



**AFRL-RH-WP-TR-2020-0019**

**IDENTIFICATION OF CRITERION CONSTRUCTS  
AND MEASURES FOR JOINT-SERVICE  
ENLISTED JOBS**

**Matthew T. Allen**

**Teresa L. Russell**

**Laura A. Ford**

**Human Resources Research Organization**

**66 Canal Center Plaza, Suite 700**

**Alexandria, VA 22314**

**Cristina D. Kirkendall**

**Army Research Institute for the Behavioral and Social Sciences**

**6000 Sixth Street**

**Fort Belvoir, MD 22060**

**Thomas R. Carretta**

**Air Force Research Laboratory**

**2210 8<sup>th</sup> Street, Area B, Bldg. 146, Room 122**

**Wright-Patterson AFB, OH 45433**

**March 2020**

**Interim Report**

**DISTRIBUTION STATEMENT A. Approved for public release; distribution unlimited.**

**AIR FORCE RESEARCH LABORATORY  
711 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 88<sup>th</sup> Air Base Wing 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-2020-0019 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signed//

THOMAS R. CARRETTA  
Work Unit Manager  
Interfaces and Teaming Branch  
Warfighter Interface Division  
Airman Systems Directorate  
711<sup>th</sup> Human Performance Wing  
Air Force Research Laboratory

//signed//

TIMOTHY S. WEBB  
Chief, Collaborative Interfaces and Collaborative  
Teaming Branch  
Warfighter Interface Division  
Airman Systems Directorate  
711<sup>th</sup> Human Performance Wing  
Air Force Research Laboratory

//signed//

LOUISE A. CARTER, Ph.D., DR-IV  
Chief, Warfighter Interface Division  
Airman Systems Directorate  
711<sup>th</sup> Human Performance Wing  
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

<b>REPORT DOCUMENTATION PAGE</b>			<i>Form Approved</i> OMB No. 0704-0188		
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to 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. <b>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</b>					
<b>1. REPORT DATE (DD-MM-YY)</b> 03-02-20		<b>2. REPORT TYPE</b> Interim		<b>3. DATES COVERED (From - To)</b> 10 Jan 18 – 03 Feb 20	
<b>4. TITLE AND SUBTITLE</b> Identification of Criterion Constructs and Measures for Joint-Service Enlisted Jobs				<b>5a. CONTRACT NUMBER</b> FA8650-14-D-6500	
				<b>5b. GRANT NUMBER</b> Not applicable	
				<b>5c. PROGRAM ELEMENT NUMBER</b> 62202F	
<b>6. AUTHOR(S)</b> *Matthew T. Allen *Teresa L. Russell *Laura A. Ford #Cristina D. Kirkendall ^Thomas R. Carretta				<b>5d. PROJECT NUMBER</b> 5329	
				<b>5e. TASK NUMBER</b> 07	
				<b>5f. WORK UNIT NUMBER</b> H0SA (532909TC)	
<b>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b> *Human Resources Research Organization / #Army Research Institute for the Behavioral & Social Sciences (HumRRO) Sciences 66 Canal Center Plaza, Suite 700 6000 Sixth Street Alexandria, VA 22314-1578 Fort Belvoir, MD 22060				<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b>	
<b>9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b> ^Air Force Research Laboratory 711 <sup>th</sup> Human Performance Wing Airman Systems Directorate Warfighter Interface Division Collaborative Interfaces and Teaming Branch Wright-Patterson AFB, OH 45433				<b>10. SPONSORING/MONITORING AGENCY ACRONYM(S)</b> 711 HPW/RHCC	
				<b>11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)</b> AFRL-RH-WP-TR-2020-0019	
<b>12. DISTRIBUTION/AVAILABILITY STATEMENT</b> DISTRIBUTION A. Approved for public release. Distribution unlimited.					
<b>13. SUPPLEMENTARY NOTES</b> Subcontract No. FPH02-S022, Prime Contract No. FA8650-14-D-6500), Report contains color; 88ABW-2020-1404, Cleared 16 Apr 2020					
<b>14. ABSTRACT</b> The Department of Defense (DoD) uses the Armed Services Vocational Aptitude Battery (ASVAB) to select over 100,000 new military recruits and place them in military occupations. Other predictors such as the Tailored Adaptive Personality Assessment System (TAPAS), interest inventories, and specialized tests to supplement the ASVAB are used to make personnel selection and assignment decisions. To ensure predictor measures are valid, the DoD and individual Services conduct rigorous, large-scale research projects to evaluate predictor measures against criterion metrics such as training/job performance or retention. However, criterion metrics are mostly Service-specific and sometimes occupation-specific, making it difficult to examine outcomes DoD-wide. This report describes recommendations for standardizing criterion measurement across the Services in order to (a) facilitate robust comparisons of results within and across the Services and (b) strengthen DoD's conclusions about the validity and utility of the ASVAB and other predictors. The Human Resources Research Organization (HumRRO) developed taxonomies of job performance, attitudes, and organizational outcomes for first term enlisted Service personnel; constructed a database of criterion measures used by the Services; linked criterion measures to the performance domain constructs; and made recommendations to develop a unified set of test evaluation criteria that can be used by all Services.					
<b>15. SUBJECT TERMS</b> Criterion measures, performance taxonomy, enlisted, joint-service, cross-service					
<b>16. SECURITY CLASSIFICATION OF:</b>			<b>17. LIMITATION OF ABSTRACT:</b> SAR	<b>18. NUMBER OF PAGES</b> 105	<b>19a. NAME OF RESPONSIBLE PERSON (Monitor)</b> Thomas R. Carretta <b>19b. TELEPHONE NUMBER (Include Area Code)</b>
<b>a. REPORT</b> Unclassified	<b>b. ABSTRACT</b> Unclassified	<b>c. THIS PAGE</b> Unclassified			

## TABLE OF CONTENTS

List of Figures .....	iv
List of Tables .....	iv
ACKNOWLEDGEMENTS .....	v
1.0 EXECUTIVE SUMMARY .....	1
1.1 Research Requirement.....	1
1.2 Procedures .....	1
1.3 Findings .....	2
13.1 Relevant, Generalizable Criterion Constructs .....	2
13.2 The Criterion Measure Database .....	2
13.3 Recommendations .....	2
1.4 Utilization and Dissemination of Findings.....	2
2.0 INTRODUCTION .....	3
2.1 Background .....	3
2.2 Research Objectives .....	3
3.0 DEVELOPMENT OF TAXONOMIC STRUCTURES FOR ORGANIZING CRITERION MEASURES .....	4
3.1 Defining the Job Performance Domain .....	4
3.1.1 Literature Review .....	4
3.1.2 Draft Performance Dimension Retranslation .....	5
3.1.3 Generalizability and Criticality of Performance Constructs.....	8
3.2 Defining the Attitudinal Domain.....	11
3.3 Defining Organizational Outcomes.....	12
3.4 Generalizability and Proximity of Attitudinal Constructs and Organizational Outcomes .....	14
4.0 DEVELOPMENT OF THE CRITERION INSTRUMENT DATABASE.....	15
4.1 Development of the Online Data Entry Tool .....	15
4.1.1 Procedures .....	15
4.1.2 Results .....	17
4.2 Collection of Criterion Instruments.....	17
4.2.1 Procedures .....	17
4.2.2 Results .....	19
4.3 Criterion Instrument Mapping (Gap Analysis) .....	19
4.3.1 Procedures .....	20

