	.29	.30	.30	.30	.31	.32	.32
Non-Delinquency	.64	.31	.33	.32	.31	.31	.31	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.31	.31	.31
Optimism	.50	.31	.32	.31	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.31	.31	.31	.31	.31	.32	.33	.33
Order	.67	.30	.34	.32	.32	.32	.32	.31	.30	.30	.30	.29	.29	.29	.28	.28	.28	.28	.29	.29	.30	.31
Physical Conditioning	.73	.28	.33	.32	.31	.30	.29	.28	.28	.27	.27	.26	.26	.26	.26	.26	.27	.27	.28	.29	.30	.30
Sociability	.73	.28	.30	.31	.31	.31	.30	.30	.29	.29	.28	.28	.27	.27	.27	.26	.26	.26	.26	.27	.27	.29
Team Orientation	.61	.28	.30	.29	.29	.28	.28	.27	.27	.27	.27	.27	.27	.27	.27	.27	.27	.27	.27	.28	.28	.30
Tolerance	.64	.30	.33	.32	.31	.30	.30	.29	.29	.29	.29	.28	.28	.28	.28	.29	.29	.29	.29	.30	.31	.32

Note: N = 79,752.

TAPAS Versions 12, 13, and 14 added a second section that included some experimental facets. It added 43 items to the 133 items on the common, Part 1 section, so overall these versions consist of 176 items. Marginal reliabilities and conditional SEs are shown below in Tables 31 to 33 for TAPAS versions 12, 13, and 14. Note that the marginal reliabilities of the facets common to Parts 1 and 2 are slightly higher and their conditional SEs are slightly lower.

Table 31. Marginal Reliabilities and Conditional Standard Errors for Version 12 (Parts 1 and 2, 176 Items)

Facet Name	Marginal Reliability	Average SE	Average Standard Errors at Each Percentile Range																			
			1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Achievement	.59	.29	.30	.29	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.29	.29	.29	.30	.30	.31	.32
Adjustment	.66	.23	.26	.23	.22	.22	.21	.21	.21	.21	.22	.22	.22	.22	.23	.23	.24	.24	.25	.26	.28	.30
Attention Seeking	.62	.32	.33	.33	.33	.33	.32	.32	.32	.32	.32	.31	.31	.31	.31	.31	.31	.31	.31	.32	.33	.34
Commitment ¹	.71	.31	.30	.32	.32	.31	.30	.30	.30	.29	.29	.29	.30	.30	.30	.31	.31	.32	.32	.32	.32	.31
Dominance	.69	.26	.32	.28	.27	.26	.26	.25	.24	.24	.24	.24	.24	.24	.24	.25	.25	.26	.26	.27	.28	.29
Even Tempered	.65	.29	.31	.29	.28	.27	.27	.27	.27	.28	.28	.28	.28	.29	.29	.30	.30	.31	.31	.32	.33	.36
Humility ¹	.42	.42	.41	.41	.41	.42	.42	.42	.42	.42	.42	.42	.43	.42	.42	.42	.42	.41	.41	.40	.40	.42
Intellectual Efficiency	.67	.29	.32	.31	.30	.29	.29	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.29	.29	.30	.31	.32
Non-Delinquency	.66	.29	.32	.30	.30	.29	.29	.29	.29	.29	.29	.28	.28	.29	.29	.29	.29	.29	.29	.30	.30	.31
Optimism	.53	.30	.31	.29	.29	.28	.28	.28	.28	.29	.29	.29	.29	.29	.29	.29	.30	.30	.31	.31	.32	.33
Order	.69	.29	.34	.32	.32	.31	.31	.30	.29	.29	.28	.28	.28	.27	.27	.27	.27	.27	.27	.28	.28	.30
Persistence ¹	.53	.37	.37	.37	.37	.36	.36	.37	.37	.36	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37
Physical Conditioning	.76	.27	.32	.31	.29	.28	.27	.27	.26	.25	.25	.25	.25	.25	.25	.25	.25	.26	.27	.28	.29	.30
Self-Efficacy ¹	.46	.38	.37	.37	.38	.38	.38	.38	.38	.38	.38	.38	.38	.39	.39	.39	.38	.39	.39	.38	.39	.40
Sociability	.76	.27	.29	.31	.30	.30	.29	.28	.28	.27	.26	.26	.26	.25	.25	.25	.25	.25	.25	.25	.26	.28
Team Orientation	.65	.26	.30	.28	.27	.27	.26	.26	.26	.25	.25	.25	.25	.25	.25	.25	.25	.25	.26	.26	.27	.29
Tolerance	.66	.28	.32	.30	.29	.28	.28	.27	.27	.27	.27	.27	.27	.27	.27	.27	.27	.28	.28	.29	.30	.31

Note: N = 33,597. ¹Experimental facet included in Section 2.

Table 32. Marginal Reliabilities and Conditional Standard Errors for Version 13 (Parts 1 and 2, 176 Items)

Facet Name	Marginal Reliability	Average SE	Average Standard Errors at Each Percentile Range																			
			1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Achievement	.57	.29	.30	.29	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.29	.29	.29	.30	.30	.31	.32
Adjustment	.65	.23	.26	.23	.22	.22	.21	.21	.21	.21	.22	.22	.22	.22	.23	.23	.24	.24	.25	.26	.28	.31
Army Self-Efficacy ¹	.60	.33	.32	.32	.32	.32	.32	.32	.32	.32	.32	.33	.33	.33	.33	.33	.33	.33	.34	.34	.34	.33
Attention Seeking	.63	.32	.33	.33	.33	.33	.32	.32	.32	.32	.32	.31	.31	.31	.31	.31	.31	.31	.32	.32	.33	.34
Commitment ¹	.68	.31	.30	.32	.32	.31	.31	.30	.30	.30	.30	.30	.30	.30	.30	.31	.31	.31	.32	.32	.32	.31
Dominance	.69	.26	.32	.28	.27	.26	.26	.25	.25	.24	.24	.24	.24	.24	.24	.25	.25	.26	.27	.28	.28	.29
Even Tempered	.63	.29	.31	.28	.28	.27	.27	.27	.28	.28	.28	.28	.29	.29	.29	.30	.30	.31	.31	.32	.33	.36
Intellectual Efficiency	.66	.29	.32	.31	.30	.29	.29	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.29	.29	.30	.31	.32
Non-Delinquency	.64	.29	.32	.31	.30	.30	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.30	.30	.31
Optimism	.50	.30	.31	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.29	.30	.30	.30	.31	.31	.32	.34
Order	.68	.29	.34	.32	.32	.31	.31	.30	.29	.29	.28	.28	.28	.27	.27	.27	.27	.27	.28	.28	.29	.30
Physical Conditioning	.75	.27	.32	.30	.29	.28	.27	.27	.26	.26	.25	.25	.25	.25	.25	.25	.26	.26	.27	.28	.30	.30
Persistence ¹	.50	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37	.37
Sociability	.74	.27	.29	.31	.31	.30	.29	.28	.28	.27	.27	.26	.26	.25	.25	.25	.25	.25	.25	.25	.26	.28
Team Orientation	.63	.26	.30	.28	.27	.27	.26	.26	.26	.25	.25	.25	.25	.25	.25	.25	.25	.26	.26	.26	.27	.29
Tolerance	.66	.28	.32	.31	.29	.29	.28	.28	.27	.27	.27	.27	.27	.27	.27	.27	.27	.28	.28	.29	.30	.31
Virtue ¹	.51	.35	.35	.34	.34	.34	.34	.34	.34	.34	.34	.34	.35	.35	.35	.35	.35	.35	.35	.35	.35	.39

Note: N = 33,678. ¹Experimental facet included in Section 2.

Table 33. Marginal Reliabilities and Conditional Standard Errors for Version 14 (Parts 1 and 2, 176 Items)

Facet Name	Marginal Reliability	Average SE	Average Standard Errors at Each Percentile Range																			
			1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Achievement	.58	.29	.30	.29	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.29	.29	.29	.30	.30	.31	.32
Adjustment	.65	.23	.26	.23	.22	.22	.22	.21	.21	.21	.21	.22	.22	.22	.23	.23	.24	.24	.25	.26	.28	.30
Army Self-Efficacy ¹	.60	.33	.32	.32	.32	.32	.32	.32	.32	.32	.32	.33	.33	.33	.33	.33	.33	.34	.34	.34	.34	.33
Attention Seeking	.62	.32	.33	.33	.33	.33	.32	.32	.32	.32	.32	.31	.31	.31	.31	.31	.31	.31	.31	.32	.33	.34
Commitment ¹	.68	.31	.31	.32	.32	.31	.31	.30	.30	.30	.30	.30	.30	.30	.30	.31	.31	.31	.32	.32	.32	.31
Dominance	.69	.26	.32	.28	.27	.26	.26	.25	.25	.24	.24	.24	.24	.24	.25	.25	.25	.26	.27	.28	.28	.29
Even Tempered	.63	.29	.31	.29	.28	.27	.27	.27	.28	.28	.28	.28	.29	.29	.29	.30	.30	.31	.31	.32	.33	.36
Humility ¹	.37	.43	.42	.42	.42	.42	.43	.43	.43	.43	.43	.43	.43	.43	.43	.43	.43	.42	.42	.41	.40	.42
Intellectual Efficiency	.66	.29	.32	.31	.30	.29	.29	.28	.28	.28	.28	.28	.28	.28	.28	.28	.29	.29	.29	.30	.31	.32
Non-Delinquency	.64	.29	.32	.30	.30	.29	.29	.29	.29	.28	.28	.28	.28	.28	.28	.29	.29	.29	.29	.30	.30	.30
Optimism	.52	.30	.31	.29	.29	.29	.29	.29	.28	.29	.29	.29	.29	.29	.29	.30	.30	.30	.31	.31	.32	.33
Order	.69	.29	.34	.32	.32	.31	.31	.30	.29	.29	.28	.28	.28	.27	.27	.27	.27	.27	.27	.28	.29	.30
Physical Conditioning	.74	.27	.32	.30	.29	.28	.28	.27	.26	.25	.25	.25	.25	.25	.25	.25	.25	.26	.27	.28	.29	.30
Sociability	.74	.27	.29	.31	.30	.30	.29	.28	.28	.27	.27	.26	.26	.26	.25	.25	.25	.25	.25	.25	.26	.28
Team Orientation	.64	.26	.29	.28	.27	.27	.26	.26	.26	.25	.25	.25	.25	.25	.25	.25	.25	.25	.26	.26	.27	.29
Tolerance	.65	.28	.32	.30	.29	.28	.28	.27	.27	.27	.27	.27	.27	.27	.27	.27	.27	.28	.28	.29	.30	.31
Virtue ¹	.55	.35	.35	.34	.34	.34	.34	.34	.34	.34	.34	.34	.35	.35	.35	.35	.35	.35	.35	.35	.35	.39

