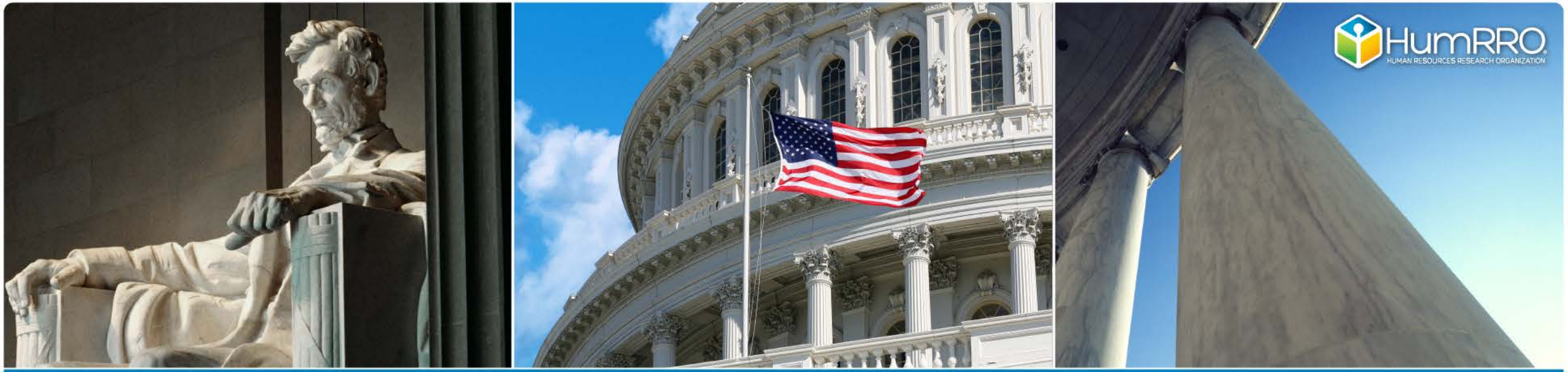

TAPAS Evaluation Project: Results and Way Forward

Presented to: Defense Advisory Committee on Military Personnel Testing (DACMPT)

Presenters: Tim McGonigle, HumRRO

September 17, 2020

TEP Process and Objectives

- DPAC contracted with HumRRO to independently review the body of TAPAS research and make recommendations regarding the readiness of TAPAS for operational use in selection, classification, or other decision making.
- The evaluators had two objectives:
  - Recommendations on readiness of TAPAS for operational use/implementation
    - Operational use/implementation indicates policy/issuance directing permanent implementation
  - Recommendations on future research and development, to include operational analyses/pilots
- HumRRO's method for conducting the review involved assembling a team of nationally known experts in psychometrics, personality theory and measurement, and operational testing:
  - Dr. James Robert (Chair), Georgia Tech
  - Dr. Winfred Arthur, Texas A&M University
  - Dr. Mark Reckase, Michigan State University
  - Dr. Paul Sackett, University of Minnesota
  - Dr. April Zenisky, University of Massachusetts

Key Points

- Four meetings (Oct. 2018–July 2019)
  - Attended by DPAC and Service representatives
  - Presentations from TAPAS developers, RAND, and Service representatives
- Report delivered October 2019
  - Organized around *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014)

TEP Process and Objectives (cont.)

- Organized findings using the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) as a framework. The *Standards* provide
  - "… criteria for the development and evaluation of tests and testing practices and provide guidelines for assessing the validity of interpretations of test scores for the intended test uses," (p. 1) and
  - guidance about current best professional practice in demonstrating the quality of tests.
- Section evaluation (i.e., satisfactory, minimally sufficient, insufficient)
  - Multi-Unidimensional Pairwise Preference (MUPP) Model
  - Scores, Scales, Norms, Score Linking, and Cut Scores
  - Reliability
  - Validity
  - Mitigating Social Desirability
  - Fairness and Subgroup Differences
  - Test Design and Development and Documentation
  - General Recommendations

TEP Process and Objectives (cont.)

