OPA

OFFICE OF PEOPLE ANALYTICS

Refinement of the Joint-Service TAPAS Instrument

Dan Putka

Human Resources Research Organization

Briefing presented to the DACMPT

January 23, 2025

Briefing Agenda

- Joint-Service (JS) TAPAS Background Refresh
  - JS TAPAS Composites, Instrument, and Development Phases
- Recap of Preliminary Phase 1 JS TAPAS Composite Recommendations
- FY24 Research to Inform Phase 1 JS TAPAS Revisions
- Finalizing the Phase 1 JS TAPAS Design
- Next Steps for JS TAPAS
  - Operations and Maintenance (O&M) Track (FY25)
  - R&D Track (FY25-26)
- Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service TAPAS Background Refresh

OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service TAPAS Mission

1. Develop a composite for military compatibility
   - Designed to predict alignment with military core values — various forms of misconduct
   - DoD directive that applies to enlisted personnel

2. Develop a composite for enlisted selection
   - Designed to predict first-term enlisted job performance
   - Expand qualified applicant pool without compromising valued outcomes

3. Develop a Joint-Service TAPAS instrument

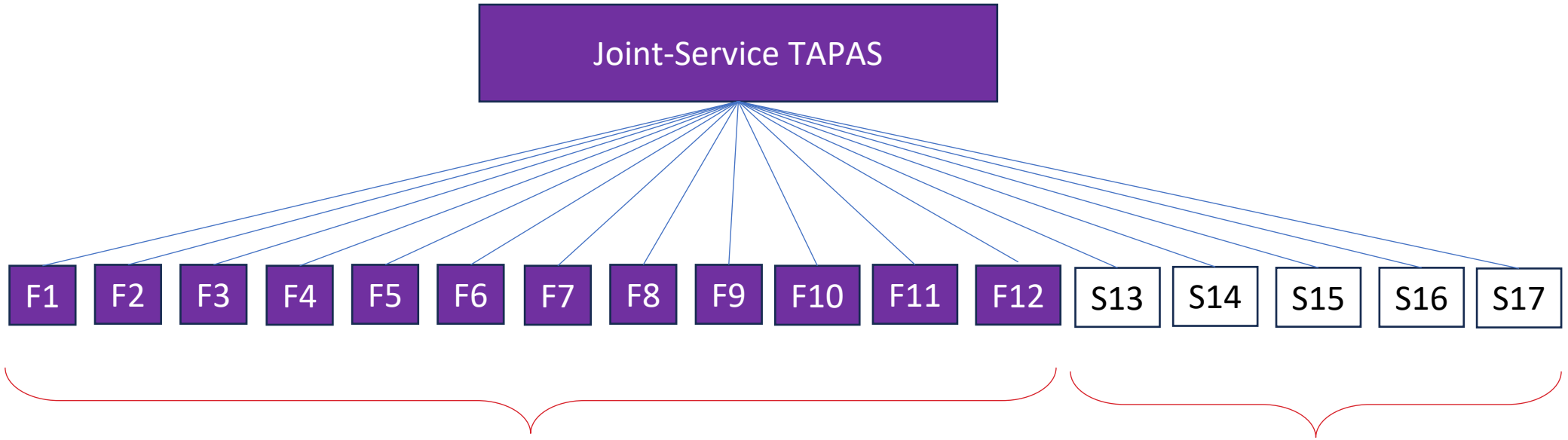
OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service (JS) TAPAS Concept

- The JS TAPAS “instrument” is modular and will include:
  - A common core of facets that support scoring of the military compatibility (MC) and enlisted (ENL) composites
  - Service-specific facets to support Service-specific use cases

OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service TAPAS Concept

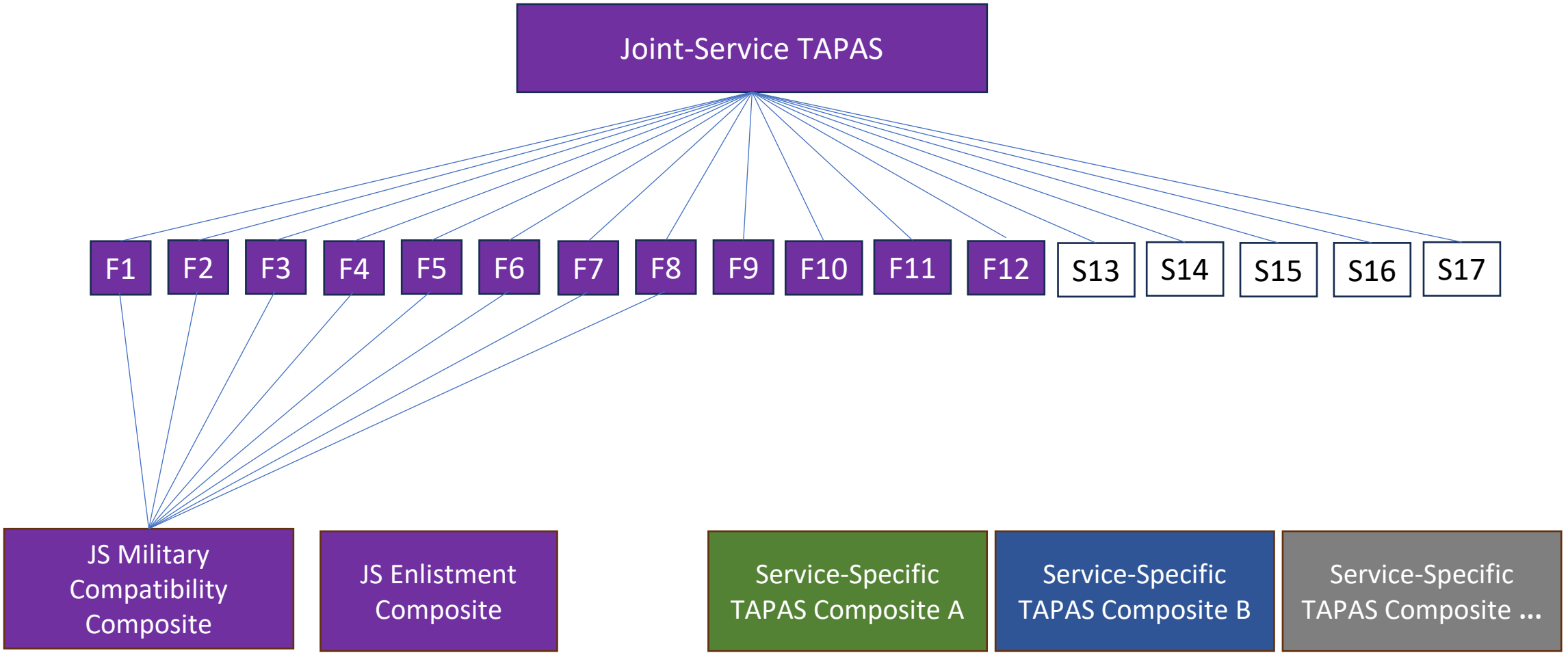


JS MC and ENL composite facets shared across all TAPAS versions

Slots reserved for Service-specific facets

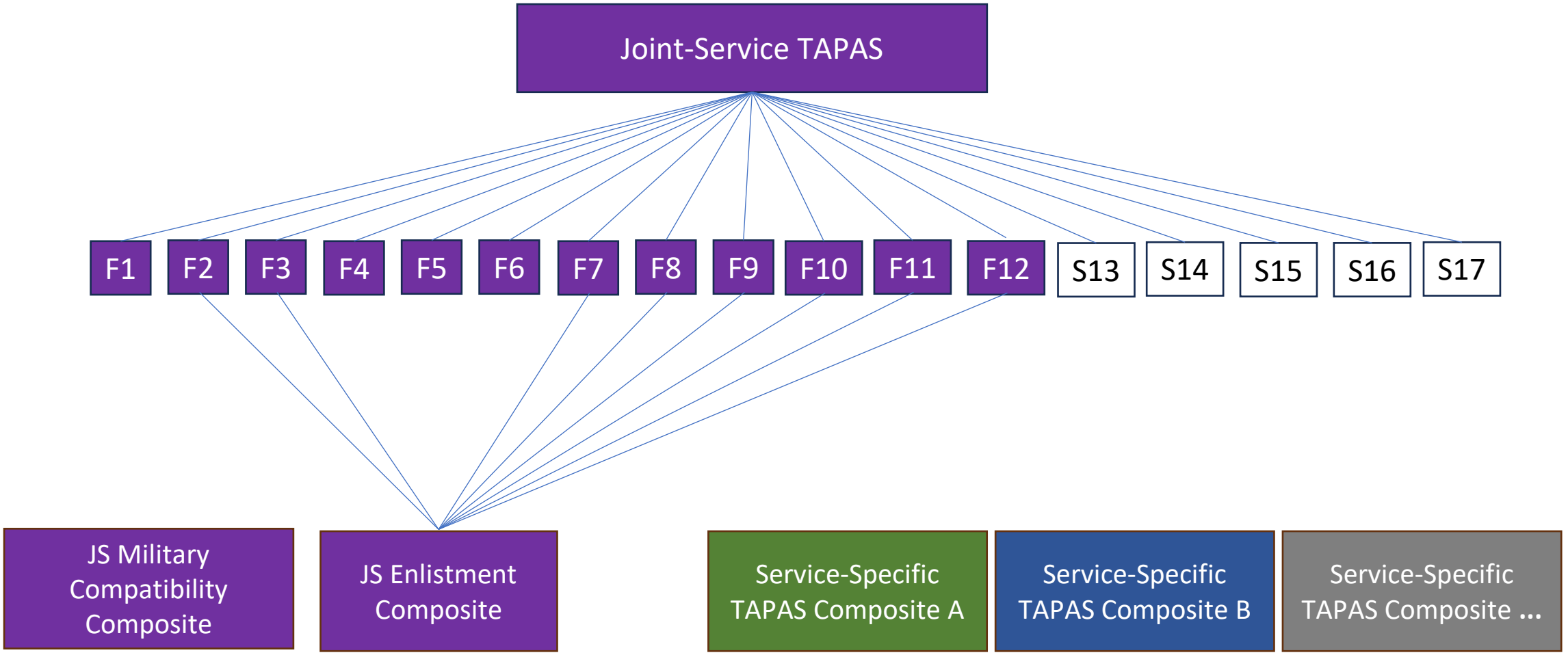
OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service TAPAS Concept