432	Results .....	20
4.4	Finalizing the Database .....	24
441.	Procedures .....	24
442	Results .....	25
4.5	Estimating Psychometric Quality and Feasibility of Measurement Methods .....	25
451	Procedures .....	25
452	Results .....	27
4.6	Criterion Measure Scores .....	27
5.0	RECOMMENDATIONS .....	29
5.1	Use and Maintenance of Research Products .....	29
5.1.1	Use of the Criterion Constructs .....	29
5.1.2	Use and Maintenance of the Criterion Instrument Database.....	29
5.2	Criterion Measurement Recommendations Overview .....	30
5.3	Tier 1 - Maximize Use of Administrative Data.....	31
531	Tier 1 Overview.....	31
532	Tier 1 Recommendations.....	32
533	Tier 1 Summary.....	36
5.4	Tier 2 Improve Measurement of Attitudes and Outcome Variables .....	36
541	Tier 2 Overview.....	36
542	Tier 2 Recommendations.....	37
543	Tier 2 Summary.....	40
5.5	Tier 3 Improve Measurement of Job Performance.....	41
551	Tier 3 Overview.....	41
552	Tier 3 Recommendations.....	42
553	Tier 3 Summary.....	48
6.0	SUMMARY AND CONCLUSIONS .....	50
7.0	REFERENCES .....	52
8.0	LIST OF ACRONYMS .....	60
APPENDIX A.	Retranslation Survey.....	62
APPENDIX B.	Retranslation Results .....	66
APPENDIX C.	Final Performance Construct Definitions .....	68
APPENDIX D.	Attitudinal and Organizational Outcome Construct Definitions.....	72
APPENDIX E.	Data Entry Survey Tool .....	74
APPENDIX F.	Criterion Instruments Included in Mapping Exercise.....	90
APPENDIX G.	Sub-Dimension Results for Measurement Mapping.....	92

APPENDIX H. Measurement Method Full Ratings..... 97

## LIST OF FIGURES

Figure 1.	Online Performance Dimension Survey .....	9
Figure 2.	Screenshot from the Criterion Instrument Database Output .....	25
Figure 3.	Overview of Composite Development Approach.....	28
Figure 4.	Three-tiered Recommendations .....	31
Figure 5.	Criterion Domain Constructs Measured in Tier 1. ....	36
Figure 6.	Criterion Domain Constructs Measured in Tier 2. ....	41
Figure 7.	Criterion Domain Constructs Measured In Tier 3. ....	49

## LIST OF TABLES

Table 1.	Summary of Job Performance Taxonomic Literature Reviewed.....	5
Table 2	Hierarchical Trainee and 1st Term Performance Taxonomy .....	8
Table 3.	Job Performance Category and Subcategory Mean Relevance Ratings .....	11
Table 4.	Attitudinal Criterion Domain Taxonomy .....	12
Table 5.	Organizational Outcome Taxonomy .....	13
Table 6.	Number of Measures Mapped to Organizational Outcomes .....	20
Table 7.	Number of Measures Mapped to Attitudinal Constructs .....	21
Table 8.	Summary Statistics for Instruments Mapped to Job Performance Dimensions.....	23
Table 9.	Mean Psychometric Quality and Application Ease Ratings.....	27
Table 10.	Gaps in Job Performance Measurement Tiers 1 & 2.....	42

## **ACKNOWLEDGEMENTS**

We want to acknowledge the invaluable support of representatives from each Service and the Department of Defense (DoD) who shared their insights and expertise through their participation on the Criterion Measures Advisory Panel (CMAP): Mr. Tom Blanco, Dr. Eric Charles, Dr. Donna Duellberg, Lt Heidi Keiser, Commander Henry (Hank) Phillips, Capt Alex Ryan, Ms. Mary Ellen (Ellie) Stone, Dr. Sofiya Velgach, and Mr. Johnny Weissmuller.



## **1.0 EXECUTIVE SUMMARY**

### **1.1 Research Requirement**

The Department of Defense (DoD) uses the Armed Services Vocational Aptitude Battery (ASVAB) to select over 100,000 new military recruits every year and place them in military occupations. In recent years, the Services have begun using other selection and classification tools to supplement the ASVAB in making personnel selection and assignment decisions. To ensure those tools are valid, the DoD and individual Services conduct rigorous, large-scale research projects evaluating the tools against criterion metrics such as training/job performance or retention. Currently, criterion metrics and measures are mostly Service-specific and sometimes occupation-specific, making it difficult to examine outcomes DoD-wide. Standardizing criterion measurement across the Services would (a) facilitate robust comparisons of results within and across the Services and (b) strengthen DoD's conclusions about the validity and utility of the ASVAB and other predictors.

The purpose of the *Development of Criterion Measures for Evaluating Accession and Classification Testing* project was to develop a unified set of test evaluation criteria that can be used by all Services to conduct validation research.

### **1.2 Procedures**

Three primary questions guided our research procedures:

- (1) What criterion constructs (e.g., job performance, attitudes, outcomes) are relevant/important and generalizable across first-term enlisted occupations in all military Services?
- (2) What criterion metrics and instruments exist to measure those constructs?
- (3) What criterion measurement practices are recommended based on extant tools and research findings?

A panel of military accession experts drawn from members of the Manpower Accession Policy Working Group (MAPWG), known as the Criterion Measures Advisory Panel (CMAP), was formed to assist research staff in accessing information and provide input and advice.

An extensive literature review identified job performance constructs, job attitudes, and organizational outcomes that were well-grounded in research. The constructs were organized into a taxonomic structure based on the literature review. In turn, research personnel completed a survey to evaluate the taxonomy, and the taxonomy was revised based on the results. A second survey was constructed to determine the relevance and generalizability of the criterion constructs across first-term enlisted occupations in all military Services. Subject Matter Experts (SME) representing the Army, Navy, Air Force, and Marine Corps rated the relevance and criticality of the constructs. Statistical analyses showed that the SME judgments were highly reliable.

With help from the CMAP, research staff identified over 200 criterion measures used by the Services over the last 20 years. Staff developed an online data entry tool and populated a database describing key characteristics of those criterion measures. Staff mapped the criterion measure to the job performance, attitudinal, and organizational outcomes taxonomy. Research staff also estimated the psychometric quality of existing criterion measures based on prior research.

Analyses of the criterion measure database helped staff identify criterion constructs that are not well-measured with existing criterion metrics or instruments. Those analyses facilitated development of recommendations for measurement practices.

### **1.3 Findings**

#### **1.3.1 Relevant, Generalizable Criterion Constructs**

The criterion taxonomy developed in this project has three broad domains: (a) job performance (behaviors that are relevant to the Services' goals), (b) attitudes (such as commitment, satisfaction, career intentions), and (c) organizational outcomes such as reducing attrition and enhancing reenlistment.

The job performance taxonomy includes training and in-unit performance in the first term of enlistment. It is organized hierarchically with four performance categories at the highest level: (a) Technical Proficiency, (b) Organizational Citizenship & Peer Leadership, (c) Psychosocial Well-Being, and (d) Physical Performance. The lowest level of the performance taxonomy has 33 specific dimensions.

SME survey results indicated that the performance categories and dimensions within those categories are relevant and generalizable across Services, albeit with some differences in emphasis. Psychosocial Well-Being was consistently rated as the most critical performance category across the Services. The US Marine Corps (USMC) and US Army SMEs rated Physical Performance as the next most critical performance category while the US Navy (USN) and US Air Force (USAF) SMEs rated Organizational Citizenship & Peer Leadership performance as second highest. Regardless, it is important to note that all four of the broad performance categories were rated as critical within and across Services.

