TAPAS Validity Framework and Joint-Enlistment Composite

Deirdre J. Knapp and Dan J. Putka
*Human Resources Research Organization
(HumRRO)*

Briefing presented to the DACMPT
August 16, 2023

23-S-2673

Briefing Agenda

- Introduction to TAPAS
- TAPAS Validity Argument Framework
- Development of a Joint-Service Selection Composite
- Generation of Additional Validity Evidence

Introduction to TAPAS

OPA
OFFICE OF PEOPLE ANALYTICS

Tailored Adaptive Personality Assessment System (TAPAS)

- Developed by Drasgow Consulting Group (DCG) under a Small Business Innovation Research grant with the US Army Research Institute

- Promising research led to Army use of TAPAS to support enlistment selection decisions

- Findings prompted other Services to initiate their own TAPAS research programs

- RAND report identified some technical concerns which prompted an independent TAPAS Evaluation Project (TEP) review that was briefed to DAC in September 2020

- The TEP recommended establishment of a theory of action validity argument framework for TAPAS

TAPAS Description

- DoD-owned statement pools for 27 personality facets, with 13–15 facets typically included on a given TAPAS version

- Uses multidimensional pairwise preference (MDPP) items

- Most items present two statements from different personality dimensions

- Statements matched on the strength of the dimension and on the socially desirable nature of the response options

- Items generated on-the-fly by selecting from pools of pre-calibrated personality statements that measure construct dimensions relevant to performance in the military

- Scored using multi-unidimensional pairwise preference IRT (ideal point) model

**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."

OPA
OFFICE OF PEOPLE ANALYTICS

TAPAS Validity Argument Framework

Purpose of Validity Argument Frameworks

- Compile, organize, and review existing evidence related to the use of assessments
  - Relevant information defined much more broadly than psychometric properties or criterion-related evidence
  - Includes all aspects of a measure's design, development, administration, score reporting, etc.

- Evaluate whether available evidence supports the use of the assessments for their intended purposes

- Identify ways to strengthen evidence supporting the use of the assessments

- Help inform improvements to the assessments in terms of content, scoring, administration, and/or interpretation