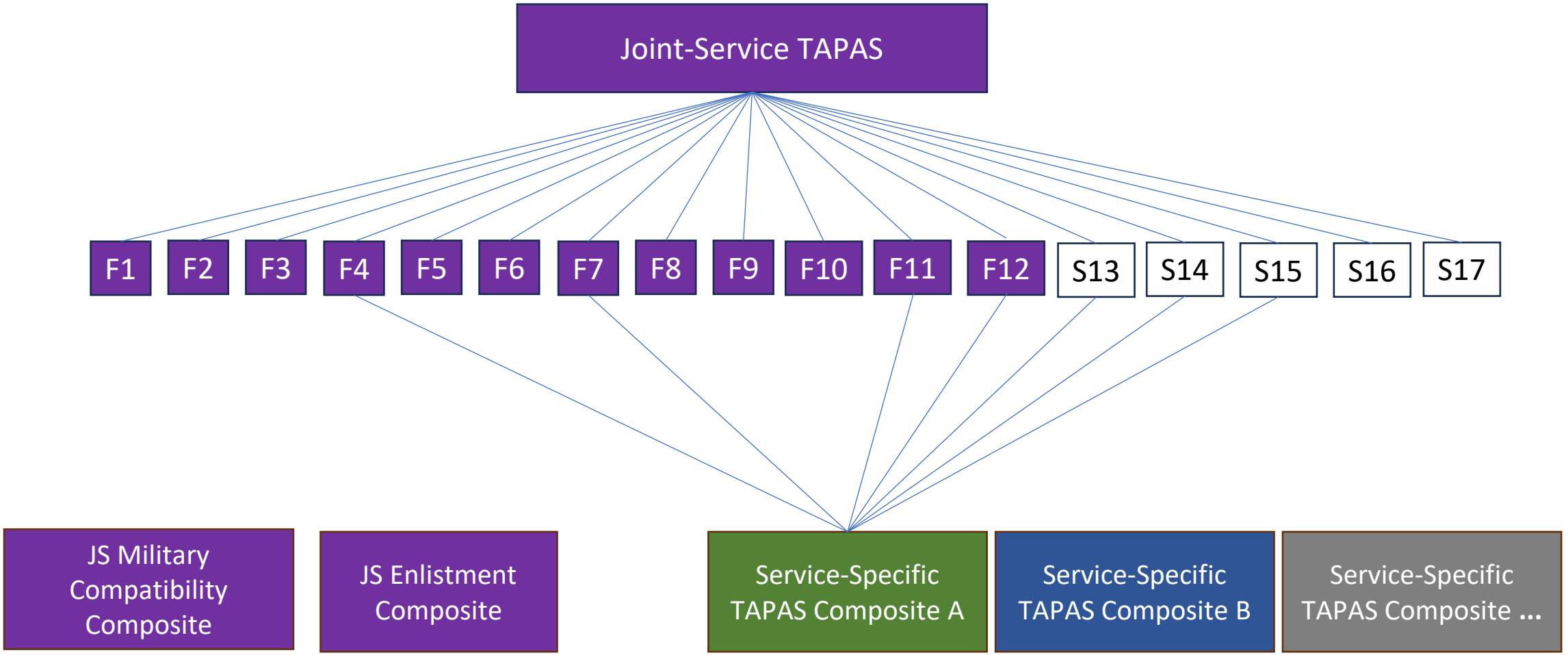
OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service TAPAS Concept



OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service TAPAS Concept



OPA
OFFICE OF PEOPLE ANALYTICS

Phased Development Approach

- **Phase 0** JS TAPAS instrument and composites
  - FY23 work designed to address immediate OSD tasking
  - Features **interim** MC and ENL composites
  - Added facets to USAF and USMC TAPAS needed for scoring of interim MC composite
  - Implemented at MEPS in September 2024
    - Phase 0 MC and ENL composites scored but not used for operational decision making

OPA
OFFICE OF PEOPLE ANALYTICS

Phased Development Approach

- **Phase 1** JS TAPAS instrument and composites
  - **Preliminary** recommendations for Phase 1 composites (facets, weighting) made based on FY23 research
  - **Refined** recommendations for Phase 1 instrument (design, JS facet set) made based on FY24 research
  - Content development and psychometric work to occur in FY25
    - **Refined** composition and facet weighting Phase 1 MC and ENL composites
    - Updating of TAPAS statements pools
    - Calibrating TAPAS statement pools with a joint-Service sample
    - Develop provisional joint-Service norms for JS and SS facets
  - IT work to enable implementation at MEPS sometime in FY27 (TBD)

Phased Development Approach

- **Phase 2**: Evaluation and refinement of Phase 1 JS composites for operational decision making
  - Update joint-Service norms for JS and SS facets
    - Informed by FY27 applicant data and subsequent evaluation work
  - Revisit composition and weights for each Phase 1 composite and adjust as needed
  - Establish an evidentiary base for use of final Phase 2 composites for enlistment and military compatibility-related screening decisions (e.g., criterion-related validity study for enlistment composite)

FY23	Foundational research to inform Phase 0 and preliminary recommendations for Phase 1 JS TAPAS composites	
FY24	Follow-up research to inform refined recommendations for Phase 1 JS TAPAS instrument (design, facet set)	IT work to support implementation of Phase 0 JS TAPAS at MEPS
FY25	Development work to support Phase 1 JS TAPAS instrument and refined composites + JS TAPAS R&D	Began administering Phase 0 JS TAPAS at MEPS
FY26	JS TAPAS R&D (continued) Evaluation and refinement of Phase 1 JS TAPAS composites	IT work to support implementation of Phase 1 JS TAPAS at MEPS
FY27		Begin administering Phase 1 JS TAPAS at MEPS
FY28		