Note: N = 33,732. ¹Experimental facet included in Section 2.

Table 34 provides a summary of the marginal reliabilities for the facets that have been administered in the MEPS from V4 to V14.

Table 34. Summary of Marginal Reliabilities Across All TAPAS MEPS Army Versions

Facet Name	V4	V5	V7	V8	V9	V10	V11	V12,13,14 (part 1 only)	V12 (part1&2)	V13 (part1&2)	V14 (part1&2)
Achievement	.58	.58	.55	.55	.62	.60	.62	.59	.59	.57	.58
Adjustment	.56	.62	.55	.54				.64	.66	.65	.65
Adventure Seeking			.73								
Army Self-Efficacy										.60	.60
Attention Seeking	.67	.66	.66	.65			.63	.61	.62	.63	.62
Commitment to Serve			.50		.66		.72		.71	.68	.68
Cooperation	.38	.65	.64		.70						
Courage				.49		.56					
Dominance	.76	.69	.68	.66	.73	.72	.71	.68	.69	.69	.69
Even Tempered	.59	.53	.55	.57	.64	.67	.68	.63	.65	.63	.63
Humility									.42		.37
Intellectual Efficiency	.68	.64	.63	.64	.67	.66	.65	.65	.67	.66	.66
Non-Delinquency	.52	.56	.60	.59		.62		.64	.66	.64	.64
Optimism	.50	.44	.43	.42	.47	.47	.52	.50	.53	.50	.52
Order	.69	.65	.66		.69	.69	.69	.67	.69	.68	.69
Physical Conditioning	.78	.72	.72	.71	.74	.75	.75	.73	.76	.75	.74
Persistence									.53	.50	
Responsibility				.58	.65						
Self-Efficacy									.46		
Self-Control	.47	.48		.48							
Selflessness	.51	.65	.62		.66	.66	.67				
Situational Awareness			.43			.49					
Sociability	.64	.74		.72	.78	.78	.76	.73	.76	.74	.74
Team Orientation				.61			.64	.61	.65	.63	.64
Tolerance	.59	.64		.61	.67	.67	.67	.64	.66	.66	.65
Virtue										.51	.55

10.0 TEST-RETEST RELIABILITY⁴

Although it has been used for several years now, little systematic information is available about the reliability of the TAPAS. The previous section provides IRT marginal reliabilities and conditional SEs for trait estimates. Another form of reliability that can be computed for the TAPAS is test-retest reliability. Note that estimates of reliability based on the test-retest approach are meaningful to the extent that respondents are equally and identically motivated on both occasions. Specifically, respondents need to be motivated in the same way on both occasions: They cannot be motivated to answer honestly on one occasion and motivated to fake good on the other. Also, it is important that the adaptive algorithm produces forms that are similar (i.e., parallel) on both occasions. If the forms differ in content or extremity, the observed correlation will be lower than the true reliability of either form. Thus, the correlation here is an alternate form reliability estimate and should be viewed as a lower bound to the facets' reliabilities.

As described above, different versions of the TAPAS have been administered in various settings. Although each version varies the specific facets that are assessed, they typically assess a large number of facets (i.e., typically 13-15) in a relatively short period of time (e.g., MEPS TAPAS generally takes about 20 minutes). In addition, the number of items that can be administered is also limited (i.e., typically between 120 and 130 depending on the number of facets) due to the available testing time and test-taker fatigue. Each of these factors can limit the potential reliability of the TAPAS. Therefore, factors that influence the reliability of the TAPAS are explored in this section.

We systematically varied the characteristics of the TAPAS versions administered across five studies to examine the influence of several factors including the administration mode (i.e., static versus adaptive assessment), the specific facets that were assessed, the number of items administered, and the number of statements per facet being assessed. Here, we differentiate between test length and the amount of information gathered on each facet. Because TAPAS is administered in a forced-choice format, test length is defined as the number of forced-choice items. However, each of these items includes statements from two different TAPAS facets and therefore the number of statements administered equals twice the number of items. Consequently, the number of statements administered for each facet influences the amount of information obtained and the overall reliability of each facet. The methods and results of the studies used to examine these methods are described next.

⁴ This section is reproduced from Nye, Chernyshenko, Stark, Drasgow, O'Brien, and White (2022) *Improving the Tailored Adaptive Personality Assessment System (TAPAS) for enlisted and officer selection* (Research Report XXXX). Fort Belvoir: U.S. Army Research Institute for the Behavioral and Social Sciences. Reproduced with permission. We gratefully acknowledge ARI's support for this work.

10.1 Study 1

10.1.1 Method

The data for Study 1 were collected from 946 Soldiers in the U.S. Army between December 2016 and January 2017 at several locations within and outside the continental United States. This sample was approximately 88% male and 51% White. Each of these individuals completed two static versions of the TAPAS administered in the same research session. Soldiers first completed a version of the TAPAS that included 94 items. This was followed by a version of the ALQ, which is an assessment of Soldiers' attitudes, experiences, and background in the Army. The purpose of including the ALQ in this session was to provide Soldiers with a short break between TAPAS administrations. After completing the ALQ, Soldiers then completed a second 83-item form of the TAPAS that was created to be as similar as possible to the first form in terms of overall content. An alternate form of the TAPAS (i.e., rather than the same form) was used to reduce the potential for response biases due to Soldiers responding to the same items twice within a two-hour period. The two static TAPAS forms used for this study assessed 10 facets that were selected from the Army version of the MEPS TAPAS, and all assessments were administered as static paper-and-pencil forms.

10.1.2 Results

The data for this study were first screened for unmotivated responding using several response check items (e.g., "Please select option B for this pair") embedded in the test form. We removed anyone who missed one or more of these response check items, resulting in a sample size of 845 for the test-retest analyses. The results of these analyses are shown in Table 35. As shown here, several of the TAPAS facets had modest test-retest reliabilities. For example, the Dominance facet had a reliability of .68 and the Sociability facet had a reliability of .64. However, several facets also had low test-retest reliabilities. The Achievement facet had a reliability of .48 while the Order and Selflessness facets had reliabilities of .49 and .41, respectively.

Despite these lower reliabilities, individual TAPAS facets are not used for making selection or assignment decisions in the Army. Instead, composites of TAPAS scales are used. Table 35 also reports the test-retest reliabilities of the TAPAS Can-Do, Will-Do, and Adaptation composites developed for screening entry-level Soldiers (Nye et al., 2013). The test-retest reliabilities for these composites were generally higher than the reliabilities for the constituent facets. The reliabilities for these composites were .62 for the Can-Do composite, .64 for the Will-Do composite, and .52 for the Adaptation composite.

The test-retest reliabilities in this sample are lower than desired. However, several factors potentially mitigate these reliability estimates. First, administering two versions of the TAPAS in the same session could have resulted in test-taker fatigue and boredom, which might have influenced the consistency of responses. A second factor that could have influenced the reliability estimates in this study is that two alternate forms of the TAPAS were administered rather than administering the same version twice. In other words, the items were not the same at each time point. Thus, rather than test-retest reliability, this study provided an estimate of alternate form reliability. Such reliability estimates tend to be lower, on average, than test-retest reliabilities. Third, the rather small number of statements used to assess each facet could have

also lowered the reliabilities. As CTT shows, test length influences reliability. In this study, the ten-dimensional form included about 17 statements per facet; increased reliability would be expected if the number of statements per facet were increased. Fourth, a static administrative format was used; Stark et al. (2012) showed that adaptive administration can substantially increase reliability given a fixed test length.

In Study 2 we examined the effect of increasing the number of statements administered for each facet on the test-retest reliabilities. To mitigate the potential effects of fatigue and boredom, fewer facets were administered so that the overall burden for the Soldiers was held constant.

Table 35. TAPAS Test-Retest Reliabilities in Study 1

TAPAS Facet	Test-Retest Reliability
Achievement	.48
Dominance	.68
Even Tempered	.57
Intellectual Efficiency	.60
Optimism	.60
Order	.49
Physical Conditioning	.59
Selflessness	.41
Sociability	.64
Tolerance	.50
Can-Do Composite	.62
Will-Do Composite	.64
Adaptation Composite	.52
<i>Note.</i> N = 845. Both TAPAS forms administered in the same research session.	