- Scope of review: While the TEP team considered evidence from the entire TAPAS system and all listed uses (i.e., selection, classification, and special assignment), the scope of this evaluation was limited to the operational use of TAPAS for selection decisions.
- Conceived of TAPAS as
  - a specific instantiation of a model developed by the Army with other users (Air Force, Navy, Marine Corps);
  - having a library of 27 facets with individual versions consisting of 13 to 15 of these facets; and
  - primarily evaluated for use in selection decisions to date, although other uses have received limited research.
- Identified 9 recommendations that should be addressed prior to operational use and 15 that can be addressed after operational use has begun
  - Assuming no adverse findings result from addressing the pre-operational recommendations, the TEP team believes TAPAS can be used operationally for selection purposes.
  - Post-implementation recommendations will improve the quality of information provided by TAPAS.
- **AP, DPAC, and Services commented on the report and provided their recommendations/ intentions for way forward.**
- **Concentration is on the pre-implementation recommendations.**

MUPP Model: **Minimally sufficient** evidence

- The MUPP model is based on the idea that choices are determined by the degree to which each statement in the pair best represents the respondent. TAPAS is designed to prevent faking by making the “correct” response not readily apparent because the statements are matched on strength of association with a single dimension and on social desirability.
- Found sufficient reasoning for the use of the MUPP model, but also identified 3 pre-implementation and 7 post-implementation recommendations:
  - **Pre-implementation recommendations**
    1. Test the proportionately redistributed probabilities assumptions when respondents agree or disagree with both statements in an item
    2. Provide evidence that an unfolding model is an appropriate choice for disagree-agree responses to statements in the TAPAS item pool, including examination of principal components to single-stimulus responses along with generation of empirical item characteristic curves
    3. Provide technical documentation about the computerized adaptive testing algorithm
  - **Post-implementation recommendations**
    1. Incorporate a population variance-covariance matrix estimate for TAPAS facets into the procedure used to estimate TAPAS theta values
    2. Recalibrate the Generalized Graded Unfolding Model (GGUM) item parameter estimates for each facet and the corresponding social desirability estimates using a larger dataset consisting of more recent data; monitor parameters for drift from version to version

MUPP Model: **Minimally sufficient** evidence (cont.)

- Found sufficient reasoning for the use of the MUPP model, but also identified 3 pre-implementation and 7 post-implementation recommendations (cont.):
  - **Post-implementation recommendations**
    3. Investigate the possibility of estimating TAPAS item parameters from the forced-choice responses rather than relying on those developed from responses to single statements using the GGUM
    4. Demonstrate algebraically, and document, any dependence of the TAPAS latent metric on unidimensional statement pairings
    5. Reanalyze simulation data using a Root Mean Squared Deviation (RMSD) index between estimated and true parameters
    6. Document item selection procedures to show that the precision of facet score estimation for core facets is psychometrically sufficient given that the number of statements per facet varies by TAPAS version
    7. Calculate and document the item pool information function for each facet of the TAPAS using the information function
- **Way forward**
  - **ARI will lead the effort for addressing the pre-implementation recommendations, using data from across the Services.**

Scores/Scales/Norms/Score Linking/Cut Scores: **Insufficient** evidence

- Versions 9, 10, and 11 are used interchangeably despite variation in facets between versions. Additional variations occur across Services, as does the use of composite scores versus facet scores. There is no single source of information about the scaling and comparability of TAPAS versions. Identified 4 pre-implementation recommendations:
  - **Pre-implementation recommendations**
    1. Demonstrate and document version equivalency, develop a comprehensive documentation on norming and equating, including a detailed description of the size and characteristics of the samples used to norm each version
    2. Investigate the comparability of facet and composite estimates across versions to determine the impact of changes in construct, by version (movement in and out of facets/statements)
    3. Document the process by which the cut scores on the facets and composites were derived; document argument and measurement precision near any cut scores that are established to make decisions using TAPAS results
    4. Provide clear guidance on score interpretation, including information about score meaning and score precision
- **Way forward**
  - **AFPC will take the lead in developing/collecting required documentation on norming and equating, to include impact on facets and composite scores. If additional analyses are required, data from across the Services will be utilized.**
  - **Services will develop clearly articulated process for cut score development, score interpretation, and process for score reports.**
  - **Stakeholders (AP, DPAC, and Services) will develop a centralized standardization process/policy on identification of facets, norming, equating, and version control. Effort for developing Theory of Action has been initiated.**