Validity Argument Framework Method
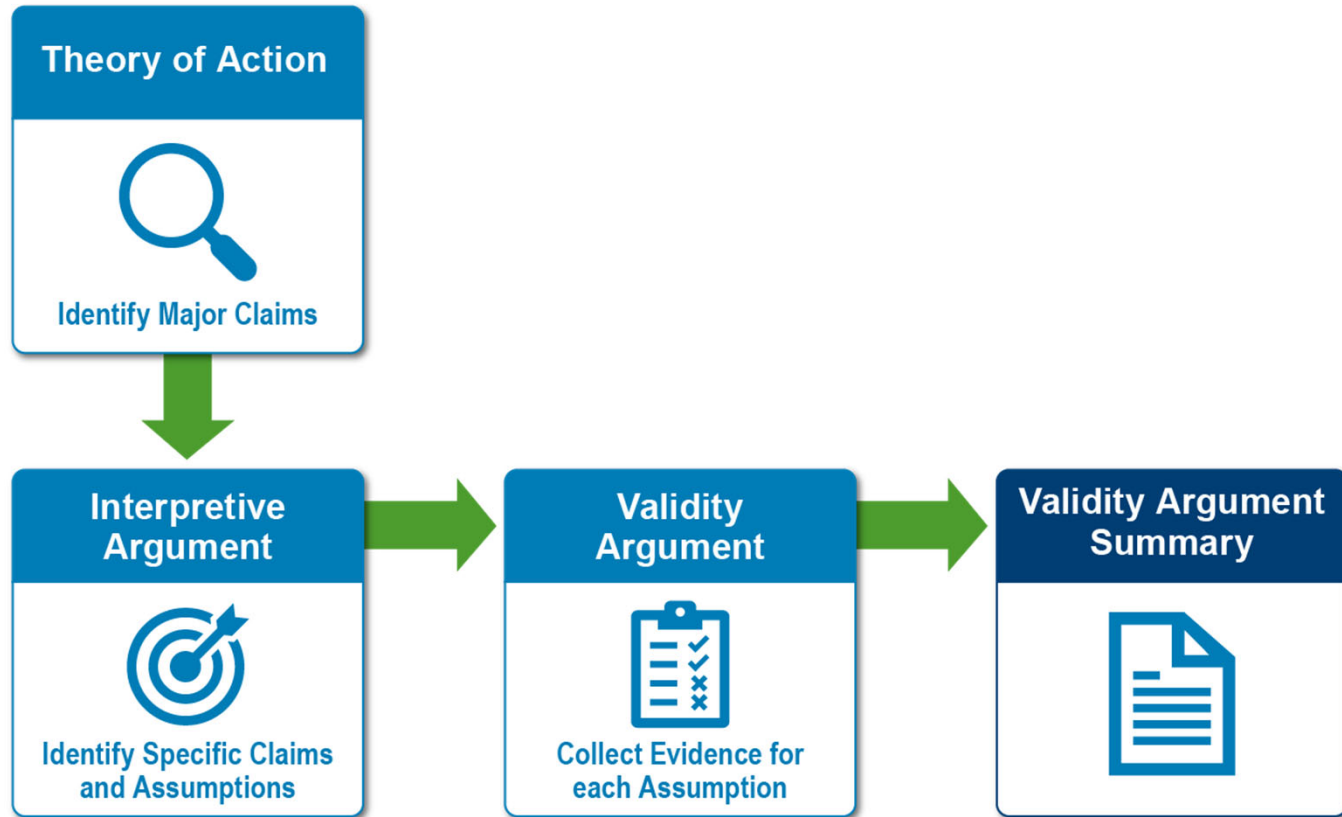
Overview of Pre-Enlistment Assessment Validity Arguments

- AFQT Validity Argument 2.0 (Knapp et al., in review)
- ASVAB Validity Argument 1.0 (Sinclair et al., 2023)
  - Both AFQT and ASVAB work briefed to DAC December 2022
- TAPAS Validity Argument 1.0 (Ford et al., 2022)
  - Covers both selection and classification
  - Not tied to a specific use case

Nature of this work is dynamic as validity argument evidence accumulates over time.
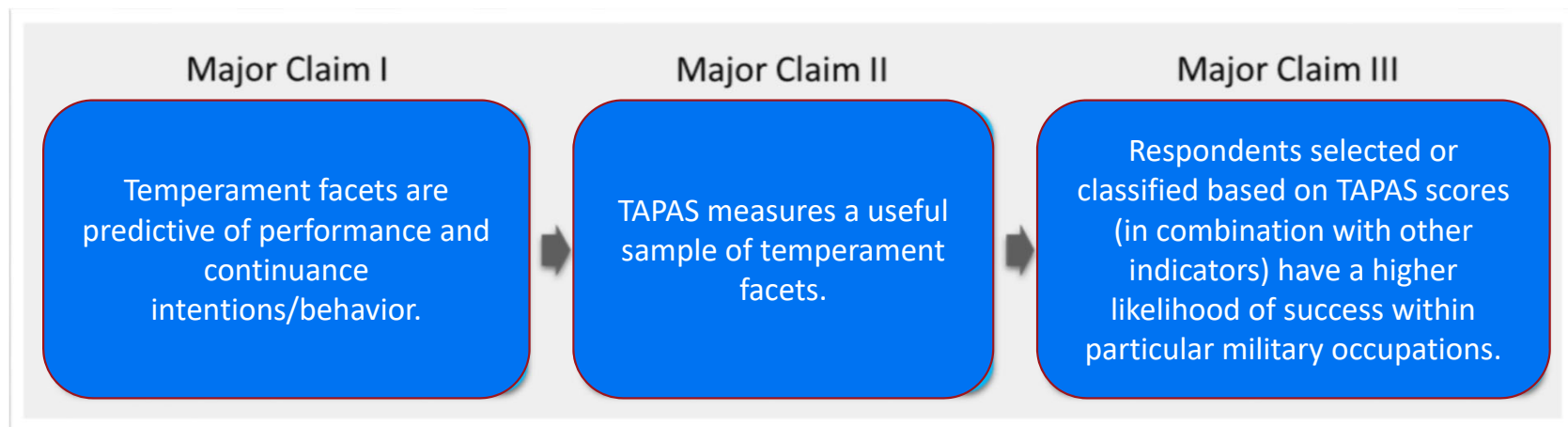
TAPAS Theory of Action