10.2 Study 2

10.2.1 Method

The data for Study 2 were collected from 585 Soldiers in the U.S. Army at three different locations. This sample was approximately 85% male, 54% White, and 89% were E-4 or below. As in Study 1, each of these individuals completed two static versions of the TAPAS administered in the same research session. Soldiers first completed a version of the TAPAS, followed by the ALQ. After completing the ALQ, participants were administered a second form of the TAPAS. Consequently, the reliabilities estimated in this study were also alternate form reliabilities. The two alternate forms of the TAPAS each assessed 8 facets, which was fewer than in Study 1. This was done so that we could increase the number of statements assessing each facet while maintaining approximately the same number of items

used in Study 1 (i.e., the forms used in the current study included 84 items). Both TAPAS forms included about 20 statements assessing each facet and were administered in a paper-and-pencil format.

10.2.2 Results

The data for this study were first screened for unmotivated responding as in Study 1 using several response check items (e.g., “Please select option B for this pair”) embedded in the test form. We removed anyone who missed one or more of these response check items resulting in an effective sample size of 411 for the test-retest analyses. The results of these analyses are shown in Table 36. Not surprisingly, increasing the number of statements assessing each facet also increased the reliabilities of the TAPAS facets. The reliabilities for nearly all the facets that were also assessed in Study 1 were higher in Study 2. This included the TAPAS composites where both the Will-Do and Adaptation composites had reliabilities above .70. These results indicate that increasing the number of statements assessing each dimension improves the test-retest reliabilities of the facets.

Table 36. TAPAS Test-Retest Reliabilities in Study 2

TAPAS Facet	Test-Retest Reliability
Achievement	.56
Dominance	.75
Even Tempered	.66
Intellectual Efficiency	.58
Non-Delinquency	.58
Optimism	.63
Physical Conditioning	.75
Sociability	.65
Can-Do Composite	.61
Will-Do Composite	.73
Adaptation Composite	.72
<i>Note.</i> N = 411. Both TAPAS forms administered in the same research session.	

Although the test-retest reliability estimates were higher in this study than in Study 1, the design of this study still had several limitations that could lower the reliabilities. Again, two alternate forms of the TAPAS were administered. Although these two forms were designed to be as similar as possible in terms of their content and item characteristics, it is nonetheless unlikely that these two forms were truly parallel. Therefore, the reliability estimates shown in Table 36 represent test-retest with alternate forms, which tend to be lower than test-retest reliabilities. A second limitation was that both forms were again administered in the same session, raising questions about the potential effects of Soldier fatigue and boredom on item responses and their consistency across TAPAS forms.

10.3 Study 3

10.3.1 Method

Study 3 represents a combination of Studies 1 and 2 in that the TAPAS forms included 10 facets as in Study 1, but with the increased number of statements per facet used in Study 2. As in the previous studies, Soldiers were administered two alternate forms of TAPAS. In Study 3, some Soldiers were administered paper-and-pencil forms and other Soldiers responded via computer. The paper-and-pencil format was used for a sample of 211 Soldiers at two different locations. This sample was approximately 39% White, 25% Black, and 25% Hispanic. In addition, 83% were male and 96% were E-4 or below. The computerized format was utilized for a sample of 258 Soldiers from two different locations. This sample was approximately 38% White, 28% Black, and 22% Hispanic. The sample was also 80% male and 93% were E-4 or below.

Both the computerized and the paper-and-pencil forms were administered as static assessments. As with Studies 1 and 2, both the test and the retest were administered in the same session for both formats. Soldiers completed one version of the TAPAS, which was followed by the ALQ. Following the ALQ, participants completed a second alternate form of the TAPAS. Consequently, the reliabilities estimated in this study were again test-retest with alternate forms. The two alternate forms of the TAPAS each assessed 10 facets. These facets were selected to provide information on facets that were not administered in Studies 1 or 2. However, consistent with Study 2, each form contained an increased number of statements per facet to increase the reliability.

10.3.2 Results

The data for this study were first screened for unmotivated responding using several response check items (e.g., “Please select option B for this pair”) embedded in the test form. We removed anyone who missed any of these response check items resulting in an effective sample size of 155 for analyses with the paper form and 164 for analyses with the computerized form. The results of these analyses are shown in Tables 37 and 38 for the paper and computerized forms, respectively. The results shown in these tables are consistent with the results of Study 1. This is not surprising given that Study 3 used the same procedures as in Study 1—that is, same session test-retest with alternate forms. However, the test-retest reliabilities for some facets were also higher than in Study 1 due to the increased number of statements per facet. In other words, adding more statements per facet clearly has a positive effect on the reliability of the TAPAS.

Table 37. TAPAS Test-Retest Reliabilities for the 10D Paper Form in Study 3

TAPAS Facet	Test-Retest Reliability
Achievement	.56
Adjustment	.51
Attention Seeking	.66
Even Tempered	.61
Non-Delinquency	.58
Optimism	.63
Order	.52
Physical Conditioning	.65
Sociability	.58
Team Orientation	.59
Can-Do Composite	.60
Will-Do Composite	.73
Adaptation Composite	.59
<i>Note.</i> N = 155. Both TAPAS forms administered in the same research session	

Table 38. TAPAS Test-Retest Reliabilities for the 10D Computerized Form in Study 3

TAPAS Facet	Test-Retest Reliability
Achievement	.55
Adjustment	.53
Attention Seeking	.66
Even Tempered	.64
Non-Delinquency	.68
Optimism	.73
Order	.50
Physical Conditioning	.70
Sociability	.66
Team Orientation	.59
Can-Do Composite	.52
Will-Do Composite	.72
Adaptation Composite	.56
<i>Note.</i> N = 164. Both TAPAS forms administered in the same research session.	

Although the test-retest reliabilities were higher in this study than in Study 1, the design of this study had several limitations that could lower the reliabilities. For example, the same session study design is still an important limitation of these results given the potential mitigating effects of test-taker fatigue. In addition, although the two alternate forms that were administered were designed to be as similar as possible in terms of their content and item characteristics, it is still not possible to make these forms truly parallel. To address these issues, we next conducted a fourth study with a different research design.

10.4 Study 4

10.4.1 Method

The data for Study 4 were collected from 580 respondents via Amazon's Mechanical Turk (MTurk). This sample was approximately 70% female and the mean age was 40.12. In contrast to Studies 1, 2, and 3, participants in Study 4 completed the *same* static TAPAS form at two different time points approximately 10 days apart. In other words, Study 4 was the first true test-retest reliability study of the TAPAS. The version of TAPAS used in this study assessed the same 10 facets examined in Study 1 and was administered on-line at both time points. There were about 17 statements per facet.

10.4.2 Results

The data for this study were first screened for unmotivated responding as done previously using several response check items (e.g., "Please select option B for this pair") embedded in the test form. We removed anyone who missed more than one of these response check items resulting in an effective sample size of 562 for the test-retest analyses. The results of these analyses are shown in Table 39. Although the same 10 facets assessed in Study 1 were measured, the test-retest reliabilities in the MTurk sample were substantially higher. All but one of the reliabilities were above .70 (and the exception was .69) and several were .80 or above. Thus, these results suggest that the TAPAS has sufficient reliability for assessing these 10 facets.

Table 39. TAPAS Test-Retest Reliabilities in Study 4

TAPAS Facets	Test-Retest Reliability
Achievement	.69
Dominance	.80
Even Tempered	.77
Intellectual Efficiency	.73
Optimism	.77
Order	.83
Physical Conditioning	.79
Selflessness	.73
Sociability	.83
Tolerance	.77
Can-Do Composite	.78
Will-Do Composite	.78
Adaptation Composite	.75

Notes. N = 562. Identical TAPAS forms administered approximately 10 days apart.

The reliability estimates found in this study were substantially higher than in the earlier studies. This is perhaps not surprising given that the same TAPAS form was administered at both time points (i.e., in contrast to the studies that used different forms) and were administered 10 days apart rather than in the same session, which likely reduced test-taker fatigue. Importantly, the test-retest reliabilities of the TAPAS composites, which are typically used for screening Soldiers, ranged from .75 to .78. These results indicate that the TAPAS composites have substantial test-retest reliability. Despite these promising results, we conducted an additional study to examine the test-retest reliabilities of a computer adaptive version of TAPAS. The versions of TAPAS administered in Studies 1 through 4 were all static forms. However, a computer adaptive version of the TAPAS is administered at the MEPS. Therefore, we next examined the reliability of adaptive TAPAS. In addition, the sample used in Study 4 was older (mean age was 40.12) than the typical applicant at the MEPS. Therefore, we also examined the test-retest reliability of the adaptive version of the TAPAS in a younger sample.

10.5 Study 5

10.5.1 Method

The data for Study 5 were collected from 314 respondents on Amazon's MTurk. This sample was 69% female and the mean age was 23.27. All participants completed a computer adaptive version of the TAPAS assessing 13 facets. The purpose of assessing these 13 facets was to simulate the number of facets assessed at the MEPS. To estimate test-retest reliabilities, this adaptive version of TAPAS was administered twice approximately 10 days apart.

10.5.2 Results

The data for this study were first screened for unmotivated responding as in previous studies using several response check items (e.g., “Please select option B for this pair”) embedded in the test form. We removed anyone who missed more than one of these response check items resulting in an effective sample size of 310 for the test-retest analyses.