Begin operational use of TAPAS composites based on FY26–FY28 work

OPA
OFFICE OF PEOPLE ANALYTICS

Recap of Preliminary Phase 1 JS TAPAS Composite Recommendations

OPA
OFFICE OF PEOPLE ANALYTICS

Military Compatibility (MC) Composite – Focal Criterion

- Focal criterion reflects 10 categories of misconduct
  - Violent Behavior
  - Sexual Violence/ Assault
  - Sexual Harassment
  - Harassment and Non-Violent Abuse
  - Disclosing Classified or Sensitive Information
  - Rebellious/Extremist Behavior
  - Unethical Behavior
  - Vandalism/Sabotage
  - Theft
  - Production Deviance
- Informed by literature and expert review
  - Counterproductive work behavior (CWB) literature (e.g., Spector et al., 2006)
  - Uniform Code of Military Justice
  - OPA/PERSEREC reports
  - DoD instruction 1304.26

OPA
OFFICE OF PEOPLE ANALYTICS

Preliminary Phase 1 MC Composite Recommendations

- Subject matter experts (SMEs) evaluated conceptual and empirical evidence of alignment between TAPAS facets and 10 categories of misconduct
  - Rated alignment as strong, moderate, or weak
- Reached consensus on facet composition and weighting for a preliminary Phase 1 MC composite *[facets withheld for test security]*
  - See June 2023 DACMPT slides for more details

OPA
OFFICE OF PEOPLE ANALYTICS

Enlistment Composite – Focal Criterion

- Focal criterion reflects first-term enlisted job performance composite
- Based on performance dimensions from Russell et al. (2023)[1] taxonomy
- Captured "overall performance" policy from Service stakeholders
  - Task Performance, Decision Making, Problem Solving, and Innovation (m = 17.0)
  - Organizational Support (m = 12.8)
  - Support for Peers (m = 10.2)
  - Conscientious Initiative (m = 10.2)
  - Communication (m = 9.8)
  - Adjusting to Stressful Situations (m = 9.2)
  - Physical Performance (m = 9.2)
  - Safety and Security Consciousness (m = 8.2)
  - Initiating Structure for Self and Others (m = 8.0)
  - Counterproductive Work Behavior (m = 5.4)

Note. Parenthetical values reflect distribution of 100 points across dimensions.

[1]Russell, T., Allen, M., Ford, L., Carretta, T., & Kirkendall, C. (2023). Development of a performance taxonomy for entry-level military occupations. *Military Psychology*, *35*(4), 283-294. https://doi.org/10.1080/08995605.2022.2050163.

OPA
OFFICE OF PEOPLE ANALYTICS

Preliminary Phase 1 ENL Composite Recommendations

- Gathered archival and SME data to support development and validation
  - Developed regression-weighted composite based on mix of archival and SME-estimated correlations
  - See June 2023 DACMPT slides for more details
- Identified subset of facets for predicting first-term enlisted job performance based on regression models *[facets withheld for sensitivity]*

OPA
OFFICE OF PEOPLE ANALYTICS

FY24 Research to Inform Phase 1 JS TAPAS Revisions

OPA
OFFICE OF PEOPLE ANALYTICS

Focus of FY24 Research

- Largely focused on refining preliminary Phase 1 JS TAPAS recommendations and identifying needs for FY25 development work
  - Conducted multiple research efforts pertinent to evaluating TAPAS facets and their statement pools
  - Engaged in multiple rounds of discussion with OSD and Services to arrive at an agreed-upon set of JS facets and JS instrument design/configuration
    - Factoring in FY23 recommendations *AND* FY24 research results
  - Established plans for recalibration of TAPAS statements with a joint-Service sample

OPA
OFFICE OF PEOPLE ANALYTICS

Outline of FY24 Research Activities

- Retranslation of facet statements
- Bias and sensitivity review of facet statements
- Susceptibility of facet statements to transient error
- Revisiting marginal IRT reliability of facet scores
- Equivalence of facet scores across TAPAS versions
- Composite shortening analyses

The above research provided additional perspectives on the functioning of TAPAS facets beyond what was known when preliminary Phase 1 composites recommendations were made in FY23.

OPA
OFFICE OF PEOPLE ANALYTICS

Retranslation of Facet Statements

- Purpose
  - Evaluate whether TAPAS statements are clear indicators of their intended facets
- Method
  - Leveraged natural language processing (NLP) methods to identify items most in need of review by SMEs (*n* = 482 out of 1,200+ statements in DoD TAPAS statement pool)
    - Focused on statements that were more semantically similar to statements of another facet rather than their intended facet
  - Eight psychologist SMEs independently indicated which facet each statement primarily measured
    - At least 6 of 8 (75%) SMEs had to agree on the facet a statement was designed to measure for it to be considered “translated” to that facet