#### **1.3.2 The Criterion Measure Database**

The criterion measure database contains information on 226 criterion measures that have been used experimentally or operationally in military research since 1980 and are relevant to first-term enlisted occupations. For each criterion measure, the database (a) indicates what constructs are measured, (b) describes the content of the measure, (c) provides references/citations, and (d) summarizes psychometric properties.

#### **1.3.3 Recommendations**

Research staff developed recommendations for a unified set of test evaluation criteria that can be used by all Services based on (a) what constructs need to be measured and (b) extant measures. The recommendations for development of measures are summarized in three tiers. Tier 1 is the most basic—its measures address only a few facets of the criterion taxonomy. Tiers 2 and 3 add measures that require more extensive levels of effort. Each new tier expands measurement of the criterion domain and measurement quality.

### **1.4 Utilization and Dissemination of Findings**

The recommendations were presented to the CMAP for consideration. Once the CMAP provided feedback, we began the development of a new set of joint-service criterion measures. The criterion measures will cover as much of the job performance, attitudinal, and organizational outcome domains as possible, within the time and budget constraints of the project. The development of new measures is described in a follow-on report (Ford, Yu, Graves, Huber & Wilmot, 2020).

## **2.0 INTRODUCTION**

The objective of this project was to identify criterion constructs and measures for joint-service enlisted occupations. This report documents the first of two phases for the *Development of Criterion Measures for Evaluating Accession and Classification Testing* project.

### **2.1 Background**

The DoD uses the ASVAB to select over 100,000 new military recruits every year and place them in military occupations. In recent years, the Services have begun using other tools such as the Tailored Adaptive Personality Assessment System (TAPAS), interest inventories (e.g., Job Opportunities in the Navy [JOIN]; Air Force Work Interest Navigator [AF-WIN]), and specialized tests to supplement the ASVAB in making personnel selection and assignment decisions. To ensure those tools are valid, the DoD and individual services conduct rigorous, large-scale research projects evaluating the tools against criterion metrics such as training/job performance or retention. Currently, the criterion metrics and measures are mostly Service-specific and sometimes occupation-specific, making it difficult to examine outcomes DoD-wide. Standardizing criterion measurement across the Services would (a) facilitate robust comparisons of results within and across the Services and (b) strengthen DoD's conclusions about the validity and utility of the ASVAB and other predictors.

### **2.2 Research Objectives**

The purpose of the *Development of Criterion Measures for Evaluating Accession and Classification Testing* project is to develop a unified set of test evaluation criteria that can be used by all Services to conduct validation research. The objectives for the first phase were to:

- organize a panel of representatives from each Service component to serve as SME on a review/recommendation panel;
- develop a taxonomic structure for attitudinal, organizational outcome, and job performance domains for joint-Service, first-term enlisted personnel;
- document key features of criterion instruments currently in use and identify any measurement gaps and redundancies; and
- organize criterion instruments into a taxonomic structure of outcomes of interest (e.g., job performance).

A necessary component of accomplishing the above objectives was garnering input from representatives from each Service (Air Force, Army, Coast Guard, Marine Corps, and Navy). The CMAP was formed early in the project to provide support and feedback. The CMAP met regularly throughout the project via teleconference briefings and provided input through discussions and emails.

### **3.0 DEVELOPMENT OF TAXONOMIC STRUCTURES FOR ORGANIZING CRITERION MEASURES**

A taxonomic structure was developed for use in organizing criterion instruments, mapping criterion instruments onto the taxonomic structure, and consequently identifying measurement gaps – that is, mapping existing criteria measures to the constructs to identify those constructs that are not being measured.

The taxonomy included three criterion domains:

- *Job performance* – behaviors that are relevant to the Services’ goals and that can be scaled in terms of individuals’ proficiency (Campbell, McCloy, Oppler, & Sager, 1993). The taxonomy captures training and in-unit performance during the first term of enlistment. The taxonomy does not include occupation-specific behaviors.
- *Attitudes* – cognitions that are relevant to individuals’ job plans and performance (e.g., commitment, satisfaction, career intentions).
- *Organizational outcomes* – outcomes that are important to the Services at an organizational level, such as reducing attrition and enhancing reenlistment.

#### **3.1 Defining the Job Performance Domain**

The job performance taxonomy had three purposes. First and foremost, it was intended to describe the entire domain of early career, enlisted job performance, including performance in training and through the end of the first term. Second, it provided a structure for describing the content of criterion instruments, allowing comparison of the content of instruments. Third, the taxonomy was used to guide development of new criterion instruments (Ford, Yu, Graves, Huber, Russell, & Wilmot, 2020).

##### **3.1.1 Literature Review**

To develop the performance dimensions, we gathered and integrated literature describing performance taxonomies. As shown in Table 1, some of the taxonomies contained many constructs, others focused on a domain of constructs, and others were military specific. We placed definitions from all of the taxonomies into a spreadsheet.

The Campbell Model is the most extensively researched and documented of the taxonomies and early versions were developed from military work. Therefore, we used it as a scaffold, sorting dimensions from other categorization schemes into it and adjusting as needed to better capture dimensions. This process resulted in 33 draft dimension definitions organized into ten broader categories. Draft materials were reviewed by our contract monitors, CMAP members, Dr. Deirdre Knapp, and Dr. John Campbell.

**Table 1. Summary of Job Performance Taxonomic Literature Reviewed**

Target	Dimension Set	Key References
Many constructs	The Campbell Model	Campbell, 2012; Campbell, Hanson, & Oppler, 2001; Campbell, McCloy, Oppler, & Sager, 1993; Campbell & Wiernik, 2015
	The Great Eight	Bartram, 2005
	Model of Work Role Performance	Griffin, Neal, & Parker, 2007
	Attributes of Successful Leaders	Zaccaro, Laport, & Jose, 2012
Specific constructs	Teamwork	O’Shea, Goodwin, Driskell, Salas, & Ardison, 2009; Shuffler, Pavlas, & Salas, 2012
	Task Performance	Borman, Grossman, Bryant, & Dorio, 2017
	Adaptability	Pulakos, Arad, Donovan, & Plamondon, 2000
	Self-Directed/Active Learning	Garrison, 1997; Russell, Rosenthal, Paullin, & Putka, 2006
	Employee Engagement	Macey & Schneider, 2008
	Organizational Citizenship	Dorsey, Cortina, Allen, Waters, Green, & Luchman, 2017; Goffin, Woycheshin, Hoffman, & George, 2013; Organ, 1988
	Counterproductive Work Behavior	Dalal, 2005; Rotundo & Spector, 2017; Spector, Bauer, & Fox, 2010; Spector et al., 2006
	Ethical Performance	Russell, Sparks, Campbell, Ramsberger, Handy, & Grand, 2017
Military-specific constructs	Cross Cultural Performance	Klafehn, Anderson, Taylor, Ingerick, & Ford, 2018
	Combat Performance	Wasko, Owens, Campbell, & Russell, 2012
	Situational Awareness	Matthews, Eid, Johnsen, & Boe, 2011
	1 <sup>st</sup> term Performance	Campbell, Hanson, & Oppler, 2001; Sager, Russell, Campbell, & Ford, 2005
	Air Force-Wide Rating Dimensions	Lance, Teachout, & Donnelly, 1992
	Training Performance	Waugh & Russell, 2005

Note. Full citations appear in the Reference section.

### 3.1.2 Draft Performance Dimension Retranslation

To evaluate the dimensions and hierarchical structures for them, we asked 17 researchers with substantial experience in performance measurement and/or military criterion development to categorize the 33 dimensions according to (a) two categories (Can-do/Technical and Will-do/Contextual), (b) four categories, and (Technical Performance, Counterproductive Work Behavior, Citizenship & Peer Leadership, and Physical Performance, and (c) ten categories.