Major Claim I	Major Claim II	Major Claim III
Temperament facets are predictive of performance and continuance intentions/behavior.	TAPAS measures a useful sample of temperament facets.	Respondents selected or classified based on TAPAS scores (in combination with other indicators) have a higher likelihood of success within particular military occupations.

OPA
OFFICE OF PEOPLE ANALYTICS

Illustration of Specific Claims and Assumptions

Specific Claims and Assumptions Supporting "TAPAS Measures a Useful Sample of Temperament Facets"	
Major Claim	
II. TAPAS measures a useful sample of temperament facets.	
Specific Claims	
Selection	**Classification**
II.1. TAPAS measures temperament facets associated with work requirements across a broad range of military occupations.	II.2 TAPAS measures temperament facets associated with work requirements differentially across military occupations.
Assumptions	**Assumptions**
II.1.a. If TAPAS measures temperament facets associated with work requirements across a broad range of military occupations, then job analysis information across a broad sample of occupations should support the relevance of those facets. (Content relevance)	II.2.a. If TAPAS measures temperament facets associated with work requirements differentially across military occupations, then occupational analysis information should show differences in which facets are most relevant for different occupations. (Content relevance of facets)
II.1.b. If TAPAS measures temperament facets associated with work requirements across a broad range of military occupations, then the facets should cover a broad range of established temperament constructs.	

Summary of Interpretive Argument

- Three major claims
- 18 specific claims
  - 6 Selection
  - 7 Classification
  - 5 Selection and classification
- 47 assumptions
  - 14 Selection
  - 14 Classification
  - 19 Selection and classification
- *See handout for full interpretive argument*

TAPAS Validity Argument 1.0 Abbreviated Claim I Summary

Major Claim I: Temperament facets are predictive of performance and continuance intentions/behavior.

- "The preponderance of research evidence strongly supports the claim that temperament facets are predictive of performance and continuance intentions/behavior, particularly when temperament is used for selection purposes" (Ford et al., 2022, p. 2).

- "Overall, it appears that research evidence for classification uses of temperament is limited either by lack of accumulated evidence or lack of sufficient relationships" (p. 3).

- "As individual trait levels change across time, they change somewhat in tandem across people, thereby maintaining the trait-criterion correlation over time" (p. 3).

TAPAS Validity Argument 1.0 Abbreviated Claim II Summary

Major Claim II: TAPAS measures a useful sample of temperament facets.

- "Evidence in support of the assumptions associated with this broad claim is generally positive, but important information is not sufficiently documented to judge and much of the available evidence could be strengthened with additional research" (p. 3).

- "While there is evidence that TAPAS facets are relevant for selection, there is less job analysis evidence to suggest that personality measures will be useful for differentiating among occupations as would be needed to support classification decisions" (p. 3).

- "There is insufficient documentation to determine whether the TAPAS statement pools sufficiently cover the range of extremity and social desirability parameters to support reliable and accurate measurement, or to critique the computerized adaptive testing algorithm used for statement selection, pairing, and scoring. And, while there is some evidence of cross-format and cross-version score correspondence, there is insufficient evidence to more broadly conclude that scores are not affected by use of different statement pools and/or different sets of facets included on a given version of TAPAS" (p. 4).

- "Subgroup differences on TAPAS scores are generally minimal, especially compared to cognitive ability measures" (p. 4).