Retranslation of Facet Statements

Target Facet	% of the Target Facet Statements Retranslated to		
	Target Facet	Non-Target Facet	No Clear Translation
Physical Conditioning	100.0	0.0	0.0
Team Orientation	100.0	0.0	0.0
Tolerance	100.0	0.0	0.0
Order	96.0	0.0	4.0
Sociability	95.8	2.1	2.1
Non-Delinquency	95.7	2.2	2.2
Army Self Efficacy	95.6	4.4	0.0
Selflessness	94.4	0.0	5.6
Cooperation	93.5	0.0	6.5
Dominance	92.2	2.0	5.9
Persistence	88.9	0.0	11.1
Even Tempered	84.9	1.9	13.2
Intellectual Efficiency	84.1	4.5	11.4
Courage	78.6	3.6	17.9
Self Control	76.2	7.1	16.7
Commitment to Serve	75.0	13.5	11.5
Virtue	74.0	6.0	20.0
Adjustment	73.6	3.8	22.6
Humility	71.1	6.7	22.2
Optimism	70.2	12.8	17.0
Achievement	70.2	12.3	17.5
Self Efficacy	69.6	2.2	28.3
Responsibility	65.9	2.4	31.7
Situational Awareness	64.6	6.3	29.2
Attention Seeking	61.7	14.9	23.4

Key Findings

- Facets varied in the % of statements translated by SMEs into their target facet, with some facets (e.g., Physical Conditioning) exhibiting perfect retranslation and others (e.g., Attention Seeking) exhibiting relatively poor retranslation (61.7%)

Recommendations for FY25

- Have humans retranslate remainder of statements in pool
- Move statements to proper facet as needed and recalibrate
- Revise statements so they have a clear translation and recalibrate

Note. Statements not flagged for retranslation by the NLP methods for rating by SMEs were considered as translated to their target facet for purposes of the percentages in this table. Facets are sorted in descending order based on the percentage of statements in their pool that was successfully retranslated to their target facet. Cells are color coded to facilitate interpretation. Green/red indicates better/poorer retranslation results.

Bias and Sensitivity Review of Facet Statements

- Purpose
  - Identify TAPAS statements that may be problematic from a bias or sensitivity perspective
- Method
  - Each statement was evaluated by two external SMEs with expertise in bias and sensitivity review (a total of four external SMEs participated in this exercise)
    - Five categories of biased-sensitive language considered (see next slide)
  - Statements flagged by at least one external SME underwent a second round of review by three internal experts who indicated whether statements should be revised or dropped, and reason(s) for doing so

OPA
OFFICE OF PEOPLE ANALYTICS

Bias and Sensitivity Categories

1. **Unfamiliar Term:** The item uses simple, familiar terms that most people can understand and avoids unnecessarily complex or obscure language. For example, if an item is attempting to describe something that is uninteresting, it would be more appropriate to use words such as *boring*, *uninteresting*, or *dull* than to use words such as *jejune*, *pedestrian*, or *humdrum*.
2. **Colloquial:** The item avoids informal and figurative expressions such as colloquialisms ("wicked good"), slang ("nuts"), idioms ("break a leg"), aphorisms ("*when it rains, it pours*"), and technical jargon ("masthead"). The meaning of such terms may not be clear to examinees from a wide variety of backgrounds. Instead, clearly describe the concept of interest in a way that can be reasonably considered comprehensible to all examinees. For example, a more appropriate phrasing of the item "I am easily thrown off" would be "I am easily distracted."
3. **Unfamiliar Situation:** The item avoids situations, contexts, behaviors, and/or other content that will likely not be familiar or accessible to, or feasible for, examinees from a wide variety of cultural, social, and economic backgrounds. For example, "I regularly attend opera performances" would be a less appropriate measure of individuals' artistic interests than "I enjoy listening to classical music."
4. **Controversial Language:** The item avoids language that could be reasonably considered controversial, inflammatory, offensive, insensitive, or otherwise likely to distract examinees by inducing strong emotional reactions (e.g., anger, distress, sadness). This includes avoiding invoking potentially upsetting or controversial topics and concepts (e.g., abortion, colonialism, death, extreme pain, religion, sexuality, violence, illegal activities) explicitly *or* implicitly. For example, the item "I always choose the master bedroom" does not directly concern the controversial topic of slavery, but the historical association of slavery with the term "master bedroom" means it would be inappropriate to phrase the item in this way.
5. **Discrimination**: The item avoids explicit *or* implicit reference to groups that could potentially be discriminated against. Such groups include those related to characteristics such as age, appearance (e.g., attractiveness, height, weight), citizenship status, culture, disability, ethnicity, gender (including gender identity or gender representation), national or regional origin, native or primary language, political beliefs, race, religion (or its absence), sexual orientation, socioeconomic status. For example, the item "I often feel gypped by my friends" indirectly refers to "gypsy," a derogatory term sometimes applied to the Romani people.