For the two-category judgment, we defined Can-do/Technical and Will-do/Contextual performance as follows, based on definitions from Borman and Motowidlo (1993) and Campbell and Knapp (2001).

- *Can-do/Technical Performance* – performance of activities that contribute to the organization’s technical core. Task activities usually vary between different jobs in the same organization. Technical task performance is usually predicted by knowledge, skills, and abilities. Technical task performance is role-prescribed, that is formally recognized

as part of the job. Can-do performance is typically measured using maximal performance instruments.

- *Will-do/Contextual Performance* – performance of activities that support the organizational, social, and psychological environment (e.g., organizational citizenship behaviors). Contextual activities are important across jobs. Motivational and personality characteristics are key determinants of contextual performance. Contextual activities may not be role-prescribed. Will-do performance is most often measured using measures of typical (as opposed to maximal) performance.

Borman and Motowidlo (1993) talked about specific performance dimensions as being saturated with either task or contextual elements. Dimensions can vary somewhat in terms of how reliant performance is on task or contextual elements. So, for example, some dimensions might be saturated with both task and contextual elements. Managerial dimensions that involve both planning and dealing with people might be the best example of dually-saturated dimensions.

Therefore, we asked raters to judge the Can-do/Technical vs. Will-do/Contextual saturation of each of the 33 specific dimensions using the following scale:

- 2 = Can-do/Technical task
- 1 = Can-do/Technical with some Will-do/Contextual saturation
- 0 = Equally Can-do/Technical and Will-do/Contextual
- +1 = Will-do/Contextual with some Can-do/Technical saturation
- +2 = Will-do/Contextual

For the 4-category solution, we provided definitions of four categories derived from the Campbell Model: (a) Technical Performance, (b) Counterproductive Work Behavior, (c) Citizenship & Peer Leadership, and (d) Physical Performance. Raters were asked to categorize each of the 33 dimensions into one of the four categories. Similarly, for the ten-category solution, we defined ten performance categories that had emerged as we organized dimensions from the literature review. Raters were asked to sort each specific performance dimension into one of the ten categories. Specific instructions given to raters appear in Appendix A, and Appendix B provides a table of results.

Statistics showed that consistency across raters on the judgments was high.

- *Can-do/Technical and Will-do/Contextual rating.* Reliability estimates suggested strong consistency across raters (Interclass correlations [ICC]<sup>1</sup> of .69 [single-rater] and .95 [all raters]).
- *4-Category judgments.* The percent of raters agreeing on the categorization of the specific 33 dimensions ranged from 47% to 100%, with a mean of 88%. For 14 of the 33 dimensions, 100% of the raters agreed on the 4-category grouping.
- *10-Category judgments.* The percent of raters agreeing on the categorization of the specific 33 dimensions ranged from 50% to 100%, with a mean of 86%. For six of the 33 dimensions, 100% of the raters agreed on the ten-category grouping.

Results of the rating exercise showed that a two-category structure did not work well. Raters had difficulty making judgments for dimensions that were thought to be related to both Can-do and

---

<sup>1</sup> Intraclass correlation coefficients.

Will-do performance, as evidenced by higher standard deviations in Can-do/Will-do judgments for some of the dimensions. The 4- and 10-category structures both worked well. However, the results suggested some fine tuning was needed.

After concluding that the 2-category taxonomy (Can-do/Will-do) was insufficient, we turned to the 4-category taxonomy. We placed the 33 specific dimensions into the 4-category structure based on where they were classified by the raters. Then, subcategories of the 4- category solution were created using the Can-do/Will-do and 10-category data to make refinements. Our goal was to create the four categories and dimensions such that they would be relatively homogeneous with regard to Can-do/Will-do saturation. For example, before the retranslation exercise, we had grouped the two safety consciousness dimensions and a dimension on well-being (stress adjustment) into a category we called health and safety. However, the safety and well-being dimensions were rated as having very different Can-do/Will-do saturations, with Can-do playing a much stronger role in safety consciousness than well-being. Safety Consciousness thus became a sub-category in the broader Technical Proficiency category. We moved well-being to a broader Psychosocial Well-being category along with Counterproductive Work Behaviors. The final four categories were:

- Technical Proficiency,
- Organizational Citizenship & Peer Leadership,
- Psychosocial Well-Being, and
- Physical Performance.

In the 10-category solution, two categories - Individual Work Responsibility and Health and Safety Conscientiousness - did not hold together well. We moved their constituent dimensions to other categories with similar categorizations by raters. We also broke out Decision Making, Problem Solving, and Innovation into its own dimension, and made minor wording changes to dimension titles based on rater feedback. The final version of the mid-level solution in the performance hierarchy has 12 subcategories.

*Final Fine-Tuning of Dimensions.* We mapped a large sample of criterion instruments to the taxonomy (as described in Section 4.3). In doing so, we held a consensus meeting of researchers who had done the mapping, and we identified a few areas needing clarification. Based on that discussion, we made final tweaks to the titles and definitions of a few performance constructs. Table 2 provides the resulting recommended performance taxonomy for First Term and Training Job Performance. The taxonomy has three levels that vary in breadth. The broadest level has four categories, and the next level has 12 subcategories. The specific level has 33 specific dimensions. The final performance construct definitions appear in Appendix C.

**Table 2. Hierarchical Trainee and 1st Term Performance Taxonomy.**

<b>Performance Category</b>	<b>Subcategory</b>	<b>Specific Dimension</b>
<b>A. Technical Performance</b>		
	A.1. Task Performance	Job-Specific Proficiency General Proficiency
	A.2. Decision Making, Problem Solving, and Innovation	Decision Making, Problem Solving, and Innovation
	A.3. Communication	Oral Communication Written Communication Nonverbal Communication
	A.4. Safety and Security Consciousness	Safety and Security Consciousness in Everyday Work Safety and Security Consciousness during Mission Operations
<b>B. Organizational Citizenship &amp; Peer Leadership</b>		
	B.1. Planning and Structuring Work	Providing Structure Teamwork Self-Management Learning/Training Self-Management
	B.2. Conscientious Initiative	Classroom Learning Self-Development Persistence Initiative
	B.3. Support for Peers	Helping Peers Cooperating Courtesy & Respect Accepting Differences Motivating Serving as a Model
	B.4. Organizational Support	Military Presence Selfless Service Support for the Organization Integrity/Moral Courage
<b>C. Psychosocial Well-Being</b>		
	C.1. Adapting to Stressful Situations	Adapting to Stressful Situations
	C.2. Counterproductive Work Behavior	Loafing and Tardiness Abusing Substances and Other Self-Destructive Behavior Bullying, Harassing, or Hurting Others Delinquency
<b>D. Physical Performance</b>		
	D.1. Physical Endurance	Physical Endurance
	D.2. Physical Fitness	Physical Fitness

### **3.1.3 Generalizability and Criticality of Performance Constructs**

Having developed a taxonomic structure that has a great deal of support based on research literature and input from military research staff, we wanted to assess the generalizability of the



constructs across the Services and the criticality of the performance constructs. Toward that end, we solicited experts from each branch of the Services to complete a survey to rate performance dimensions.

We sought experts who were not just knowledgeable of one occupation; we were looking for experts who had a good understanding of the non-technical performance requirements of many or all occupations in the Service. CMAP members were tasked to identify at least three experts who would meet the following criteria:

- Have at least five years of experience in their organization.
- Be broadly knowledgeable of their Service’s occupations.
- Be highly knowledgeable of their Service’s mission.
- Be able to respond to questions based, cumulatively, on their years of experience.