Reliability: **Minimally sufficient** evidence

- Because TAPAS uses an Item Response Theory (IRT) model for test design, it should be evaluated using indices of measurement precision (e.g., conditional standard errors and marginal reliability) rather than test-retest or alternate forms reliability alone.
  - Non-operational analyses reported good marginal reliability and sufficient test-retest correlations. Operational analyses show lower test-retest correlations, but may confound several sources of error (e.g., test items within and across test takers, testing conditions/retest after failure versus not, and inconsistent retest intervals).
- Preliminary results provide minimally sufficient evidence of measurement precision, but also identified 1 pre-implementation and 2 post-implementation recommendations:
  - **Pre-implementation recommendation**
    1. Calculate the marginal reliability and conditional standard errors for new item pools, along with information about the distribution of precision of estimates across a given sample (when developing new item pools)
  - **Post-implementation recommendation**
    1. Design and conduct an operational study to estimate the test-retest correlation of TAPAS facet and composite scores in a context where test administrations can be considered replications and incorporate conditional standard error and marginal reliability indices
    2. Calculate and document conditional standard errors of test scores based on Fisher information in addition to those from the replication method
- **Way forward**
  - **AFPC will lead the effort in addressing the pre-implementation recommendations, using data from across the Services.**

Validity: **Minimally sufficient** evidence

- TAPAS research has focused on adding prediction beyond ASVAB scores for a number of performance criteria (e.g., turnover, training performance, supervisor ratings) for selection. Evidence shows that TAPAS composite scores contribute small but consistent increases in prediction of attrition. The value of this increment must be determined by the Services themselves.
- Existing research evidence provides minimally sufficient evidence of incremental validity, but also identified 3 pre-implementation and 4 post-implementation recommendations:
  - **Pre-implementation recommendations**
    1. Develop and document a validity argument for each operational version and use of TAPAS, specifying the outcomes the test is intended to predict, the intended population, and evidence in support of intended interpretation
    2. Conduct operational pilots evaluating impacts on critical Service performance outcomes
    3. Provide documentation on construct validity at the composite level
  - **Post-implementation recommendations**
    1. If retests are permitted by policy, evaluate and demonstrate the psychometric properties and validity of the testing system, including policies about retest intervals, number of retests, and length of time for which scores are valid
    2. Calculate mean differences in scores by sex and ethnicity when sample sizes are sufficient for separate examination of validity by subgroup
    3. Document the extent to which external judgment has been utilized to evaluate the TAPAS item pools for content, construct, and sensitivity reasons
    4. Reduce the number of latent constructs to those that are absolutely essential (e.g., those weighted in composites)

Validity: **Minimally sufficient** evidence

- **Way forward**
  - **Services will continue to develop and propose additional analyses for validity assessments using the TEP team's recommended study designs. The pilots/analyses design documentation should include critical outcomes, the intended population, evidence in support of intended interpretation, and impact of any confounding artifacts.**
  - **Stakeholders will develop a comprehensive centralized technical infrastructure for summarizing all validation efforts to date. All applicable documentation must be included in the summary.**
  - **Services will provide existing validity documentation in support of current and proposed inferences to be included in the centralized infrastructure.**
  - **Establish a nomological network of the composites to provide additional evidence of construct validity at the composite level.**
    - ***Services had concerns on the appropriateness of this recommendation, since construct validity is established at the facet level.***
  - **Services will conduct validity analyses by sub-group.**
  - **Stakeholders will develop centralized standardization process/policy for including constructs/facets in future TAPAS versions. Theory of Action has been initiated, which will drive toward development of a centralized version.**
  - **Stakeholders will discuss feasibility of a centralized retest policy.**

Social Desirability/Fairness/Subgroup Diff: **Satisfactory evidence**

- Mitigating Social Desirability
  - Personality measures are notorious for response distortion because candidates attempt to determine what response is expected and present themselves in a way that they believe will make them a more attractive candidate.
  - TAPAS was designed to reduce the effectiveness of this strategy, such that item pairs consist of equally desirable or undesirable traits from different facets.
  - TAPAS utilizes a design that effectively mitigates social desirability and seems to be impermeable to coaching.
  - Recommend evaluation of the extent to which the social desirability responding parameter is fixed over time.
- Fairness and Subgroup Differences
  - Personality measures tend to show small to non-existent differences between sex, race, and ethnic groups.
  - TAPAS generally follows this pattern, but the effect size differences for the composites may be worthy of additional study depending on different uses among the Services.
  - Additionally, examination of validity by subgroup will further enhance the fairness evidence.