The results of the analysis of Study 5 data are shown in Table 40. The reliabilities of the 13 TAPAS facets assessed in this study varied substantially. The lowest reliability was observed for Non-Delinquency with a test-retest reliability of .39, but several of the facets also had test-retest reliabilities above .70. More importantly, the reliabilities of the three TAPAS composites were above .70. Again, the reliabilities of these composites are most important given that they are used to make important personnel decisions.

Table 40. TAPAS Test-Retest Reliabilities in Study 5

TAPAS Facets	Test-Retest Reliability
Intellectual Efficiency	0.66
Achievement	0.60
Adjustment	0.68
Attention Seeking	0.61
Dominance	0.65
Even Tempered	0.61
Non-Delinquency	0.39
Optimism	0.64
Order	0.71
Physical Conditioning	0.69
Sociability	0.72
Team Orientation	0.59
Tolerance	0.64
Can-Do Composite	0.72
Will-Do Composite	0.74
Adaptation Composite	0.74

Notes. N = 310. Adaptive version of TAPAS administered twice approximately 10 days apart

Surprisingly, the test-retest reliabilities for the adaptive version of TAPAS were slightly lower than the test-retest reliabilities of the static form used in Study 4. There are several factors that might have influenced the reliability estimates in this study. First, due to the greater number of facets assessed in this study, fewer statements per facet were administered. As demonstrated in previous studies and shown with CTT, the number of statements per facet is an important determinant of reliability.

If the assumptions of IRT are correct, then an adaptive assessment should substantially increase reliability for a fixed test length. However, the assumptions are not necessarily true and one likely culprit is the assumption of unidimensionality. The Big Five personality domains are nowhere close to unidimensional and even their underlying facets sometimes appear noticeably multidimensional. To the extent that the facets are multidimensional, adaptive administrations yield tests that are not identical. In this case, the reliability estimates reported in Table 40 would represent test-retest with alternative forms reliabilities. As noted above, such estimates of reliability would be lower than true test-retest reliabilities because the alternate forms are not truly parallel.

10.6 Discussion

Several lessons can be learned from the five TAPAS reliability studies described in this section. First, across the studies, the test-retest reliabilities of the facets were often acceptable. In particular, although the reliabilities of some scales were unacceptably low (e.g., Non-Delinquency in Study 5), the vast majority of the scales had much higher reliabilities. Importantly, the test-retest reliabilities of the TAPAS composites, which were designed to identify high potential Soldiers for selection and assignment, were generally above .70. This is important because it suggests that the composites that are potentially used for making personnel decisions do have acceptable reliabilities, despite sometimes lower reliabilities for individual TAPAS facets. Nevertheless, there were also several factors that affected the reliabilities across these studies.

First, the length of time between assessments appeared to have a substantial effect on the test-retest reliability. In studies where the TAPAS was administered twice in a single session, the test-retest reliabilities tended to be lower. One possible explanation for this effect was that respondents became more fatigued or bored when responding to the same test twice in the same session. These effects likely affected their responses and lowered the test-retest reliability. In contrast, when the two administrations were separated by 10 days, the reliabilities tended to be larger. It is important to note that the studies in which the TAPAS was administered twice in the same session were constrained by the realities of testing active-duty Soldiers. In these studies, it was not possible to separate the two administrations across multiple sessions. Nevertheless, this constraint does seem to have influenced the reliability estimates.

Not surprisingly, the number of statements administered per facet also had a substantial effect on reliability. TAPAS forms with more statements per facet showed higher reliabilities than versions with fewer statements per facet. This has important implications for future versions of the TAPAS and suggests that new TAPAS forms should include an adequate number of statements per facet. Based on the results presented here, it appears that at least 20 statements per facet should be included. Nevertheless, the increased number of statements will need to be balanced with the number of facets so that the total amount of testing time is reasonable.

11.0 SUMMARY AND DISCUSSION

This report began with a summary of the history of the TAPAS, including the reasons for its original development. The initial version, TAPAS-95s, was described and research investigating its properties was summarized. Then TAPAS assessment at the MEPS was reviewed and a discussion of the facets that have been developed was given. The process of developing pools of statements for TAPAS facets was outlined and additional TAPAS forms created to support DoD personality research were summarized. Then the development of TAPAS composites was described. This was followed by a detailed description of configurable specifications and the adaptive algorithm. Next, TAPAS scoring, norming, and score diagnostics were described. The next two sections provided information on the reliability and conditional standard errors of scores.

Over the past 15 years, numerous reports and journal papers have been written. For example, details of the original development of TAPAS are given in Drasgow et al. (2012). This report provides information about origin of the facet framework and, specifically, the data and analyses that led to the framework. Studies have been conducted for specific MOS; for example there are reports for Special Forces (Nye et al., 2014), Recruiters (Nye et al., 2018a), and DS (Nye, et al., 2020a). Research has not been limited to the Army. For example, Trent et al., (2020) and Chernyshenko et al. (2019) describe research conducted with Air Force personnel.

So, what does the future hold for TAPAS? There are certainly many avenues for additional research. For example, it was found that customizing the TAPAS composites for a few specific MOS led to improved prediction (Nye, et al., 2020b) of important outcomes. It may be possible to group MOS into a manageable number of clusters and develop optimal composites for each cluster.

Another avenue for research is to move from the 2AFC format to triples for each item where the respondent indicates the statement that is “Most like me” and the statement that is “Least like me.” From a triple, three pseudo-2AFC items can be created. For example, for statements A, B, and C, if statement A is most preferred and statement C is least preferred, we know A is preferred to B and B is preferred to C in addition to A being preferred to C. A small demonstration study reported in Nye et al. (2022) showed substantial gains in reliability for an 84 item three-alternative forced choice (3AFC) instrument relative to a 120 item 2AFC measure with the same facets: Marginal reliabilities were generally in the range of .75 to .85 for facets assessed with the 3AFC measure in comparison to .60 to .75 for the 2AFC measure.

A third line of possible inquiry is contextualization. In the civilian research literature, it has been found that adding the tag “at work” improves prediction of job performance (Shaffer & Postlethwaite, 2012). Thus, the Agreeableness statement “I get along well with others” statement would be revised to be “I get along well with others at work.” Would adding an appropriate tag for the military context improve validity here as well?

Ultimately, the value of TAPAS to the Services will lie in its usefulness in improving selection and classification. For example, it has been found that individuals with good TAPAS scores and ASVAB scores placing them in AFQT category 3B perform like individuals with AFQT 3A

scores (Nye et al., 2013). But individuals with problematic TAPAS scores and AFQT 3B scores perform substantially worse. Of course, the extent to which individuals with AFQT 3B scores are recruited and enlisted is a policy decision, but from an empirical standpoint, the data appears to support tapping into this source of applicants. Research has also shown that placing individuals into a MOS consistent with their personalities can yield improved performance. For example, Nye et al. (2020b) found that approximately 40% of individuals might have performed substantially better (a half-SD or more) if they were assigned to a different MOS.

In conclusion, there has been a great deal of research conducted with TAPAS over the past 15 years. We hope this work proves to be of value to the US Military Services.

12.0 REFERENCES

- Allen, M.T., Cheng, Y. A., Putka, D. J., Hunter, A., & White L. (2010). Analysis and findings. In D. J. Knapp & T. S. Heffner (Eds.). *Expanded enlistment eligibility metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement*, 19(3), 269–290. <https://doi.org/10.1177/014662169501900306>
- Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44, 1-26. <https://doi.org/10.1111/j.1744-6570.1991.tb00688.x>
- Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, 15, 263-272.
- Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences*, 11(5), 515–524. [https://doi.org/10.1016/0191-8869\(90\)90065-Y](https://doi.org/10.1016/0191-8869(90)90065-Y)
- Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, 111(2), 218–229. <https://doi.org/10.1037/pspp0000085>
- Campbell, J. P., & Knapp, D. J. (Eds.) (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Erlbaum
- Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, 104, 1347–1368. <https://doi.org/10.1037/apl0000414>
- Chernyshenko, O. S., Stark, S., & Chan, K. Y. (2001). Investigating the hierarchical factor structure of the Fifth Edition of the 16PF: An application of the Schmid-Leiman orthogonalization. procedure. *Educational and Psychological Measurement*, 61, 290-302.
- Chernyshenko, O. S., Stark, S., & Drasgow, F. (2011). Individual differences, their measurement and validity. In S. Zedeck (Ed.) *APA Handbook of industrial and*

organizational psychology. pp. 117-151. Washington: American Psychological Association.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumption of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, 19, 88-106. <https://doi.org/10.1037/1040-3590.19.1.88>

Chernyshenko, O. S., Zhang, B., Drasgow, F., Stark, S., & Nye, C. D. (2019). *The development of three static personality research forms and on-line scoring tools for the U.S. Air Force* (Research Report AFRL-RH-WP-TR-2019-0090). Wright-Patterson AFB, OH: Air Force Research Laboratory.

Conn, S., & Reike, M. L. (Eds.). (1994). *The 16PF fifth edition technical manual*. Champaign, IL: Institute for Personality and Ability Testing.

Costa, P. T., Jr., & McCrae, R. R. (1988). From catalog to classification: Murray's needs and the five-factor model. *Journal of Personality and Social Psychology*, 55(2), 258–265. <https://doi.org/10.1037/0022-3514.55.2.258>

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and the NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., McCrae, R. R., & Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences*, 12, 887-898. [https://doi.org/10.1016/0191-8869\(91\)90177-D](https://doi.org/10.1016/0191-8869(91)90177-D)

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41, 417-440.