We created an online survey that asked the military experts to make the following two ratings for each of the 33 job performance dimensions.

- Importance across enlisted, first term occupations on a five-point rating scale ranging from “Not Important” to “Extremely Important”
- Criticality to the Service’s mission accomplishment on a five-point rating scale ranging from “Not at all Critical” to “Extremely Critical”

A screenshot of the online survey appears in Figure 1.

**Trainee and 1st Term Enlisted Servicemember Performance:  
A Survey of the Importance and Criticality of Performance Dimensions**

46% Completed

[Navigation Instructions](#) | [Continue Later](#)

1. Background   2. **Technical Performance**   3. Organizational Citizenship and Peer Leadership   4. Psychosocial Well-Being   5. Physical Performance

Please rate the importance and criticality of each element.

	How important is this performance element across enlisted, first-term occupations?					To what extent is successful enlisted, first-term performance in this element critical to your service's mission accomplishment?				
	Not important	Somewhat important	Important	Very important	Extremely important	Not at all critical	Somewhat critical	Critical	Highly critical	Extremely critical
a) <b>Job-Specific Proficiency</b> - Being able to perform job-specific tasks at the appropriate skill level.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
b) <b>General Proficiency</b> - Being able to perform service-wide tasks at the appropriate skill level (e.g., navigation in the Army and Marine Corps).	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
c) <b>Oral Communication</b> - Conveying information in a clear, understandable, organized manner when speaking.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Figure 1. Online Performance Dimension Survey.**

Our thinking was that the two judgments offered somewhat different information about the relevance of the performance dimension. Importance was intended to simply focus on the day-to-day performance requirement. In making the criticality judgement, however, experts could consider ongoing initiatives and future plans. We averaged the ratings for the two scales to create an overall relevance score.

Generalizability was defined in terms of the grand mean of the Services' mean relevance (i.e., importance and criticality) ratings. In other words, the grand mean across the Services was used as an indicator for the extent to which the performance requirement generalized across Services.

Twenty-six military experts responded to the survey. Analysis of their ratings revealed the following:

- Data from the US Army, USN, USAF, and USMC yielded high ICCs coefficients. ICCs (C,k) were .95 (US Army), .73 (USN), .98 (USAF), and .85 (USMC).
- Eight of 33 performance dimensions had average importance / criticality ratings<sup>2</sup> less than 3.0 (on a 5-point scale) in one or more of the Services, suggesting these subdimensions were not generalizable across Services. We dropped these eight subdimensions from further analysis: Nonverbal Communication, Written Communication, Learning/Training, Classroom Learning, Motivating, Serving as a Model, Military Presence, and Endurance. Many of these subdimensions were rated as being important for one or more of the Services, but they were not generalizable across all Services.
- Means across Services were computed for the remaining dimensions. Psychosocial Well-being subdimensions were rated most highly across the Services (i.e., they were the most generalizable). Technical Performance subdimensions were, on average, rated least highly (i.e., least generalizable). However, the grand mean importance / criticality differences among the latter three broad performance factors (Physical Performance, Organizational Citizenship & Peer Leadership, and Technical Performance) were very small.
- All four of the broad performance categories were relevant within and generalizable across the Services. As shown in Table 3, all four of the categories received high mean ratings within and across Services.

---

<sup>2</sup> Average importance / criticality was computed by first taking the average across ratings within service, then computing a unit-weighted average across services, and finally averaging importance and criticality metrics.

**Table 3. Job Performance Category and Subcategory Mean Relevance Ratings**

Job Performance Categories and Dimensions	Within Service Means				Across Services	
	Army	Navy	Air Force	Marine Corps	Grand Mean	SD
C. Psychosocial Well-Being	4.15	4.70	4.43	4.22	4.38	0.28
C.2. Counterproductive Work Behavior	4.24	4.88	4.56	4.24	4.48	0.32
C.1. Adapting to Stressful Situations	3.72	3.83	3.75	3.75	3.77	0.06
D. Physical Performance	3.83	3.67	3.00	4.11	3.65	0.44
D.1 Fitness	3.83	3.67	3.00	4.11	3.65	0.44
B. Organizational Citizenship & Peer Leadership	3.41	3.96	3.45	3.68	3.62	0.15
B.4. Organizational Support	3.41	4.17	4.00	3.81	3.85	0.40
B.2. Conscientious Initiative	3.31	4.17	3.04	3.73	3.56	0.59
B.1. Planning and Structuring Work	3.50	3.83	3.25	3.61	3.55	0.29
B.3. Support for Peers	3.42	3.67	3.50	3.56	3.54	0.13
A. Technical Performance	3.58	3.82	3.33	3.49	3.58	0.24
A.4. Safety Consciousness	3.94	4.17	3.63	3.78	3.88	0.27
A.1. Task Performance	3.86	3.63	3.44	3.68	3.65	0.24
A.2. Oral Communication	3.11	3.83	3.25	3.50	3.42	0.38
A.3. Decision Making, Problem Solving, and Innovation	3.39	3.67	3.00	3.00	3.35	0.33

*Notes.* Sample sizes were 9 Army, 3 Navy, 4 Air Force, and 10 Marines. ICCs ( $C,k$ ) were .95 (Army), .73 (Navy), .98 (Air Force), and .85 (Marines). The survey had 33 specific dimensions. Dimensions were retained only if they had mean ratings of at least 3.0 (on a 5-point scale) within each Service. Twenty-five dimensions were retained and eight were dropped. The means presented here include only those 25 generalizable dimensions.

While the sample size for the survey was small, the respondents were selected experts with Service-wide experience. We are, therefore, confident in these data. Even so, we recommend that the Services use the performance taxonomy in future job analytic work and continue to make judgments regarding its efficacy. Over time the importance of different performance dimensions can change. For example, we suggest that one reason that Technical Performance was, on average, the least important category, is because applicants are screened on the ASVAB and undergo rigorous technical training for their jobs. So, the Services are not likely to observe as many problems with Technical Performance in the field as they might with other dimensions for which applicants are not as directly selected or trained (e.g., Psychosocial Well-Being). Selection and training initiatives or even other external factors such as reduced or enhanced physical demands could change the priority of performance categories over time.

### 3.2 Defining the Attitudinal Domain

We reviewed research literature to identify attitudinal variables that have served as important criterion instruments in military and civilian research. This review led to five primary constructs defined in Table 4: (a) work satisfaction, (b) morale, (c) organizational commitment, (d) withdrawal cognitions/intentions, and (e) person-environment fit. Table D.1 in Appendix D provides full construct definitions.