OPA
OFFICE OF PEOPLE ANALYTICS

TAPAS Validity Argument 1.0 Abbreviated Claim III Summary

Major Claim III: Respondents selected or classified based on TAPAS scores (in combination with other indicators) have a higher likelihood of success within particular military occupations.

- ". . . most of the research reviewed addresses notional rather than current or proposed operational selection and classification decision-making systems. More evidence would be needed to support specific operational applications of TAPAS scores" (p. 4).

- "Evidence suggests that multiple TAPAS facet and composite scores (other than Can-Do) tend to show incremental validity over AFQT, particularly for motivational and retention-related outcomes" (p. 5).

- "Available evidence from controlled studies suggests TAPAS scores display moderate to moderately high test-retest correlations, though results using ad hoc military retest samples are low. Additional carefully designed research on military samples is needed to assess the reliability of TAPAS scores more confidently throughout the score distribution, and particularly near proposed cut scores" (p. 5).

- "Available evidence shows that TAPAS facets and composites exhibit differing levels of criterion-related validity across occupations. Patterns of incremental validity over ASVAB aptitude area scores are similar to those observed with AFQT scores. There is no direct evidence, however, relevant to the efficacy of using a finite number of TAPAS composites to identify the types of occupations for which enlisted personnel would be most successful" (p. 5).

Validity Argument Report Recommendations

- Ideas for strengthening TAPAS validity evidence were organized as follows:

  1. Expand on and improve documentation
  2. Strengthen content and construct validity evidence
  3. Clarify and revisit suitability of the measurement model
  4. Strengthen development procedures
  5. Strengthen psychometric evidence
  6. Broaden criterion-related validity investigations
  7. Investigate classification efficiency as warranted
  8. Extend test fairness investigations
  9. Make administrative improvements

- TEP and validity argument report recommendations integrated to produce an updated R&D agenda (summer 2023)

Other Work in Progress

- Technical reports providing further documentation on TAPAS development and psychometric properties nearing completion

- Developing design recommendations for joint-service TAPAS

- Developing an interim joint-service selection composite
  - Objective is to widen the aperture for identifying qualified recruits without compromising outcomes important to the Services

- Developing a research plan to collect cross-service criterion-related validation data on the joint-service TAPAS selection composite

- Developing a strategy to evolve the TAPAS validity argument framework

Development of an Interim Joint-Service Composite

Background on a Joint-Service TAPAS Composite

- Objective is to develop a TAPAS composite that can be used to inform general enlisted selection and qualification decisions

- Composite would presumably complement other measures and data Services use when making such decisions during applicant screening, e.g.,
  - AFQT scores
  - Medical, physical, conduct-related data

- Focus is on building a composite that will be predictive of first-term enlisted job performance

Key Development Steps

- Step 1: Identify first-term enlisted performance dimensions
- Step 2: Capture "overall performance" policy
- Step 3: Define universe of potential TAPAS facets for interim composite
- Step 4: Establish interim composite development and validation strategies
- Step 5: Gather archival and SME data to support development and validation
- Step 6: Build and provide initial evaluation of interim composite
- Step 7: Evaluate composite based on archival data

Step 1: Identify first-term enlisted performance dimensions

- Focus on 10 joint-service performance dimensions adapted from the joint-service performance taxonomy for entry-level enlisted occupations (Russell et al., 2022)

Task Performance, Decision Making, Problem Solving, and Innovation
Communication
Safety and Security Consciousness
Initiating Structure for Self and Others
Conscientious Initiative
Support for Peers
Organizational Support
Adjusting to Stressful Situations
Counterproductive Work Behavior
Physical Performance

Step 2: Capture "overall performance" policy

- To form a TAPAS composite optimized to predict overall performance, we first needed to gain an understanding of the relative importance Services place on various performance dimensions when defining overall performance

- We administered a "policy capturing" exercise to each Service's MAPWG policy representatives and key stakeholders the policy reps identified who could help define overall performance priorities for first-term enlisted Servicemembers in their Service