Bias and Sensitivity of Facet Statements

Target Facet	% Fair	% Revise	% Drop	Reason for Revision/Drop				
				% Unfamiliar Term	% Colloquial	% Unfamiliar Situation	% Controversial Language	% Discrimination
Situational Awareness	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dominance	98.0	2.0	0.0	2.0	0.0	0.0	0.0	2.0
Army Self Efficacy	97.8	2.2	0.0	0.0	2.2	0.0	0.0	0.0
Team Orientation	94.3	5.7	0.0	1.9	3.8	0.0	0.0	1.9
Humility	91.1	8.9	0.0	6.7	0.0	2.2	0.0	2.2
Intellectual Efficiency	90.9	9.1	0.0	2.3	4.5	0.0	0.0	0.0
Selflessness	90.7	9.3	0.0	7.4	5.6	1.9	0.0	3.7
Cooperation	89.1	10.9	0.0	6.5	10.9	0.0	0.0	0.0
Commitment to Serve	88.5	11.5	0.0	1.9	11.5	1.9	0.0	0.0
Virtue	88.0	8.0	4.0	6.0	12.0	2.0	0.0	0.0
Achievement	87.7	7.0	5.3	3.5	10.5	0.0	1.8	0.0
Tolerance	86.4	9.1	4.5	2.3	9.1	9.1	0.0	2.3
Courage	85.7	7.1	7.1	1.8	8.9	1.8	3.6	0.0
Sociability	83.3	16.7	0.0	10.4	16.7	0.0	0.0	0.0
Responsibility	82.9	17.1	0.0	2.4	14.6	4.9	0.0	0.0
Even Tempered	81.1	17.0	1.9	1.9	18.9	0.0	0.0	0.0
Adjustment	81.1	18.9	0.0	5.7	17.0	0.0	0.0	0.0
Self Control	81.0	16.7	2.4	4.8	14.3	2.4	2.4	0.0
Self Efficacy	80.4	19.6	0.0	8.7	10.9	4.3	0.0	2.2
Non-Delinquency	80.4	8.7	10.9	2.2	17.4	2.2	0.0	4.3
Physical Conditioning	79.6	13.0	7.4	1.9	11.1	0.0	5.6	3.7
Optimism	78.7	17.0	4.3	8.5	19.1	0.0	0.0	0.0
Attention Seeking	72.3	23.4	4.3	4.3	27.7	4.3	0.0	0.0
Order	70.0	26.0	4.0	20.0	12.0	6.0	0.0	0.0
Persistence	62.2	37.8	0.0	6.7	28.9	0.0	2.2	0.0

Key Findings

- Almost all facets had statements that were flagged for one or more reasons, though **most flags were related to use of unfamiliar/colloquial terms** rather than use of controversial or discriminatory language

Recommendations for FY25

- Have internal experts review remainder of pool
- Write new statements to replace drops and calibrate
- Revise statements flagged for revision and recalibrate

Note. Facets are sorted in descending order of the percentage of statements in their pool deemed fair by external and internal experts. Green/red indicates higher/lower percentages of facet statements deemed fair. Percentages under Reasons for Revision/Drop columns reflect the percentages of all statements in the facet's statement pool flagged by internal SMEs for the given reason (a statement could be flagged for more than one reason).

Susceptibility of Facet Statements to Transient Error

- Goal
  - Evaluate TAPAS statements for susceptibility to transient error variance
- Method
  - Eight psychologist SMEs independently rated each statement on the following scale:

    "Please rate how much you think applicants' responses to the following statements would be influenced by their psychological/physical state at the time of testing (e.g., based on their mood, how they physically feel, etc.), using a scale of 1 (not at all influenced), 2 (slightly influenced), 3 (moderately influenced), and 4 (very influenced)"

Susceptibility of Facet Statements to Transient Error

Target Facet	Percentage of Statements with Mean Ratings in the Given Range				
	1.0 to 1.5	1.6 to 2.0	2.1 to 2.5	2.6 to 3.5	3.6 to 4.0
Intellectual Efficiency	100.0	0.0	0.0	0.0	0.0
Order	100.0	0.0	0.0	0.0	0.0
Team Orientation	98.1	1.9	0.0	0.0	0.0
Tolerance	97.7	2.3	0.0	0.0	0.0
Attention Seeking	95.7	4.3	0.0	0.0	0.0
Non-Delinquency	94.4	5.6	0.0	0.0	0.0
Selflessness	94.4	5.6	0.0	0.0	0.0
Responsibility	92.6	4.9	2.4	0.0	0.0
Situational Awareness	91.7	9.3	0.0	0.0	0.0
Persistence	88.9	11.1	0.0	0.0	0.0
Self Control	88.1	11.9	0.0	0.0	0.0
Virtue	88.0	12.0	0.0	0.0	0.0
Dominance	86.3	13.7	0.0	0.0	0.0
Courage	83.9	16.1	0.0	0.0	0.0
Sociability	83.3	16.7	0.0	0.0	0.0
Cooperation	82.6	17.4	0.0	0.0	0.0
Achievement	82.5	17.5	0.0	0.0	0.0
Humility	82.2	17.8	0.0	0.0	0.0
Commitment to Serve	73.1	23.1	3.8	0.0	0.0
Physical Conditioning	61.1	37.0	1.0	0.0	0.0
Even Tempered	56.6	35.8	7.5	0.0	0.0
Self Efficacy	50.0	41.3	8.7	0.0	0.0
Adjustment	38.9	37.7	22.6	0.0	0.0
Army Self Efficacy	31.1	60.0	8.8	0.0	0.0
Optimism	19.1	53.2	25.5	2.1	0.0

Key Findings

- Overall, SMEs viewed responses to TAPAS statements as NOT very susceptible to transient error (low mean ratings)
- Statements rated as slightly more susceptible were consistent with expectations, given affective elements associated with those facets (e.g., Optimism, Adjustment, Even Tempered)

Recommendations for FY25

- Revisit/revise statements with ratings 2.0 or greater, if deemed warranted, and recalibrate

Note. Scale points: 1 (not at all influenced), 2 (slightly influenced), 3 (moderately influenced), and 4 (very influenced). Facets are sorted in descending order of the percentage of statements in their pool that had mean ratings in the range of 1.0 to 1.5. Green indicates higher percentages of facet statements were deemed not susceptible to transient error, and red indicates lower percentages of facet statements were deemed not susceptible to transient error.