**Table 4. Attitudinal Criterion Domain Taxonomy**

<b>Construct</b>	<b>Definition</b>	<b>Facets</b>
Work Satisfaction	An individual's satisfaction with work.	<ul style="list-style-type: none"> <li>- Whole job satisfaction</li> <li>- Job facet satisfaction</li> <li>- Career satisfaction</li> </ul>
Morale	A holistic judgment of one's own morale	
Organizational Commitment	An individual's psychological bond with the organization, as represented by an affective attachment to the organization, internalization of its values and goals, and a behavioral desire to put forth effort to support it.	<ul style="list-style-type: none"> <li>- Affective</li> <li>- Continuance</li> <li>- Normative</li> </ul>
Withdrawal Cognitions/Intentions	Thinking about or intending to quit one's job.	<ul style="list-style-type: none"> <li>- Attrition cognitions</li> <li>- Short-term active duty career continuance intentions</li> <li>- Long-term active duty career continuance intentions</li> <li>- Post-active duty plans</li> <li>- Deployment-attributed change in career intentions</li> </ul>
Person-Environment Fit (PE Fit)	Congruence between the individual's abilities, needs, and expectations and characteristics of the organization, job or group.	<ul style="list-style-type: none"> <li>- Person-Job, Needs-supplies fit</li> <li>- Person-job, Demands-abilities fit</li> <li>- Person-organization fit</li> <li>- Person-team fit</li> </ul>

*Note.* Based primarily on Allen, Knapp, & Owens (2016); Arthur, Bell, Villado, & Doverspike (2006); Cable & Edwards (2004); Cable & Judge (1997); Dawis & Lofquist (1984); Edwards (1996); Greenhaus, Parasuraman & Wormley (1990); Hom (2011); Hom, Lee, Shaw, & Hausknecht (2017); Judge, Cable, Boudreau, & Bretz (1994); Judge & Kammeyer-Mueller (2012); Judge, Weiss, Kammeyer-Mueller, & Husin (2017); Meyer & Allen (1991); Meyer, Kam, Goldenberg, & Bremner (2013); and Weiss, Dawis, Lofquist, & England (1966).

### 3.3 Defining Organizational Outcomes

An important organizational effectiveness concept for the military is readiness, “defined as the ability of individual units in the armed forces to execute their assigned missions promptly and competently” (O’Hanlon, 2017, p. 1). Readiness is a complex topic with many facets, and our focus is only on human resources. From a human resources standpoint, readiness means having personnel with the experience, training, skill, and aptitude needed to accomplish missions (DoD, 2018; Forrester, O’Hanlon, & Zenko, 2001). We examined military literature to identify a draft list of outcome variables that are indicative of experience, training, skill (and consequently facilitators of military readiness). Next, the CMAP reviewed the draft list and identified organizational outcomes that are important for each of the Services. We also reviewed reports provided by the CMAP and identified in our own literature review. Table 5 provides the resulting outcome taxonomy. Table D.2 provides full construct definitions.

**Table 5. Organizational Outcome Taxonomy**

<b>Outcome Construct</b>	<b>Facet</b>	<b>Example Indicators</b>
Attrition	Delayed entry program (DEP)	<ul style="list-style-type: none"> <li>• DEP attrition</li> </ul>
	Boot Camp	<ul style="list-style-type: none"> <li>• Attrition from boot camp</li> </ul>
	Advanced Training	<ul style="list-style-type: none"> <li>• Attrition from advanced training</li> </ul>
	In-unit	<ul style="list-style-type: none"> <li>• Attrition in-unit (premature attrition)</li> </ul>
	Re-enlistment	<ul style="list-style-type: none"> <li>• Re-enlistment for second term; propensity to re-enlist</li> </ul>
Reprimands	Reprimands	<ul style="list-style-type: none"> <li>• Articles 15/ reprimands</li> </ul>
Experience	Tenure	<ul style="list-style-type: none"> <li>- Time in grade/rank</li> <li>- Time in uniform/Length of service</li> </ul>
	Rank	<ul style="list-style-type: none"> <li>• Rank</li> </ul>
Initiative	Awards	<ul style="list-style-type: none"> <li>• Merit-based awards and commendations</li> </ul>
Performance	Advanced Training	<ul style="list-style-type: none"> <li>- Training school grades</li> <li>- Pass/Fail</li> <li>- Rank in class</li> <li>- Training recycles/Wash-backs</li> </ul>
	In-unit	<ul style="list-style-type: none"> <li>- Supervisor performance ratings/ Enlisted Performance Ratings (EPR in the USAF)/ Proficiency marks (PRO marks [USMC])</li> <li>- Job knowledge test scores (e.g., USAF Skill/Knowledge Test [SKT]; USAF Promotion Fitness Examination [PFE])</li> </ul>
	Skill Upgrading	<ul style="list-style-type: none"> <li>• Skill level attainment (e.g., USAF skill level badges)</li> </ul>
	Promotion Potential	<ul style="list-style-type: none"> <li>• Promotion exam scores</li> </ul>
	Physical	<ul style="list-style-type: none"> <li>• Current physical fitness</li> </ul>
	Qualifications	<ul style="list-style-type: none"> <li>- Rifle/pistol qualification score</li> <li>- Other qualifications (swim, brown belt, Ranger)</li> </ul>
	Re-enlistment Eligibility	<ul style="list-style-type: none"> <li>• Computed Tier Score (re-enlistment eligibility composite based on a number of qualifications)</li> </ul>
Productivity	Skilled Tenure	<ul style="list-style-type: none"> <li>• Qualified man months ([QMM] - number of months in service at qualified level based on skills test)</li> </ul>
	Skilled Tenure	<ul style="list-style-type: none"> <li>• Months mission ready service (months of service at the highest skill level)</li> </ul>
	Quantity of Performance	<ul style="list-style-type: none"> <li>• Productive capacity (rate of task performance)</li> </ul>
Promotion	Rate	Promotion rate (a deviation score comparing to other Service members with the same time in service and in the same job)
	Time	Promotion time to E-4

*Note.* Indicators were drawn primarily from Alley, Pacheco, Birkelbach, Schwartz & Weissmuller (2007); Campbell & Knapp (2001); Halper, Goodman, & Alley (2010); Ingerick, Allen, Weaver, Caramagno, & Hooper (2006); Knapp & Campbell, 1993, Mayberry (1990); Sims & Hiatt (2001); and Wathen (2014).

### **3.4 Generalizability and Proximity of Attitudinal Constructs and Organizational Outcomes**

Attitudinal measures provide a surrogate measure for organizational outcomes such as separation and attrition. Separation data can take years to mature sufficiently and are not available for use in a validation study until new recruits reach specific milestones (e.g., graduation from basic or advanced training). Also, separation and attrition data tend to have low base rates. For those reasons, it is useful to identify measures that can serve as near-term surrogates for attrition measurement. Attitudes can also serve as proximal predictors of other outcomes, such as job performance (HumRRO, 2018).

We prioritized the attitudinal and outcome constructs to be measured based on their (a) proximity to either attrition or job performance and (b) generalizability across jobs and occupations. Three staff members made judgments independently and met to reach consensus on their judgments.

Attrition is critical across Services and clearly warrants the highest priority to measure. Withdrawal cognitions is highly proximal to attrition and warrants high priority. Satisfaction, morale and commitment are less proximal. They are expected to predict attrition, withdrawal cognitions, and some aspects of performance. They are slightly lower in priority for measurement.

Performance records, merit-based awards, and reprimands were deemed to be highly proximal, generalizable measures of job performance. Therefore, they are high in priority for measurement.

## 4.0 DEVELOPMENT OF THE CRITERION INSTRUMENT DATABASE

The purpose of this section is to describe the activities undertaken over the course of the project to address the following research objectives described in Section 2.2:

- Document key features of criterion instruments currently in use and identify any measurement gaps and redundancies.
- Organize criterion instruments into a taxonomic structure of outcomes of interest (e.g., job performance).