Drasgow, F., Chernyshenko, O. S., Stark, S., Nye, C. D., Zhang, B., Sun, T., & Li, L. (2020). *Development of the dark tetrad research forms for the U.S. Air Force* (Research Report AFRL-RH-WP-TR-2020-0106). Wright-Patterson AFB, OH: Air Force Research Laboratory.

Drasgow, F., Levine, M. V. & Williams, E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67-86. <https://doi.org/10.1111/j.2044-8317.1985.tb00817.x>

Drasgow, F., Nye, C. D., Chernyshenko, O. S., Stark, S., & O'Brien, E. (in press). *Evaluation of the Tailored Adaptive Personality Assessment System (TAPAS): Final Report* (Technical Report xxxx). Ft. Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

- Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support selection and classification decisions* (Technical Report 1311). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6), 1216–1229. <https://doi.org/10.1037/0022-3514.59.6.1216>
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26-34. <https://doi.org/10.1037/0003-066X.48.1.26>
- Goldberg, L. R. (1997). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe*, Vol. 7. The Netherlands: Tilburg University Press.
- Gough, H. G. (1987). *The California Psychological Inventory administrators guide*. Palo Alto, CA: Consulting Psychologists Press.
- Harvey, J., Hyland, J., Nye, C., Patterson, J., & Morath, R. (2018). *Marine Corps non-cognitive testing: Final briefing/report*. Fairfax, VA: ICF.
- Hirsh, J. B., & Peterson, J. B. (2008). Predicting creativity and academic success with a "fake-proof" measure of the Big Five. *Journal of Research in Personality*, 42(5), 1323–1333. <https://doi.org/10.1016/j.jrp.2008.04.006>
- Hogan, R. T. (1991). Personality and personality measurement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 873–919). Consulting Psychologists Press.
- Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.
- Horgen, K. E., Nye, C. D., White, L. A., Laporte, K. A., Hoffman, R. R., Drasgow, F., Chernyshenko, O. S., Stark, S., & Conway, J. S. (2013). *Validation of the Non-Commissioned Officer Special Assignment Battery (NSAB)* (Technical Report 1328). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Jackson, D. N. (1994). *Jackson Personality Inventory – Revised manual*. Port Huron, MI: Sigma Assessment Systems, Inc.
- Joo, S.H., Lee, P., & Stark, S. (2018). Development of information functions and information indices for the GGUM-RANK multidimensional forced choice model. *Journal of Educational Measurement*, 55, 357-372.

- Kilcullen, R. N., Putka, D. J., McCloy, R. A., & Van Iddekinge, C. H. (2005). Development of the Rational Biodata Inventory. In D.J. Knapp, C.E. Sager, & T.R. Tremble (Eds.). *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (pp. 105-116) (Technical Report 1168). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Kim, S., & Lee, W.-C. (2004). *IRT scale linking methods for mixed-format tests* (ACT Research Report Series 2004-5). Iowa City, IA: ACT, Inc.
- Kirkendall, C., Bynum, B., Nesbitt, C., & Hughes, M. (2020). Validation of the TAPAS for predicting in-unit soldier outcomes. *Military Psychology*, 32, 24-35.
- Knapp, D. J., & Heffner, T. S. (Eds.) (2010). *Expanded Enlistment Eligibility Metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., Heggestad, E. D., & Young, M. C. (2004). *Understanding and improving the Assessment of Individual Motivation (AIM) in the Army's GED Plus Program* (Study Note 2004-03). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Lee, K., Ashton, M. C., & de Vries, R. E. (2005). Predicting workplace delinquency and integrity with the HEXACO and five-factor models of personality structure. *Human Performance*, 18(2), 179–197. https://doi.org/10.1207/s15327043hup1802_4
- Lee, P., Joo, S. H., Stark, S., & Chernyshenko, O. S. (2019). GGUM-RANK statement and person parameter estimation with multidimensional forced choice triplets. *Applied Psychological Measurement*, 43, 226-40.
- Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences*, 123, 229-235.
- Legree, P. J., Kilcullen, R. N., Putka, D. J., & Wasko, L. E. (2014). Identifying the leaders of tomorrow: Validating predictors of leader performance. *Military Psychology*, 26, 292-309.
- Martin, C. J. (1998). Comparison of norming procedures in 1980 and 1997 (Archive Document No. CV98001). Monterey, CA: Defense Manpower Data Center.
- Martin, C. J. (1999, November). Overview of the Profile of American Youth 1997. In J. R. Welsh (Chair) *Applied issues in ASVAB development*. Symposium conducted at the annual conference of the International Military Testing Association, Monterey, CA.
- McCrae, R. R., Costa, P. T., & Piedmont, R. L. (1993). Folk concepts, natural language, and psychological constructs: The California Psychological Inventory and the Five-Factor

model. *Journal of Personality*, 61(1), 1–26. <https://doi.org/10.1111/j.1467-6494.1993.tb00276.x>

- Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 66, 574–583.
- Nye, C. D., Beal, S. A., Chernyshenko, O. S., Stark, S., Drasgow, F., White, L. A., & Dressel, J. D. (2014). *Assessing the Tailored Adaptive Personality Assessment System (TAPAS) for Special Forces personnel* (Research Report 1971). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Nye, C. D., Chernyshenko, O. S., Stark, S., Drasgow, F., O'Brien, E., & White, L. A. (in press). *Improving the Tailored Adaptive Personality Assessment System (TAPAS) for enlisted and officer selection*. Ft. Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Nye, C. D., Drasgow, F., Chernyshenko, O. S., Stark, S., Kubisiak, C., White, L. A., & Jose, I. (2012a). *Assessing the Tailored Adaptive Personality Assessment System (TAPAS) as an MOS qualification instrument* (Technical Report 1312). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Nye, C. D., Drasgow, F., Stark, S., Chernyshenko, O. S., & White, L. S. (2012b). Appendix E: Development of TAPAS composites for predicting Army-wide criteria. In D. J. Knapp & T.S. Heffner (Eds.). *Tier One Performance Screen initial operational test and evaluation: 2011 interim report* (Technical Report 1306). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Nye, C. D., Muhammad, R. S., Graves, C., Drasgow, F., Chernyshenko, O. S., Stark, S., & Butt, S. (2018a). *Examining enhanced suitability screening for predicting performance in recruiting duty assignments* (Technical Report 1366). Ft. Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Nye, C. D., Muhammad, R. S., Wolters, H. M., Drasgow, F., Chernyshenko, O. S., & Stark, S. (2018b). *New scale development for enhanced suitability screening (ESS)* (Research Note 2018-01). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Nye, C. D., Stark, S., Drasgow, F., Chernyshenko, O. S., & White, L. S. (2013). Appendix A: Developing revised TAPAS composites for predicting Army-wide criteria. In D. J. Knapp & T.S. Heffner (Eds.). *Tier One Performance Screen initial operational test and evaluation: 2012 interim report* (Technical Report 1332). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Nye, C. D., White, L. A., Drasgow, F., Chernyshenko, O. S., Stark, S., Graves, C. R., & Muhammad, R. S. (2020a). *Examining enhanced suitability screening for predicting drill*

- sergeant training and job outcomes* (Technical Report 1388). Ft. Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Nye, C. D., White, L. A., Drasgow, F., Prasad, J., Chernyshenko, O. S., Stark, S., & Kubisiak, C. (2020b). *Non-cognitive tools for military occupational specialties qualification* (Technical Report 1387). Ft. Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Nye, C. D., White, L. A., Drasgow, F., Prasad, J., Chernyshenko, O. S., & Stark, S. (2020c). Examining personality for the selection and classification of soldiers: Validity and differential validity across jobs. *Military Psychology*, 32, 60-70.
- Nye, C. D., White, L. A., Horgen, K., Drasgow, F., Stark, S., & Chernyshenko, O. S. (2020d). Predictors of attitudes and performance in U. S. Army recruiters: Does personality matter? *Military Psychology*, 32, 81-90.
- O'Neill, T. A., Lewis, R. J., Law, S. J., Larson, N., Hancock, S., Radan, J., Lee, N., & Carswell, J. J. (2017). Forced-choice pre-employment personality assessment: Construct validity and resistance to faking. *Personality and Individual Differences*, 115, 120–127. <https://doi.org/10.1016/j.paid.2016.03.075>
- Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism and psychopathy. *Journal of Research in Personality*, 36(6), 556–563. [https://doi.org/10.1016/S0092-6566\(02\)00505-6](https://doi.org/10.1016/S0092-6566(02)00505-6)
- Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1990). *Numerical recipes: The art of scientific computing*. NY: Cambridge University Press.
- Roberts, B., Chernyshenko, O.S., Stark, S., & Goldberg, L. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology*, 58, 103-139. <https://doi.org/10.1111/j.1744-6570.2005.00301.x>
- Roberts, J. S., Arthur, W. Jr., 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* (Research Report 2019 No. 076). Alexandria, VA: Human Resources Research Organization.
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses using item response theory. *Applied Psychological Measurement*, 24, 2-32. <https://doi.org/10.1177/01466216000241001>
- Roberts, J. S., Fang, H., Cui, W., & Wang, Y. (2006). GGUM2004: A Windows-based program to estimate parameters in the generalized graded unfolding model. *Applied Psychological Measurement*, 30, 64-65.