- Services were asked to distribute 100 points across the 10 performance dimensions so that the resulting point distribution reflected how their Service would "effectively" weight these dimensions for purposes of defining an overall first-term enlisted *performance* composite

Step 2: Capture "overall performance" policy

Performance Dimension	Final Consensus Profiles					
	Army	Navy	Marine Corps	Air Force	Space Force	Mean
Task Performance, Decision Making, Problem Solving, and Innovation	12.0	16.0	15.0	25.0	17.0	17.0
Communication	12.0	10.0	8.0	8.0	11.0	9.8
Safety and Security Consciousness	9.0	11.0	6.0	5.0	10.0	8.2
Initiating Structure for Self and Others	9.0	7.0	8.0	8.0	8.0	8.0
Conscientious Initiative	10.0	10.0	8.0	12.0	11.0	10.2
Support for Peers	11.0	10.0	5.0	14.0	11.0	10.2
Organizational Support	10.0	13.0	13.0	16.0	12.0	12.8
Adjusting to Stressful Situations	10.0	11.0	8.0	7.0	10.0	9.2
Counterproductive Work Behavior	4.0	6.0	14.0	0.0	3.0	5.4
Physical Performance	13.0	6.0	15.0	5.0	7.0	9.2

- Overall, Services tended to give most weight to the *Task Performance, Decision Making, Problem Solving, and Innovation* dimensions, and least weight to the *Counterproductive Work Behavior* dimension

- The reliability of the mean, cross-service profile was relatively high, *ICC*(C,5) = .76

- The Air Force, Space Force, and Marine Corps profiles were most aligned with the mean profile (*r* = .96, .92, and .86, respectively), whereas the Army and Marine Corps were least aligned (*r* = .61, and .31, respectively)

Step 3: Define universe of potential TAPAS facets for interim composite

- Currently, Services administer different versions of TAPAS during the enlisted application process
  - Only six facets are common across the versions
  - In total, there are 24 facets that appear on at least one of the versions
- The current development effort aims to build an interim joint-service TAPAS composite from a *subset* of the 24 facets, given item content is already available for those facets
  - Not enough time to administer all 24 facets as part of application process

Step 4: Establish interim composite development and validation strategies

- DTAC and Accession Policy aim to have recommendations and a preliminary evaluation of the *interim* composite available by the fall of 2023

- Challenges
  - Lack of observed criterion data on the job performance criteria of interest
  - Timeframe does not allow for execution of local criterion-related validation study

- Solution
  - Ground identification of facets and initial evaluation of composite in a well-established validation framework from the I-O psychology literature

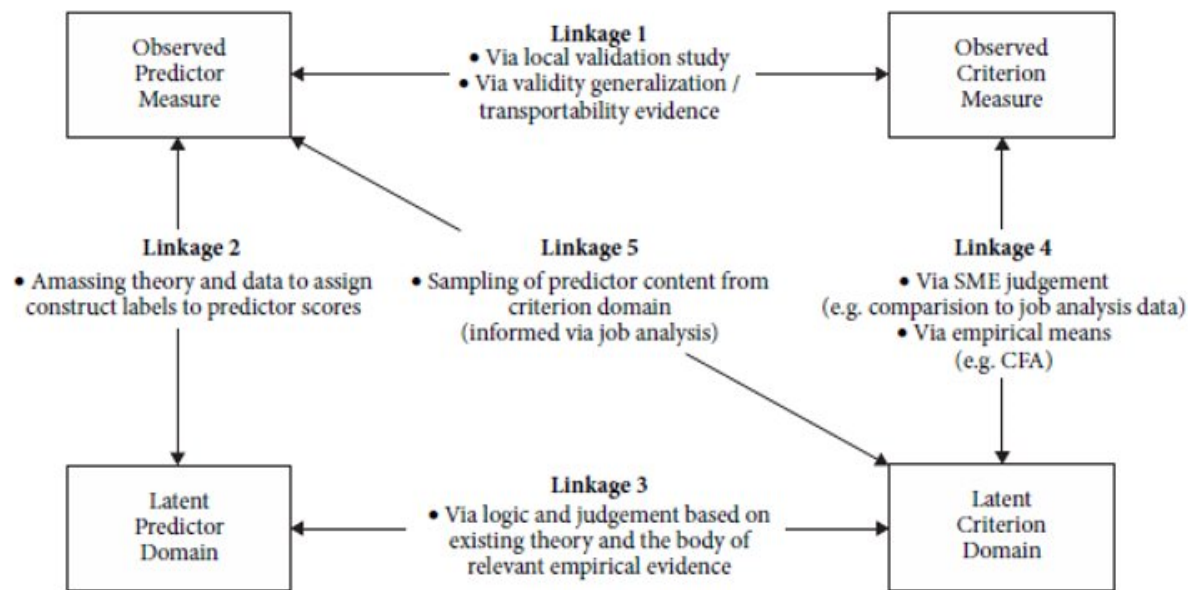# Step 4: Establish interim composite development and validation strategies