To fulfill these objectives, we completed the following tasks, which are described more fully in the remainder of this section:

1. *Developed an online data entry tool* (Section 4.1). The purpose of this tool was to provide a centralized repository for collecting and maintaining information about criterion instruments. This task involved identifying and refining metadata elements and programming those elements into an online tool.
2. *Identified criterion measures used by individual Services to validate new predictor measures* (Section 4.2). This task involved working with the CMAP to identify criterion measures currently or previously used by the Services, supplemented by input from in-house researchers experienced with military validation studies. We further supplemented this task with exploratory criteria from the academic literature.
3. *Mapped criterion measures to the job performance, attitudinal, and/or organizational outcomes taxonomy, as appropriate* (Section 4.3). To accomplish the objective, we reviewed source material about criterion measures recently used by the Services to validate predictor measures, and mapped them to elements of the job performance, attitudinal, and/or organizational taxonomy described in Section 3.0. This exercise was used to identify gaps in the measures currently used by DoD components.
4. *Populated data entry tool with criterion instrument information* (Section 4.4). This task involved researchers entering metadata elements for 226 criterion measures into an online tool. Quality assurance steps were also taken to ensure uniformity across criteria and to re-check taxonomic mapping.
5. *Applied output from the data entry tool to estimate the psychometric quality and feasibility of existing criterion instruments* (Section 4.5). To accomplish this, we combined ratings of various measurement approaches (e.g., performance ratings, attitudinal surveys) with construct relevance ratings (see Table 3) to determine the criterion instruments most likely to hold promise for future research.

### 4.1 Development of the Online Data Entry Tool

#### 4.1.1 Procedures

To develop the online data entry tool, we first identified the metadata elements relevant to the Services. For the purposes of the current study, “metadata” refers to data that describes other data - that is, it provides a description or context to other data. To identify metadata elements, we began with a high-level review of DoD projects that involved significant criterion development efforts. Specifically, we began by reviewing the following resources:

- The project “Building a joint-service classification research roadmap” (Knapp & Campbell, 1993). One of the purposes of this project was to create a taxonomy of research criteria.
- “Performance Measures for the 21st Century (PerformM21),” a project to develop an effective, affordable, Soldier assessment system, resulting in prototype assessments targeted to all Army Soldiers eligible for promotion to Sergeant (Knapp & Campbell, 2004, 2006; Moriarty & Knapp, 2007).

Based on these sources of information and input from our researchers, we drafted an initial list of metadata elements for CMAP review. The CMAP had the opportunity to review the initial draft of metadata elements in a March 2018 meeting. Most of the suggested revisions were minor (e.g., wording). Once these edits were complete, we moved on to pilot testing the metadata elements.

Inputs for applying the metadata elements to criterion measures of interest generally came from one of two types of resources: (a) technical documentation, such as technical reports or operational presentations internal to a DoD Service, or (b) academic documentation, such as journal articles or conference presentations (see Section 4.2.1. for details). Pilot testing of the metadata elements was carried out in three ways. First, two staff researchers practiced coding several instruments from technical reports/operational presentations independently and made adjustments to the metadata elements based on difficulties encountered. Second, four staff researchers then practiced applying the metadata elements to criterion instruments identified from academic resources. Specifically, the researchers coded two academic articles, then discussed difficulties in applying the metadata elements to those resources. Finally, project sponsors from the U.S. Army and the USAF reviewed the metadata elements and provided additional suggestions for changes. At each pilot phase, the metadata elements were updated, refined, or clarified as appropriate.

Additional adjustments to the metadata elements were made based on new information uncovered during the activities described in Sections 4.2 and 4.3. In general, the tool was designed for maximum flexibility in describing the characteristics of a wide variety of tools. For example, one identified criterion instrument was the Initial Military Training Army Life Questionnaire (IMT ALQ). The IMT ALQ is a self-report measure that assesses both attitudinal and performance constructs using several scales (e.g., open-ended responses, Likert scales, frequency counts), and has been refined and adjusted over roughly the last 15 years. The IMT ALQ stands in contrast to criterion measures gleaned from administrative records, such as the USAF’s Months of Mission-Ready Service (MM-RS). MM-RS is determined through a weighted combination of months spent at different Service levels (as determined by level upgrades, typically fulfilled through training requirements). We designed flexibility in the metadata elements to accommodate these disparate criteria through the following mechanisms:

- Generally relying on items that allow researchers to “select all that apply” rather than selecting one option,
- Liberal use of “free response” text boxes and “Other (please specify)” options within metadata elements, and
- Clear descriptions of what is sought by each element, and how to respond under different conditions.



## 4.1.2 Results

Once the metadata elements were identified, we designed and built a web-based survey to populate the database. The survey is designed to “live on” past the life of the project such that users will be able to enter new criteria as discovered / developed to populate a database. The database is easily exportable and helps to ensure that the elements are input in the database consistently. The final metadata elements as presented in the data entry tool are provided in Appendix E.

## 4.2 Collection of Criterion Instruments

### 4.2.1 Procedures

The United States Military has been conducting personnel assessment research since the advent of the Army Alpha and Army Beta tests in the early 20<sup>th</sup> century (Sellman, Russell, & Strickland, 2017). Given this history, to comprehensively examine all potential criteria would quickly become unmanageable and result in a lot of noise not useful to the ultimate purpose of the effort. Based on internal deliberations with CMAP members, we set the following boundary conditions for this task:

1. ***Joint-Service measures.*** Our search was limited to criteria that could potentially be used across Services. This excluded job-specific training and performance criteria but includes Service-specific criteria. For example, the Army developed Army-wide job knowledge tests for validating cognitively-oriented predictor measures (e.g., Knapp & Tremble, 2007). While Service-specific, the content is general enough that some sections of the measure could be adapted to other Services.
2. ***Developed 1980 or later.*** In 1980, individual Services undertook Joint Performance Measurement (JPM) projects for all enlisted ranks (Wigdor & Green, 1991). Criterion measures were developed for all Services as part of these projects, representing the last cross-DoD initiative to develop performance criteria. Thus, 1980 served as a useful benchmark for limiting our search.
3. ***First term enlisted outcomes.*** We limited our search to outcomes related to personnel in their first term of enlistment. This limitation excluded criteria developed for officer and non-commissioned officer populations. We considered training outcomes; however, given that training outcomes tend to be job-specific, we limited consideration of those outcomes to generalized measures (e.g., overall pass/fail rates).
4. ***Include Active Duty, Reserve, and National Guard.*** Our search considered criteria developed for all Service components.
5. ***Operational emphasis.*** Based on feedback from the CMAP, our search emphasized operational criteria that can be gleaned from administrative records and implemented across the Services. That said, anticipating that these operational criteria would not cover the entire criterion space identified in Section 3.0, we did not exclude criteria that were developed for research purposes only.

Based on the above parameters, we initially focused our search on what we referred to loosely as “operational” criteria. Operational criteria are those that are used by stakeholders for decision-making, such as whether a selection measure should be used or not. These criteria were gleaned primarily from the Services themselves through the CMAP and supplemented by additional

searching online through resources such as the Defense Technical Information Center (DTIC).<sup>3</sup> We supplemented this information with a broad-based search for additional criterion measures that we referred to as “exploratory” criteria. These search and processing procedures are described in more detail below.

***Search, Review, and Processing of Key Criteria.*** Our first task was to identify operational criteria used by the Services. To accomplish this, we started by asking the CMAP for leads. These leads included (a) technical reports or presentations that used operational criteria, (b) the names of other individuals that knew about specific criterion measures, and/or (c) copies of the measures themselves. For the written resources (e.g., technical reports), we triaged the provided resources by (a) determining relevance to the current project, (b) pulling out the titles of relevant potential criteria, and (c) providing a recommendation for further pursuit. To the third activity, we used the resources provided as a starting point to search for additional potential resources online. For example, a report may use a particular criterion measure for validation, but the details of its construction are provided in a separate report. In these cases, the researcher would seek out the relevant technical report to supplement the information from the original resource as needed. We also searched these reports and other resources (e.g., Service policies) for details on the construction of the criterion measure (in the case of administrative criteria) or copies of the measures themselves.