Revisiting Marginal IRT Reliability of Facet Scores

- Goal
  - Provide updated estimates of marginal IRT reliability of facet scores based on large, current sets of applicant data (or published data when not available)

- Method
  - Evaluated marginal IRT reliability of TAPAS facet scores for TAPAS versions used by Army, USAF, and USMC in 2021–2023 and that were current as of February 2024
  - Based on applicant records where no more than one TAPAS response check item was incorrect
    - Army *n* = 212,726: TAPAS taken between 11/30/21 – 12/1/23
    - USAF *n* = 108,063: TAPAS taken between 6/30/21 – 1/3/24
    - USMC *n* = 82,794: TAPAS taken between 6/27/21 – 12/29/23

OPA
OFFICE OF PEOPLE ANALYTICS

Revisiting Marginal IRT Reliability Estimates of Facet Scores

Facet	*n* TAPAS Versions	Estimate
Physical Conditioning	3	0.76
Sociability	3	0.76
Commitment to Serve*	2	0.75
Order	2	0.69
Dominance	3	0.68
Cooperation	1	0.68
Courage	1	0.67
Adjustment	2	0.65
Selflessness	2	0.65
Intellectual Efficiency	1	0.64
Attention Seeking	2	0.63
Team Orientation	2	0.63
Non-Delinquency	2	0.62
Tolerance	3	0.61
Even Tempered	2	0.61
Responsibility	1	0.60
Achievement	3	0.58
Persistence*	2	0.57
Situational Awareness	1	0.55
Self-Control	1	0.54
Virtue*	1	0.53
Optimism	3	0.49
Self-Efficacy*	1	0.46
Humility*	1	0.40

OPA
OFFICE OF PEOPLE ANALYTICS

Key Findings

- Facets exhibited relatively low to middling reliability (average estimates = .40 –.76) compared to suggested reliability standards for high-stakes testing (e.g., Lance et al., 2006; Nunnally, 1978)[1]
- Low levels of reliability suggest not using individual facet scores for decision making — composites would be more defensible

Recommendations for FY25

- Carefully examine statement pools for low reliability facets during FY25 content development (e.g., evidence of heterogeneity, multiple clear dimensions within a facet) and aim to bolster/refine statement pool for those facets

Note. Reliability estimates reflect simple point estimates or averages across TAPAS versions used by the Army, USAF, and USMC in the 2021–2023 timeframe. Green/red indicates relatively higher/lower levels of reliability. Facets with an asterisk are those for at least one reliability estimate sourced from Drasgow et al (2023) based on experimental Part 2 of the Army TAPAS versions.

[1]Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria. What did they really say? *Organizational Research Methods, 9(2),* 202–220. https://doi.org/10.1177/1094428105284.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Equivalence of Facet Scores Across TAPAS Versions

- Goal
  - *Start to evaluate* the comparability of facet scores from TAPAS versions that differed in their facet composition

- Method
  - Examined comparability of TAPAS facet intercorrelations across versions (e.g., Is the Facet A-B correlation the same across versions?)
  - Examined comparability of TAPAS facet — other variable correlations across versions (e.g., Is the Facet A-AFQT correlation the same across versions?)
  - Examined seven different versions of Army TAPAS used at MEPS over time that partially overlapped in their facet composition

A Note on Research Design Limitations

- Considered multiple potential approaches to examining equivalence…most of which were not feasible within the study timeframe
  - Administer multiple TAPAS versions to same respondents with facet composition systematically varied
  - Implement multigroup CFA based approaches to studying measurement invariance
    - Issues with applying factor analysis to partially ipsative data
  - IRT/item based-approaches
    - Examining item equivalence or understanding differences in scores based on items administered
  - Simulation based approaches
    - Identifying true thetas and running them through different TAPAS versions and to see how the observed thetas for a facet varied across versions
- Given time and feasibility — we adopted a simpler (albeit more limited) approach that focused only on similarity of TAPAS facet intercorrelations and TAPAS facet—other variable correlations (AFQT, 6-month attrition, 24-month attrition) across TAPAS versions that differed in their facet composition

Equivalence of Facet Scores Across TAPAS Versions

Key Findings

- TAPAS facet intercorrelations and TAPAS facet—other variable correlations were generally quite similar across versions, indicating facet mix may NOT have notable impact on a target facet's measurement
- When differences were found, they tended to be for TAPAS facet intercorrelations between TAPAS versions from different Army TAPAS development "stages"
  - Stage 2 (least use of cleaning/quality flags) → Stage 4 (most use of cleaning/quality flags)
  - Average absolute differences between same-facet correlations across versions
    - .014 WITHIN stages
    - .054 (Stage 2 vs. Stage 3 versions) and .047 (Stage 2 vs. Stage 4 versions)
- Between-stage differences in facet-intercorrelations didn't translate into differences in TAPAS facet-AFQT and TAPAS facet-attrition correlations