- Rose, M. R., Manley, G. B., & Weissmuller, J. J. (2013). *Development of two- and three-factor classification models for Air Force Battlefield Airmen (BA) and related AFSs* (AFCAPS-TR-2013-0007). Randolph Air Force Base, TX: Air Force Personnel Center.
- Sackett, P. R., Putka, D. J., & McCloy, R. A. (2012). The concept of validity and the process of validation. In N. Schmitt (Ed.), *Oxford handbook of assessment and selection* (pp. 91-118). Oxford: Oxford University Press.
- Saucier, G. (1993). Openness versus intellect: Much ado about nothing? *European Journal of Personality*, 6, 381-386.
- Saucier, G., & Ostendorf, F. (1999). Hierarchical subcomponents of the Big Five personality factors: A cross-language replication. *Journal of Personality and Social Psychology*, 76, 613-627. <https://doi.org/10.1037/0022-3514.76.4.613>
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262-274. <https://doi.org/10.1037/0033-2909.124.2.262>
- Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology*, 65, 445-494.
- Stark, S. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment* [Doctoral Dissertation]. University of Illinois at Urbana-Champaign.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise preference model. *Applied Psychological Measurement*, 29, 184-201. <https://doi.org/10.1177/0146621604273988>
- Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods*, 15, 463-487. <https://doi.org/10.1177/1094428112444611>
- Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, 91, 25-39. <https://doi.org/10.1037/0021-9010.91.1.25>
- Stark, S., Chernyshenko, O. S., Nye, C. D., Drasgow, F., & White, L. A. (2017). *Moderators of the Tailored Adaptive Personality Assessment System validity* (Technical Report 1357). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

- Stark S., & Drasgow, F. (2002). An EM approach to parameter estimation for the Zinnes and Griggs paired comparison IRT model. *Applied Psychological Measurement*, 26, 208–227.
- Tellegen, A. (1982). *A brief manual for the Multidimensional Personality Questionnaire*. Unpublished manuscript, University of Minnesota.
- 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, 51-59. <https://doi.org/10.1080/08995605.2019.1652481>
- Tsai, W.-C., Chen, C.-C., & Liu, H.-L. (2007). Test of a model linking employee positive moods and task performance. *Journal of Applied Psychology*, 92(6), 1570–1583. <https://doi.org/10.1037/0021-9010.92.6.1570>
- U.S. Department of Defense. (1982a). *Profile of American youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery*. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics).
- U.S. Department of Defense. (1982b). *The profile of American youth: Results and implications* (Technical Memorandum 82-2). Washington, DC: Office of the Secretary of Defense.
- White, L. A., & Young, M. C. (1998). *Development and validation of the Assessment of Individual Motivation (AIM)*. Paper presented at the Annual Meeting of the American Psychological Association, San Francisco, CA.
- Woo, S. E., Chernyshenko, O. S., Longley, A., Zhang, Z-W., Chiu, C-Y., & Stark, S., E. (2014). Openness to experience: Its lower-level structure, measurement, and cross-cultural equivalence. *Journal of Personality Assessment*, 96, 29-45.
- Woo, S. E., Chernyshenko, O. S., Stark, S. E., & Conz, G. (2014). Validity of six openness facets in predicting work behaviors: A meta-analysis. *Journal of Personality Assessment*, 96(1), 76–86. <https://doi.org/10.1080/00223891.2013.806329>
- Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods*, 23, 569-590. <https://doi.org/10.1177/1094428119836486>

APPENDIX A

TAPAS MEPS IMPLEMENTATION AUTHORIZATION



REPLY TO
ATTENTION OF

DEPARTMENT OF THE ARMY
OFFICE OF THE DEPUTY CHIEF OF STAFF FOR PERSONNEL
300 ARMY PENTAGON
WASHINGTON, DC 20310-0300

DAPE-MPA

APR 03 2009

MEMORANDUM FOR SEE DISTRIBUTION

SUBJECT: Implementation of the Tier 1 Performance Screen (TOPS)

1. Background. The Army has developed a Tier 1 Performance Screen (TOPS) pilot program. This program is a non-cognitive measure designed to screen Regular Army (RA), United States Army Reserve (USAR) and Army National Guard (ARNG) Soldiers with Tier 1 credentials. Authority for this pilot program runs through 3QFY12 with semi-annual progress reports due to the Deputy Chief of Staff, G1.

2. Implementation. All RA, USAR and ARNG nonprior Service Tier 1 applicants will be required to take the Tailored Adaptive Personality Assessment System (TAPAS) test at the Military Entrance Processing Station (MEPS).

a. TAPAS Testing.

(1) TAPAS testing will begin on 4 May 09, at the following MEPS: Indianapolis, IN; Jackson, MS; Kansas City, MO; Salt Lake City, UT; Pittsburgh, PA; and Omaha, NE.

(2) The remaining MEPS will begin TAPAS testing on 8 Jun 09.

(3) As an exception to paragraph 2 above, nonprior Service Tier 1 Armed Forces Qualification Test (AFQT) Category I-IIIB applicants having a valid Armed Service Vocational Aptitude Battery (ASVAB) score from the Student ASVAB test or Military Entrance Test (MET) are not required to take the TAPAS.

(4) Nonprior Service Tier 1 AFQT Category IV applicants having a valid ASVAB score from the Student ASVAB test or MET are required to take the TAPAS at the MEPS.

(5) AFQT Category IV applicants testing and enlisting outside the continental United States (OCONUS), with the exception of Alaska, Hawaii, and Puerto Rico, are not required to take the TAPAS.

(6) Beginning 1 Oct 09, Tier 2 credential applicants will take the TAPAS in addition to the Assessment of Individual Motivation (AIM) and will continue until 7,500 applicants have taken both tests.

DAPE-MPA

SUBJECT: Implementation of the Tier 1 Performance Screen (TOPS)

b. TAPAS Screening Criteria.

(1) A passing TAPAS score is not required for enlistment of AFQT Category I-IIIIB applicants. There is no minimum TAPAS score required for AFQT Category I-IIIIB applicants regardless of where the ASVAB testing was done.

(2) AFQT Category IV applicants that do not pass TAPAS are not eligible for enlistment.

(3) AFQT Category IV applicants enlisting as English as a Second Language (09C) or Interpreter/Translator (09L) are eligible to enlist regardless of their TAPAS score.

c. TAPAS Retesting.

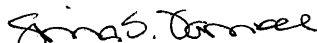
(1) Applicants who do not receive a passing score on the TAPAS or receive an incomplete result at the first test administration are authorized to retest immediately.

(2) Applicants who do not receive a passing score on the TAPAS or receive an incomplete result on the retest must wait one (1) year from the retest date before they are eligible for a second retest.

3. General coordinating instructions. The U.S. Army Accession Command (USAAC) will serve as the Army's POC for the TOPS operational test and evaluation during 3QFY09-3QFY12. During this time, USAAC will coordinate and receive test data from MEPCOM and ARI. USAAC will also coordinate and receive accession data from ARNG and USAREC (for RA and USAR input). USAAC, with supporting analyses from U.S. Army Research Institute for Behavioral and Social Sciences (ARI), will provide TOPS progress reports due to the G1 as stated above in paragraph 1.

4. Point of Contact for this office is Ms. Denise Mills, (703) 695-9262 or DSN 225-9262.

FOR THE DEPUTY CHIEF OF STAFF, G1:


GINA S. FARRISEE
Major General, GS
Director of Military
Personnel Management

DISTRIBUTION:
CDR, MEPCOM
CDR, USAAC
(CONT)

APPENDIX B

TEST BLUEPRINT FOR TAPAS VERSION V9

```
"InitialLinkedItemTypes": [  
  "Index": 1,  
  "Dimension1": "Cooperation",  
  "Dimension2": "Achievement",  
  "Id": 7917  
  "Index": 2,  
  "Dimension1": "Achievement",  
  "Dimension2": "Responsibility",  
  "Id": 7918  
  "Index": 3,  
  "Dimension1": "Responsibility",  
  "Dimension2": "Even Tempered",  
  "Id": 7919  
  "Index": 4,  
  "Dimension1": "Even Tempered",  
  "Dimension2": "Order",  
  "Id": 7920  
  "Index": 5,  
  "Dimension1": "Order",  
  "Dimension2": "Physical Condition",  
  "Id": 7921  
  "Index": 6,  
  "Dimension1": "Physical Condition",  
  "Dimension2": "Optimism",  
  "Id": 7922  
  "Index": 7,  
  "Dimension1": "Optimism",  
  "Dimension2": "Tolerance",  
  "Id": 7923  
  "Index": 8,  
  "Dimension1": "Tolerance",  
  "Dimension2": "Dominance",  
  "Id": 7924  
  "Index": 9,  
  "Dimension1": "Dominance",  
  "Dimension2": "Selflessness",  
  "Id": 7925  
  "Index": 10,  
  "Dimension1": "Selflessness",  
  "Dimension2": "Intellectual Efficiency",  
  "Id": 7926  
  "Index": 11,
```

"Dimension1": "Intellectual Efficiency",
 "Dimension2": "Sociability",
 "Id": 7927
 "Index": 12,
 "Dimension1": "Sociability",
 "Dimension2": "Cooperation",
 "Id": 7928
 "Index": 13,
 "Dimension1": "Sociability",
 "Dimension2": "Sociability",
 "Id": 7929
 "Index": 14,
 "Dimension1": "Commitment to Serve",
 "Dimension2": "Commitment to Serve",
 "Id": 7930
 "MainItemTypes": [
 "Index": 1,
 "Dimension1": "Commitment to Serve",
 "Dimension2": "Commitment to Serve",
 "Id": 7931
 "Index": 2,
 "Dimension1": "Commitment to Serve",
 "Dimension2": "Commitment to Serve",
 "Id": 7932
 "Index": 3,
 "Dimension1": "Commitment to Serve",
 "Dimension2": "Commitment to Serve",
 "Id": 7933
 "Index": 4,
 "Dimension1": "Commitment to Serve",
 "Dimension2": "Commitment to Serve",
 "Id": 7934
 "Index": 5,
 "Dimension1": "Commitment to Serve",
 "Dimension2": "Commitment to Serve",
 "Id": 7935
 "Index": 6,
 "Dimension1": "Commitment to Serve",
 "Dimension2": "Commitment to Serve",
 "Id": 7936
 "Index": 7,
 "Dimension1": "Commitment to Serve",
 "Dimension2": "Commitment to Serve",
 "Id": 7937
 "Index": 8,
 "Dimension1": "Sociability",