Fig. 6.1 Routes to Establishing the Predictive Inference.

- Approach to developing and initially validating the interim joint-service TAPAS composite is based on establishing Linkages 2 and 3

- Evidence for Linkage 2 has arguably been well established (see TAPAS Validity Argument work)

- Evidence for Linkage 3 is a focus of our initial development and validation effort

Figure adapted from:
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 University Press.

Binning, J.F., & Barrett, G.V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74, 478–494.*

Step 4: Establish interim composite development and validation strategies

- Build an AFQT-TAPAS-performance dimension correlation matrix based on archival applicant data and SME-estimated correlations

- Use these data to simulate *n* "population" correlation matrices
  - Multiple potential populations that reflect uncertainty due to variation in SME estimates

- Obtain frequency distributions of AFQT scores and TAPAS facets from archival applicant data

- Use the population matrices and distributions to estimate *n* large samples of AFQT-TAPAS-performance dimension data

- Calculate overall performance composite by applying *nominal* weights for each performance dimension in each sample

- Generate full path of lasso models using the overall performance composite as the criterion and 24 TAPAS facets as the starting set of predictors
  - Lasso (Least Absolute Shrinkage and Selection Operator) is a regularized regression model that performs variable selection

- Identify which TAPAS facets tend to remain in the model as lasso constraint becomes more stringent and examine the tradeoff between number of facets included and model *R*

Step 5: Gather archival and SME data to support development and validation

- We obtained archival applicant data on AFQT and TAPAS described on the previous slide
  - Data allowed us to empirically estimate AFQT-TAPAS and TAPAS-TAPAS correlations, and create AFQT-TAPAS score distributions using very large applicant samples

- We also asked a group of 11 external Ph.D. researchers with expertise in personality/cognitive ability, job performance relations, and job performance constructs to estimate three sets of correlations
  - Correlations between each of the 24 TAPAS facets and each of the 10 performance dimensions
  - Correlations between the AFQT and each of the 10 performance dimensions
  - Correlations among the 10 performance dimensions

- SMEs provided "construct" level correlation estimates—assumed performance dimensions were free from error and no range restrictions

Step 6: Build and provide initial evaluation of interim composite

- At the end of Step 5, we had all the data we needed to build AFQT-TAPAS-performance dimension correlation matrices