OPA
OFFICE OF PEOPLE ANALYTICS

Composite Shortening Analyses

- Goal
  - Evaluate the possibility of shortening preliminary Phase 1 MC and ENL composites
- Method
  - Performed best subsets regression using preliminary Phase 1 MC and ENL composites as criteria (separate models for each criterion) and the facets that contribute to those composites as initial predictors
    - Regressions based on facet intercorrelation matrices developed during the FY23 research
  - Identified what facets were consistently retained in models as the number of features in the predictor subset was reduced and the Multiple *R* achieved by those reduced models
- Key Findings
  - There appears to be room to shorten the preliminary Phase 1 MC and ENL composites and still achieve a very high correlation with the full versions of those composites

OPA
OFFICE OF PEOPLE ANALYTICS

Finalizing the Phase 1 JS TAPAS Instrument Design

OPA
OFFICE OF PEOPLE ANALYTICS

JS Instrument Design

- Only a limited number of facets can be administered as part of the JS TAPAS due to testing time constraints at MEPS and the cognitive load associated with the use of more facets
- Tradeoff between "number of facets" and "number of statements per facet"
  - More facets mean more flexibility to cover JS MC and ENL composites and Service-specific uses
  - Due to testing time constraints, more facets also means fewer items per facet, resulting in less reliable measurement
  - Greater number of statements per facet → higher marginal IRT reliability for facets
    - Drasgow et al (2023)[1] suggests 20 statements per facet
- Targeting no more than 17 facets for the JS TAPAS instrument — key decision point was how many facets to reserve for the Joint-Service facets vs. Service-specific facets

[1]Drasgow, F., Chernyshenko, O. S., Stark, S., Nye, C. D. (2023). *Tailored Adaptive Personality Assessment System (TAPAS): Pre-implementation documentation*. (AFRL-RH-WP-TR-2023-0014). Air Force Research Laboratory.

OPA
OFFICE OF PEOPLE ANALYTICS

Key Considerations for Identifying Joint-Service Facets

Facet-level considerations

- Use in, and importance to, preliminary FY23 recommendations for Phase 1 JS composites
- Use in, and importance to, Service-specific models/composites
- Performance along FY24 research metrics (i.e., retranslation, bias/sensitivity, IRT marginal reliability, transient error)
  - Deficiencies here have the potential to be addressed via subsequent FY25 development work
- Secondary consideration — relevance to outcomes perceived to be of broad interest across Services
  - First-term attrition
  - Enlisted leadership potential/emergence

Set-level considerations

- Balance in terms of personality construct mix
- More JS facets (better prediction of JS criteria + construct coverage) ← vs → fewer JS facets (more Service-specific slots)
- More facets overall (fewer statements per facet = lower facet reliability) ← vs → fewer facets overall (more statements per facet = higher facet reliability)

Finalizing the Set of JS Facets

- SMEs from HumRRO, DCG, and DTAC reviewed information for each facet given the facet-level and set-level considerations and developed recommendations for potential sets of facets to include in the Joint-Service set
- Goal was to identify a single set of facets that could be used to support scoring of refined Phase 1 JS MC and ENL composites
  - Different facets from the set would be used to score each composite OR all may be used for each composite but differentially weighted — **TBD during FY25 development work**
- Reviewed considerations, research findings, and recommendations with Service representatives and came to consensus on a set of 12 JS facets that would be included in the Phase 1 JS TAPAS
  - 5 additional facet "slots" reserved for Service-specific facets

OPA
OFFICE OF PEOPLE ANALYTICS

Next Steps for JS TAPAS

OPA
OFFICE OF PEOPLE ANALYTICS

Operations and Maintenance (O&M) Track (FY25)

- Preparing for implementation of Phase 1 composites
  - Statement pool development
    - Needs identified in FY24 research
  - Existing statement re-calibration and new statement calibration using a joint-Service sample
  - Finalizing composition and weighting of Phase 1 composites
  - Development of provisional joint-Service norms for facets
- IT work to enable implementation at MEPS sometime in FY27 (TBD)

OPA
OFFICE OF PEOPLE ANALYTICS

R&D Track (FY25–FY26)

R&D to evaluate/enhance TAPAS adjacent to the Phase 1 JS TAPAS O&M work

- Effects of practice and coaching on TAPAS
  - Practice effects on TAPAS
  - Coaching/large language model (LLM) informed response
- AI for non-cognitive assessment
  - Review potential role of AI in bringing efficiencies to non-cognitive assessment (e.g., statement development, statement parameter estimation)
- TAPAS and supervised machine learning (ML) for attrition prediction
  - Exploration of input/features below the facet level

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC

1. Should we hold TAPAS and cognitive test scores used for high-stakes decision making to different reliability standards? What's the minimum level of reliability you believe is acceptable for defending use of TAPAS composite scores for making high-stakes selection decisions?

2. Narrowing the construct we aim to cover with a TAPAS facet can help ensure unidimensionality of a facet's statement pool (which should help with reliability issues), but doing so would make it harder to develop a statement pool of sufficient size for use in TAPAS. Thoughts on strategies to deal with this tradeoff?

3. If our research finds TAPAS is susceptible to coaching effects (e.g., elevation of scores on particular facets), what suggestions do you have for mitigating such effects?

Thank you!

For more information please contact:

Dan Putka
dputka@humrro.org
703.706.5640

OPA
OFFICE OF PEOPLE ANALYTICS