"Dimension2": "Sociability",
 "Id": 7938
 "Index": 9,
 "Dimension1": "Cooperation",
 "Dimension2": "Cooperation",
 "Id": 7939
 "Index": 10,
 "Dimension1": "Cooperation",
 "Dimension2": "Cooperation",
 "Id": 7940
 "Index": 11,
 "Dimension1": "Achievement",
 "Dimension2": "Achievement",
 "Id": 7941
 "Index": 12,
 "Dimension1": "Achievement",
 "Dimension2": "Achievement",
 "Id": 7942
 "Index": 13,
 "Dimension1": "Responsibility",
 "Dimension2": "Responsibility",
 "Id": 7943
 "Index": 14,
 "Dimension1": "Responsibility",
 "Dimension2": "Responsibility",
 "Id": 7944
 "Index": 15,
 "Dimension1": "Even Tempered",
 "Dimension2": "Even Tempered",
 "Id": 7945
 "Index": 16,
 "Dimension1": "Even Tempered",
 "Dimension2": "Even Tempered",
 "Id": 7946
 "Index": 17,
 "Dimension1": "Order",
 "Dimension2": "Order",
 "Id": 7947
 "Index": 18,
 "Dimension1": "Order",
 "Dimension2": "Order",
 "Id": 7948
 "Index": 19,
 "Dimension1": "Physical Condition",
 "Dimension2": "Physical Condition",
 "Id": 7949

"Index": 20,
 "Dimension1": "Physical Condition",
 "Dimension2": "Physical Condition",
 "Id": 7950
 "Index": 21,
 "Dimension1": "Optimism",
 "Dimension2": "Optimism",
 "Id": 7951
 "Index": 22,
 "Dimension1": "Optimism",
 "Dimension2": "Optimism",
 "Id": 7952
 "Index": 23,
 "Dimension1": "Tolerance",
 "Dimension2": "Tolerance",
 "Id": 7953
 "Index": 24,
 "Dimension1": "Tolerance",
 "Dimension2": "Tolerance",
 "Id": 7954
 "Index": 25,
 "Dimension1": "Dominance",
 "Dimension2": "Dominance",
 "Id": 7955
 "Index": 26,
 "Dimension1": "Dominance",
 "Dimension2": "Dominance",
 "Id": 7956
 "Index": 27,
 "Dimension1": "Selflessness",
 "Dimension2": "Selflessness",
 "Id": 7957
 "Index": 28,
 "Dimension1": "Selflessness",
 "Dimension2": "Selflessness",
 "Id": 7958
 "Index": 29,
 "Dimension1": "Intellectual Efficiency",
 "Dimension2": "Intellectual Efficiency",
 "Id": 7959
 "Index": 30,
 "Dimension1": "Intellectual Efficiency",
 "Dimension2": "Intellectual Efficiency",
 "Id": 7960
 "Index": 31,
 "Dimension1": "Cooperation",

"Dimension2": "Even Tempered",
 "Id": 7961
 "Index": 32,
 "Dimension1": "Cooperation",
 "Dimension2": "Order",
 "Id": 7962
 "Index": 33,
 "Dimension1": "Cooperation",
 "Dimension2": "Physical Condition",
 "Id": 7963
 "Index": 34,
 "Dimension1": "Cooperation",
 "Dimension2": "Optimism",
 "Id": 7964
 "Index": 35,
 "Dimension1": "Cooperation",
 "Dimension2": "Tolerance",
 "Id": 7965
 "Index": 36,
 "Dimension1": "Cooperation",
 "Dimension2": "Dominance",
 "Id": 7966
 "Index": 37,
 "Dimension1": "Cooperation",
 "Dimension2": "Selflessness",
 "Id": 7967
 "Index": 38,
 "Dimension1": "Cooperation",
 "Dimension2": "Intellectual Efficiency",
 "Id": 7968
 "Index": 39,
 "Dimension1": "Achievement",
 "Dimension2": "Even Tempered",
 "Id": 7969
 "Index": 40,
 "Dimension1": "Achievement",
 "Dimension2": "Order",
 "Id": 7970
 "Index": 41,
 "Dimension1": "Achievement",
 "Dimension2": "Physical Condition",
 "Id": 7971
 "Index": 42,
 "Dimension1": "Achievement",
 "Dimension2": "Optimism",
 "Id": 7972

"Index": 43,
 "Dimension1": "Achievement",
 "Dimension2": "Tolerance",
 "Id": 7973
 "Index": 44,
 "Dimension1": "Achievement",
 "Dimension2": "Dominance",
 "Id": 7974
 "Index": 45,
 "Dimension1": "Achievement",
 "Dimension2": "Selflessness",
 "Id": 7975
 "Index": 46,
 "Dimension1": "Achievement",
 "Dimension2": "Intellectual Efficiency",
 "Id": 7976
 "Index": 47,
 "Dimension1": "Achievement",
 "Dimension2": "Sociability",
 "Id": 7977
 "Index": 48,
 "Dimension1": "Responsibility",
 "Dimension2": "Order",
 "Id": 7978
 "Index": 49,
 "Dimension1": "Responsibility",
 "Dimension2": "Physical Condition",
 "Id": 7979
 "Index": 50,
 "Dimension1": "Responsibility",
 "Dimension2": "Optimism",
 "Id": 7980
 "Index": 51,
 "Dimension1": "Responsibility",
 "Dimension2": "Tolerance",
 "Id": 7981
 "Index": 52,
 "Dimension1": "Responsibility",
 "Dimension2": "Dominance",
 "Id": 7982
 "Index": 53,
 "Dimension1": "Responsibility",
 "Dimension2": "Selflessness",
 "Id": 7983
 "Index": 54,
 "Dimension1": "Responsibility",

"Dimension2": "Intellectual Efficiency",
 "Id": 7984
 "Index": 55,
 "Dimension1": "Responsibility",
 "Dimension2": "Sociability",
 "Id": 7985
 "Index": 56,
 "Dimension1": "Even Tempered",
 "Dimension2": "Physical Condition",
 "Id": 7986
 "Index": 57,
 "Dimension1": "Even Tempered",
 "Dimension2": "Optimism",
 "Id": 7987
 "Index": 58,
 "Dimension1": "Even Tempered",
 "Dimension2": "Tolerance",
 "Id": 7988
 "Index": 59,
 "Dimension1": "Even Tempered",
 "Dimension2": "Dominance",
 "Id": 7989
 "Index": 60,
 "Dimension1": "Even Tempered",
 "Dimension2": "Selflessness",
 "Id": 7990
 "Index": 61,
 "Dimension1": "Even Tempered",
 "Dimension2": "Intellectual Efficiency",
 "Id": 7991
 "Index": 62,
 "Dimension1": "Even Tempered",
 "Dimension2": "Sociability",
 "Id": 7992
 "Index": 63,
 "Dimension1": "Order",
 "Dimension2": "Optimism",
 "Id": 7993
 "Index": 64,
 "Dimension1": "Order",
 "Dimension2": "Tolerance",
 "Id": 7994
 "Index": 65,
 "Dimension1": "Order",
 "Dimension2": "Dominance",
 "Id": 7995

"Index": 66,
 "Dimension1": "Order",
 "Dimension2": "Selflessness",
 "Id": 7996
 "Index": 67,
 "Dimension1": "Order",
 "Dimension2": "Intellectual Efficiency",
 "Id": 7997
 "Index": 68,
 "Dimension1": "Order",
 "Dimension2": "Sociability",
 "Id": 7998
 "Index": 69,
 "Dimension1": "Physical Condition",
 "Dimension2": "Tolerance",
 "Id": 7999
 "Index": 70,
 "Dimension1": "Physical Condition",
 "Dimension2": "Dominance",
 "Id": 8000
 "Index": 71,
 "Dimension1": "Physical Condition",
 "Dimension2": "Selflessness",
 "Id": 8001
 "Index": 72,
 "Dimension1": "Physical Condition",
 "Dimension2": "Intellectual Efficiency",
 "Id": 8002
 "Index": 73,
 "Dimension1": "Physical Condition",
 "Dimension2": "Sociability",
 "Id": 8003
 "Index": 74,
 "Dimension1": "Optimism",
 "Dimension2": "Dominance",
 "Id": 8004
 "Index": 75,
 "Dimension1": "Optimism",
 "Dimension2": "Selflessness",
 "Id": 8005
 "Index": 76,
 "Dimension1": "Optimism",
 "Dimension2": "Intellectual Efficiency",
 "Id": 8006
 "Index": 77,
 "Dimension1": "Optimism",