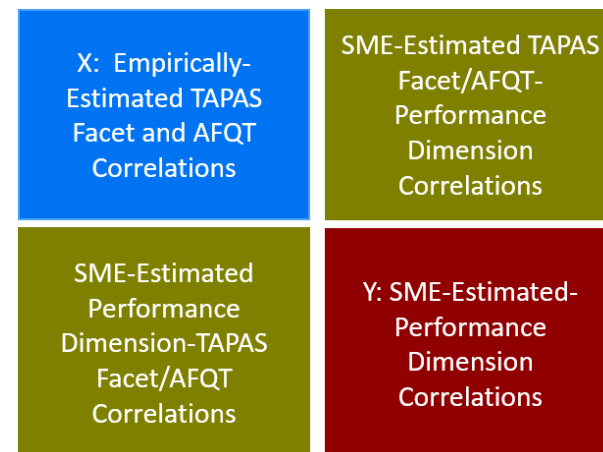| Correlations empirically estimated from existing applicant data (very large n)	X: Empirically-Estimated TAPAS Facet and AFQT Correlations	SME-Estimated TAPAS Facet/AFQT-Performance Dimension Correlations	Correlations estimated by SMEs as part of Step 5
Correlations estimated by SMEs as part of Step 5	SME-Estimated Performance Dimension-TAPAS Facet/AFQT Correlations	Y: SME-Estimated-Performance Dimension Correlations	Correlations estimated by SMEs as part of Step 5

Step 6: Build and provide initial evaluation of interim composite

- Used data gathered in Step 5 to simulate data for multiple large samples of individuals. Within each sample we:
  - Generated the full path of lasso models
  - Identified which TAPAS facets tend to remain in the model as lasso constraint becomes more stringent
  - Conducted follow-up that evaluated:
    - Sensitivity of results to how performance dimensions were weighted, using a service-specific performance dimension profile as opposed to the cross-service profile
    - How best-bet TAPAS facets change when a residualized version of the overall performance composite was used as the criterion (removing variance due to AFQT)—helps identify an interim composite that would best increment the validity of AFQT for predicting overall performance

X: Empirically-Estimated TAPAS Facet and AFQT Correlations	SME-Estimated TAPAS Facet/AFQT-Performance Dimension Correlations
SME-Estimated Performance Dimension-TAPAS Facet/AFQT Correlations	Y: SME-Estimated-Performance Dimension Correlations

Note: The work described on this slide is being conducted in the summer of 2023. Results from the analyses above were not available in time for inclusion in this slide presentation.

OPA
OFFICE OF PEOPLE ANALYTICS

Step 7: Evaluate composite based on archival data

- As a *potential* final evaluation of the interim joint-service composite, we may "score" the best-bet interim joint-service composite developed in Step 6 using archival TAPAS data from applicants to evaluate:

  - Magnitude of subgroup differences on the composite in applicant samples (and in turn the potential for adverse impact).

  - Criterion-related validity of the composite for predicting other criteria of historical interest (e.g., first-term attrition)

  - Difference between composition and weighting of the interim joint-service TAPAS vs. a TAPAS composite optimized to predict first-term attrition

X: Empirically-Estimated TAPAS Facet and AFQT Correlations	SME-Estimated TAPAS Facet/AFQT-Performance Dimension Correlations
SME-Estimated Performance Dimension-TAPAS Facet/AFQT Correlations	Y: SME-Estimated-Performance Dimension Correlations

Note: The above possibilities depend in part on whether all the TAPAS facets in for the interim joint-service composite are available in the archival data that may be available

OFFICE OF PEOPLE ANALYTICS

Generating Further Validity Evidence

Generating Further Validity Evidence

- Criterion measures
  - Administrative data, such as training completion and attrition
  - Self-report and possibly peer or instructor/NCO ratings generated in the DoD Criterion Measures project (Ellis et al., 2023; Ford et al., 2020)
- Data collection challenges
  - Administrative data sources are complex and limited
  - Outside of IRB requirements, Services other than the Army don't have established paths for obtaining cooperation to collect data from military personnel

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for DAC Consideration

- Suggestions for alternatives to criterion-related validity evidence and/or ideas for making data collection more feasible?

- Any concerns about our strategy to identify an interim TAPAS joint-service selection composite?

- Thoughts related to implementation of an interim TAPAS joint-service selection composite?

- Suggestions for additional R&D related to TAPAS that may not be covered by the TEP and TAPAS validity argument framework recommendations?

Thank you!

For more information please contact:

Deirdre Knapp and Dan Putka

dknapp@humrro.org
dputka@humrro.org
703-706-5662

PA
OFFICE OF PEOPLE ANALYTICS