Test Design and Development/Documentation: **Insufficient** evidence

- The standards for test design and development advise that the intended uses of a test are known and clearly articulated and guide the test development process.
- TAPAS was intentionally designed to be flexible to support each Service's needs. As a result, there are aspects of TAPAS test design and development that are better explained than others.
- TAPAS also suffers from a diffusion of responsibility for the evidentiary arguments for the multiple known, desired, and/or reasonably anticipated uses of TAPAS within and across the Services.
- **Post-implementation recommendation**
  - The TEP team recommends that the publisher assemble a technical manual that specifies the approved test uses, the evidence for such uses, and the examinee population associated with each use/evidentiary argument. This technical manual should include information on psychometric test design, test development, scoring, reliability, validity, and fairness
- **Way forward**
  - **Services will continue assessing potential uses for TAPAS.**
  - **Services will clearly articulate and document purpose and objective for each use, with supporting validity information, prior to implementation.**
    - **Services must utilize MAPWG-established checklist/protocol for providing validity evidence.**
  - **AP, in collaboration with DPAC, will initiate development of a centralized process for documentation repository.**
    - **Services will provide all required documentation on psychometric test design, test development, scoring, reliability, validity, and fairness to be included in the repository.**
  - **Theory of Action has been initiated, which will drive development of the validity framework similar to the ASVAB validation framework.**

General Recommendations

- **Pre-implementation recommendation**
  - Develop an infrastructure for permanent operational testing (e.g., define and assign operational roles and responsibilities, create and document quality assurance procedures, develop standardized methods for continual development)
- **Post-implementation recommendation**
  - Investigate the comparability of samples and research findings for analysis within and across Services and at points in service (pre-accession and post-accession)
- **Proposed way forward**
  - **Stakeholders will develop a centralized process/policy for future TAPAS research, development, and maintenance. The plan should include all applicable roles, responsibilities, and funding streams. Items on identification of future constructs/facets, norming, equating, version control, and implementation policies must be covered in the plan. Theory of Action effort has been initiated.**
  - **AP, in collaboration with DPAC, will initiate development of a centralized process for documentation repository.**
    - **Services will provide all required documentation on psychometric test design, test development, scoring, reliability, validity, and fairness to be included in the repository.**

Conclusion and Way Forward

- Overall Conclusions
  - Available evidence provides preliminary support for the operational use of TAPAS for selection
  - Several items should be addressed, most notably
    - scores/scales, norms, score linking, and cut scores, and
    - test design and development and documentation.
  - As noted, some of these items should be addressed prior to permanent operational implementation, while others could be addressed over time
    - Assuming no adverse findings result from addressing the pre-operational recommendations, the TEP team believes TAPAS can be used operationally for selection purposes
    - Post-implementation recommendations will improve the quality of information provided by TAPAS
- Next Steps
  - Completion of additional analyses, collection of documentation, and completion of Theory of Action to drive future development of a centralized version

Questions?

Backup Slides

Background on TAPAS

- Tailored Adaptive Personality Assessment System (TAPAS)
  - Originally developed by ARI and Drasgow Consulting Group (DCG) to measure up to 27 facets of the Big Five personality dimensions
  - Uses multidimensional pairwise preference (MDPP) items
    - Generally presents two statements from different personality dimensions
    - Matched on the strength of the dimension and on the socially desirable nature of the response options
    - Intended to make it more difficult to fake because the "correct" answer is difficult to identify
  - Items generated on-the-fly by selecting from pools of pre-calibrated personality statements that measure construct dimensions relevant to performance in the military; approximately 1M statement combinations possible
  - Scored using multi-unidimensional pairwise preference IRT (ideal point) model
- Army, Navy, Air Force, and Marines have all collected TAPAS data on applicants
  - Evidence of incremental validity beyond ASVAB for training and military success criteria (e.g., attrition)
- Some stakeholders raised technical concerns about TAPAS, especially low test-retest reliability
  - RAND conducted an independent evaluation of the reliability and validity of TAPAS
    - Analyzed data from candidates who completed TAPAS between March 2010 and April 2015, and subsequently completed at least six months of service
    - Found small, significant incremental validity over education credential in predicting attrition
    - Found low test-retest reliability in some conditions
      - r_{xx} = 0.07 (TAPAS 9/10/11, Army recruits who failed first test)
    - But not as low under other conditions
      - r_{xx} = 0.59 (TAPAS 5/7/8, Air Force all recruits)

Which of these statements is the most like you?

- People come to me when they want fresh ideas
- Most people would say I am a "good listener"

Facets associated with TAPAS versions 9, 10, and 11

TAPAS Facet	Version		
	9	10	11
Achievement	x	x	x
Attention Seeking			x
Commitment to Serve	x		x
Cooperation	x		
Courage		x	
Dominance	x	x	x
Even-Tempered	x	x	x
Intellectual Efficiency	x	x	x
Non-Delinquency		x	
Optimism	x	x	x
Order	x	x	x
Physical Conditioning	x	x	x
Responsibility	x		
Selflessness	x	x	x
Situational Awareness		x	
Sociability	x	x	x
Team Orientation			x
Tolerance	x	x	x

Note: Facet appearing on all three versions are bolded.