"Dimension2": "Sociability",
 "Id": 8007
 "Index": 78,
 "Dimension1": "Tolerance",
 "Dimension2": "Selflessness",
 "Id": 8008
 "Index": 79,
 "Dimension1": "Tolerance",
 "Dimension2": "Intellectual Efficiency",
 "Id": 8009
 "Index": 80,
 "Dimension1": "Tolerance",
 "Dimension2": "Sociability",
 "Id": 8010
 "Index": 81,
 "Dimension1": "Dominance",
 "Dimension2": "Intellectual Efficiency",
 "Id": 8011
 "Index": 82,
 "Dimension1": "Dominance",
 "Dimension2": "Cooperation",
 "Id": 8012
 "Index": 83,
 "Dimension1": "Selflessness",
 "Dimension2": "Sociability",
 "Id": 8013
 "Index": 84,
 "Dimension1": "Achievement",
 "Dimension2": "Sociability",
 "Id": 8014
 "Index": 85,
 "Dimension1": "Cooperation",
 "Dimension2": "Dominance",
 "Id": 8015
 "Index": 86,
 "Dimension1": "Responsibility",
 "Dimension2": "Sociability",
 "Id": 8016
 "Index": 87,
 "Dimension1": "Even Tempered",
 "Dimension2": "Physical Condition",
 "Id": 8017
 "Index": 88,
 "Dimension1": "Order",
 "Dimension2": "Optimism",
 "Id": 8018

"Index": 89,
 "Dimension1": "Physical Condition",
 "Dimension2": "Order",
 "Id": 8019
 "Index": 90,
 "Dimension1": "Optimism",
 "Dimension2": "Selflessness",
 "Id": 8020
 "Index": 91,
 "Dimension1": "Tolerance",
 "Dimension2": "Achievement",
 "Id": 8021
 "Index": 92,
 "Dimension1": "Dominance",
 "Dimension2": "Tolerance",
 "Id": 8022
 "Index": 93,
 "Dimension1": "Selflessness",
 "Dimension2": "Even Tempered",
 "Id": 8023
 "Index": 94,
 "Dimension1": "Intellectual Efficiency",
 "Dimension2": "Responsibility",
 "Id": 8024
 "Index": 95,
 "Dimension1": "Cooperation",
 "Dimension2": "Intellectual Efficiency",
 "Id": 8025
 "Index": 96,
 "Dimension1": "Achievement",
 "Dimension2": "Selflessness",
 "Id": 8026
 "Index": 97,
 "Dimension1": "Even Tempered",
 "Dimension2": "Sociability",
 "Id": 8027
 "Index": 98,
 "Dimension1": "Order",
 "Dimension2": "Intellectual Efficiency",
 "Id": 8028
 "Index": 99,
 "Dimension1": "Physical Condition",
 "Dimension2": "Tolerance",
 "Id": 8029
 "Index": 100,
 "Dimension1": "Optimism",

"Dimension2": "Responsibility",
"Id": 8030
"Index": 101,
"Dimension1": "Tolerance",
"Dimension2": "Order",
"Id": 8031
"Index": 102,
"Dimension1": "Dominance",
"Dimension2": "Even Tempered",
"Id": 8032
"Index": 103,
"Dimension1": "Selflessness",
"Dimension2": "Dominance",
"Id": 8033
"Index": 104,
"Dimension1": "Intellectual Efficiency",
"Dimension2": "Achievement",
"Id": 8034
"Index": 105,
"Dimension1": "Sociability",
"Dimension2": "Optimism",
"Id": 8035
"Index": 106,
"Dimension1": "Sociability",
"Dimension2": "Physical Condition",
"Id": 8036

APPENDIX C

EXAMPLE OF PRE-TEST FORM FOR ESTIMATING GGUM PARAMETERS OF TAPAS STATEMENTS

Instructions:

This section of the questionnaire asks you to respond to a series of statements describing how you typically think, feel, or act. It is very important that you **respond to the statements honestly**.

Read each statement carefully and decide the extent to which you agree or disagree. Then accurately fill in the corresponding oval on the scantron form. **Please do not write or mark on this questionnaire** – just indicate your answers on the scantron.

Work at a fairly rapid pace. **And, remember that you are to answer honestly.**

Sample Item

		STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
1.	I enjoy being part of a team.	a	b	c	d

Mark “a” on your scantron – if you strongly disagree with the statement

Mark “b” on your scantron – if you disagree with the statement

Mark “c” on your scantron – if you agree with the statement

Mark “d” on your scantron – if you strongly agree with the statement

Please Note:

- There are no right and wrong answers. Just respond to the items honestly and accurately.
- In choosing an answer, consider your life in general and not only the last few weeks or months.
- Some items may be difficult to answer. In those cases, just think a bit longer and choose the answer that best describes you.
- Some items will appear similar. This is not designed to trick you, so there’s no need to look back at your previous answers. Just continue moving forward, answering the items honestly and accurately.

- Several items will ask you to mark a specific answer on your scantron. Your answers are used to check that our scanning software is working properly. Please make sure to mark the requested oval.

Please indicate your answer to each item by marking the appropriate oval on the scantron. Remember, it is important that you **answer honestly and accurately**.

		STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
1.	I usually make a noticeable contribution to group problem-solving tasks.	a	b	c	d
2.	I am generally pretty forgiving.	a	b	c	d

APPENDIX D

EXAMPLE OF PRE-TEST FORM FOR ESTIMATING SOCIAL DESIRABILITY PARAMETERS OF TAPAS STATEMENTS

Instructions For the Remaining Section

Unlike in the previous sections where you were instructed to answer items as honestly and accurately as possible, we now want you to **PRETEND** that you are not yet in the Army, but very much want to be. Imagine that a recruiter has asked you to complete a test to determine if you are **GOOD ARMY MATERIAL**. If you score well, you will be let into the Army. If you don't score well, you won't.

For all remaining sections, we want you to answer items in a way that will make you look good from the Army's standpoint. In other words, answer in a way that will give the Army the best possible impression of you to insure that you pass the test and get accepted. Convince the Army that you will make a good Soldier!

Instructions:

This section of the questionnaire asks you to respond to a series of statements describing how one might think, feel, or act. The format is the same as one you saw earlier.

Remember, you are now trying to **create the best possible impression of yourself** from the **Army's standpoint**; so, your answers do not need to describe you accurately. Just answer in a way that you think will maximize your chances of getting accepted into the Army.

Please do not write or mark on this questionnaire – just indicate your answers on the scantron.

Work at a fairly rapid pace. **And, convince the Army that you will make a good Soldier!**

Sample Item:

		STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
1.	I enjoy being part of a team.	a	b	c	d

Mark “a” on your scantron – if you think “Strongly Disagree” makes you like good Army material

Mark “b” on your scantron – if you think “Disagree” makes you like good Army material

Mark “c” on your scantron – if you think “Agree” makes you like good Army material

Mark “d” on your scantron – if you think “Strongly Agree” makes you like good Army material

Please Note:

- Some items may be difficult to answer. In those cases, just think a bit longer and choose the answer that best serves your “goal” of getting into the Army.
- Some items will seem similar. This is not designed to trick you, so there’s no need to look back at your previous answers. Just continue moving forward, answering in a way that makes you look like good Army material.

Please indicate your answer to each item by marking the appropriate oval on the scantron.
Remember, **answer in a way that makes you like good Army material!**

	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
1. I have great respect for our legal system and support it in any way I can.	a	b	c	d
2. In group projects, I give personal and team goals equal weight and consideration.	a	b	c	d

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

%	Percent
2AFC	Two-Alternative Forced-Choice
α_i	IRT Discrimination Parameter
δ_i	IRT Location Parameter
τ_{il}	IRT Threshold Parameter
ℓ_z	Unusual Response Flag
σ_E^2	Error Variance
σ_X^2	Observed Score Variance
ρ	Reliability
θ	Theta; A Person's Standing on a Facet
$\hat{\theta}$	Theta-hat; Estimate of θ ; Facet Raw Score
ABLE	Assessment of Background and Life Experiences
APFT	Army Physical Fitness Test
AFQT	Armed Forces Qualification Test
AFSCs	Air Force Specialty Codes
AIM	Assessment of Individual Motivation
ALQ	Army Life Questionnaire
APA	American Psychological Association
ARI	Army Research Institute
ARSOF	Army Special Operations Forces
ASVAB	Armed Services Vocational Aptitude Battery
B-W	Black-White
CAT	Computer-Adaptive Test
CTT	Classical Test Theory
d	Cohen's d ; Standardized Mean Score Difference
D	Dimension (e.g., 3-D, 3-dimensional)
DoD	Department of Defense
DS	Drill Sergeant
DCG	Drasgow Consulting Group
EEEM	Expanded Enlistment Eligibility Metrics
F-M	Female-Male
GGUM	Generalized Graded Unfolding Model
H-WNH	Hispanic-White Non-Hispanic
IIFs	Item Information Functions
IRB	Institutional Review Board
IRFs	Item Response Functions
IRT	Item Response Theory
MAGE	Mechanical, Administrative, General or Electronic
MEPS	Military Entrance Processing Stations
MOS	Military Occupational Specialty
MTurk	Mechanical Turk
MUPP	Multi-unidimensional Pairwise Preference

NCO	Non-Commissioned Officer
NSAB	Noncommissioned Officer Special Assignment Battery
NSMRL	Naval Submarine Medical Research Lab
OCS	Officer Candidate School
PSM	Predictive Success Model
<i>R</i>	Multiple Correlation
RA	Regular Army
RBI	Rational Biographical Inventory
ROTC	Reserve Officer Training Corps
SBIR	Small Business Innovation Research
SDs	Standard Deviations
SE	Standard Error
TAPAS	Tailored Adaptive Personality Assessment System
USMC	United States Marine Corps
V	Version
WNH	White Non-Hispanic