Documents Reviewed by TEP Team

Technical Reports and Publications

- *An Evaluation of the Tailored Adaptive Personality Assessment System: Is It Valid for Predicting Attrition from Military Service? Is It Reliable?* (Hanser, Hardison, & Agniel, 2018)
- *Constructing Fake-Resistant Personality Tests Using Item Response Theory* (Stark, Chernyshenko, & Drasgow, 2011)
- *Tier One Performance Screen Initial Operational Test and Evaluation: 2015–2016 Biennial Report* (Knapp & Kirkendall, 2018)
- *Adaptive Testing with Multidimensional Pairwise Preference Items: Improving the Efficiency of Personality and Other Noncognitive Assessments* (Stark, Chernyshenko, Drasgow, & White, 2012)
- *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to Support Army Selection and Classification Decisions* (Drasgow, Stark, Chernyshenko, Nye, Hulin, & White, 2012)
- *Moderators of the Tailored Adaptive Personality Assessment System Validity* (Stark, Chernyshenko, Nye, Drasgow, & White, 2017)
- *Assessing the Tailored Adaptive Personality Assessment System (TAPAS) as a MOS Qualification Instrument* (Nye, Drasgow, Chernyshenko, Stark, Kubisiak, White, & Jose, 2012)
- *An IRT Approach to Constructing and Scoring Pairwise Preference Items Involving Stimuli on Different Dimensions: The Multi-Unidimensional Pairwise-Preference Model* (Stark, Chernyshenko, & Drasgow, 2005)

Documents Reviewed by TEP Team

Technical Reports and Publications (cont.)

- *Validation of the Noncommissioned Officer Special Assignment Battery* (Horgen, Nye, White, LaPort, Hoffman, Drasgow, Chernyshenko, Stark, & Conway, 2013)
- *Constructing Personality Scales Under the Assumptions of an Ideal Point Response Process: Toward Increasing the Flexibility of Personality Measures* (Chernyshenko, Stark, Drasgow, & Roberts, 2007)
- *Toward a New Attrition Screening Paradigm: Latest Army Advances* (White, Rumsey, Mullins, Nye, & LaPort, 2014)
- *From ABLE to TAPAS: A New Generation of Personality Tests to Support Military Selection and Classification Decisions* (Stark, Chernyshenko, Drasgow, Nye, White, Heffner, & Farmer, 2014)
- *Assessing the Tailored Adaptive Personality Assessment System for Army Special Operations Forces Personnel* (Nye, Beal, Drasgow, Dressel, White, & Stark, 2014)
- *Personality Assessment Questionnaire as a Pre-Accession Screen for Risk of Mental Disorders and Early Attrition in U.S. Army Recruits* (Niebuhr, Gubata, Oetting, Weber, Feng, & Cowan, 2013)
- *Examining Personality for the Selection and Classification of Soldiers: Validity and Differential Validity Across Jobs* (Nye, White, Drasgow, Prasad, Chernyshenko, & Stark, in press)
- *Tailored Adaptive Personality Assessment System (TAPAS) as an Indicator for Counterproductive Work Behavior: Comparing Validity in Applicant, Honest, and Directed Faking Conditions* (Trent, Barron, Rose, & Carretta, 2018)

Documents Reviewed by TEP Team

Presentations

- *Development of the Tailored Adaptive Personality Assessment System (TAPAS) and Ongoing Psychometric Research* (Stark, Drasgow, Chernyshenko, Nye, Heffner, & White)
- *Validity Evidence for the Tailored Adaptive Personality Assessment System* (Nye & White)
- *Stability and Validity of TAPAS Under Operational and Experimental Conditions* (Trent)
- *Evaluating the Usefulness of TAPAS: Reliability and Validity Results* (Hanser, Hardison, & Agniel)
- *TAPAS Research Review: Validity, Reliability, Demographic, and Faking Subgroup Differences* (Kantrowitz)
- *The Tailored Adaptive Personality Assessment System (TAPAS): Reliability and Validity* (White, Nye, & McMillan)
- *Use of the TAPAS in the U.S. Air Force* (Trent)
- *USMC Use of TAPAS* (Gonzales)
- *Navy's Operational Use of the Tailored Adaptive Personality Assessment System* (Keiser)
- *Operational Use of Tailored Adaptive Personality Assessment (TAPAS) in the U.S. Army* (Heffner)