We also supplemented the information provided by the CMAP with criteria from known projects. For example, one staff researcher was aware of a project to identify combat performance rating scales used by the U.S. Army. Projects such as these went through the same triaging procedure described above. Once a relevant criterion measure was identified, two senior researchers, each with a Ph.D. in Industrial and Organizational (I-O) psychology and over ten years of applied experience working with the DoD, mapped information from the criterion measures to the metadata elements. Before finalizing the list of criteria, we also reviewed a by-Service list of criteria with individual CMAP member and added criterion measures if they felt there were gaps.

While the processing of operational criteria represented the primary charge of the current effort, we recognized early on that there were likely to be gaps in these measures in terms of both content (i.e., coverage of the taxonomies described in Section 3.0) and method (i.e., some methods are less likely to be relied upon than others). Therefore, we supplemented the operational criteria with criteria from academic literature. We focused our efforts on two military-specific academic resources—*Military Psychology*, the official journal of Division 19 (Society for Military Psychology) of the American Psychological Association, and the conference proceedings of the *International Military Testing Association (IMTA)*. Both of these resources focus on human capital research in the military, and publish regularly on personnel assessment and selection issues.

The first step in identifying criteria from these two resources was sourcing papers likely to contain potential instruments of interest. For *Military Psychology*, we reviewed all article abstracts from 2013 to 2017 and identified those with the potential criterion measures (e.g., the article was about training or personnel selection). For IMTA, we reviewed abstracts for all conference presentations from 2000 to 2017. From this process, we identified 80 papers with potential criteria of interest. Three staff researchers with post-graduate degrees in I-O

---

<sup>3</sup> <https://discover.dtic.mil/>

psychology were trained to (a) identify criteria from each of the academic resources and (b) map metadata elements onto those criteria. The training was accomplished by practicing these two activities on three academic resources, then coming together for a discussion of key differences in interpretation.

Similar to the operational criteria process, these three researchers were each assigned a set of papers to review, identify potential criteria of interest, and map the metadata elements. Unlike the operational criteria however, these researchers did not seek additional resources describing each criterion measure—all of the metadata information was derived from the original source article.<sup>4</sup>

#### **4.2.2 Results**

In total, 97 operational criteria were identified for inclusion in the final database, and an additional 129 exploratory criterion measures through the above process. A complete list of these measures is provided in a memorandum for record (HumRRO, 2018).

#### **4.3 Criterion Instrument Mapping (Gap Analysis)**

As described at the start of Section 4.0, one goal of this research was to identify measurement gaps that appear when operational criterion instruments are mapped against the three taxonomies (job performance, attitudinal, organizational outcome) described in Section 3.0. While all the criterion instruments collected in Section 4.2 were initially mapped to these taxonomies as part of the data collection activities, proper mapping of core criteria to this taxonomy is a critical outcome of this research. Thus, we undertook a separate criterion mapping task for a subset of operational criteria.

Specifically, we wanted the gap analysis to focus on instruments that are currently used by the Services. A subset of 74 instruments/variables were identified that had been used in validity studies by one or more Services within the last 20 years or are currently in operational use. The 74 instruments/variables can be organized into four measurement methods as follows:

1. Performance rating scales ( $k = 13$ ), including peer, self, or supervisor rating scales.
2. Performance instruments ( $k = 13$ ), including physical performance tests, low-fidelity simulations, interviews, and job knowledge tests.
3. Attitudinal surveys ( $k = 20$ ), including surveys conducted routinely by the Services as well as surveys used to collect criterion measures in selection and classification research, and
4. Variables retained in or computed from administrative data ( $k = 28$ ) such as various types of attrition and training and in-unit performance records.

The names of the 74 criterion instruments/variables are listed in Appendix F.

---

<sup>4</sup> Another key difference with the exploratory criteria was the sourcing guidance. A criterion instrument was included if it could be adapted as a criterion instrument for first term enlisted service-members. This means that it tends to include a broader range of criterion types, such as criteria used to evaluate officers and service members from other countries. The operational criteria, by contrast, focused exclusively on criteria that met the parameters described previously.

### 4.3.1 Procedures

Highly experienced military selection and classification researchers, each with a Ph.D. in I-O psychology or a related field and more than ten years of operational military research experience, mapped the 74 instruments/variables to the 33 subdimensions in the performance taxonomy as well as the constructs in the attitudinal and outcomes taxonomies. We prepared a library of literature and instruments and a response spreadsheet for use in the mapping exercise. Researchers were instructed to review each instrument and record a “1” in a response spreadsheet if they believed the instrument’s content mapped to the construct definition.

Two researchers mapped administrative variables to the taxonomies and three researchers mapped performance rating scales, performance instruments, and attitudinal surveys to the taxonomies. We analyzed the data and held a consensus meeting to discuss areas of disagreement. After the consensus meeting, researchers re-examined their ratings and re-submitted them. We analyzed the final ratings data. In the end, we determined an instrument to be mapped to a construct if it was mapped by at least two researchers.

### 4.3.2 Results

**Organizational Outcomes.** Table 6 provides the number of instruments (by measurement method) that mapped to outcome constructs. As shown, outcome constructs are most often captured by administrative data. The Services’ physical performance tests (categorized as performance instruments) are recorded in administrative records, as reflected by the number of performance instruments listed for performance record outcomes. Attrition and performance records were the most frequently measured outcomes. Productivity, reprimands, experience, and awards were measured less frequently.

**Attitudinal Constructs.** Table 7 provides the number of instruments (by method) that mapped to attitudinal domain constructs. As would be expected, attitudinal domain constructs are measured primarily by attitudinal surveys. Withdrawal cognitions and work satisfaction are the most frequently measured constructs. Attitudinal surveys also measured morale and organizational commitment with some frequency. Person-environment fit was not measured as often as other attitudinal constructs, probably because fit, by itself, is not a particularly definitive construct. Fit is thought to predict work satisfaction and, more distally, attrition. Thus, it is less directly useful in validity studies if attrition or withdrawal cognitions/intentions data are available.

**Table 6. Number of Measures Mapped to Organizational Outcomes**

Outcome	Definition	Performance			
		Rating Scales (k=13)	Performance Instruments (k =13)	Attitudinal Surveys (k =20)	Administrative Data (k =28)
Attrition	Voluntary or involuntary separation from a Service, which may occur during a term of Service or after (e.g., re-enlistment).	0	0	0	9
Reprimands	Records of formal disciplinary action against a Service member.	0	0	1	0
Experience	Indices such as time in grade, rank, or specific experiences (e.g., deployments).	0	0	0	4

Performance Records	Training outcomes (e.g., school grades, class rank, pass/fail a course), performance evaluation scores, or qualifications (e.g., physical fitness qualification, weapon qualifications).	1	8	1	12
Promotion	Promotion outcomes include rate of promotion, time to be promoted, etc.	0	0	1	2
Awards	Merit-based awards and commendations.	0	0	0	2
Productivity	Measures of the quantity or overall amount of work or qualified time at work.	0	0	0	1
Other	Criteria derived from administrative records that don't fit into one or more of the above categories	0	0	0	0

**Table 7. Number of Measures Mapped to Attitudinal Constructs**

Attitudinal Construct	Definition	Performance			
		Rating Scales ( <i>k</i> =13)	Performance Measures ( <i>k</i> =13)	Attitudinal Measures ( <i>k</i> =20)	Administrative Data ( <i>k</i> =28)
Work Satisfaction	An individual's satisfaction with work (whole job, job facets, career)	0	0	11	0
Morale	A holistic judgment of one's own morale	0	0	8	0
Organizational Commitment	An individual's psychological bond with the organization, as represented by an affective attachment to the organization, internalization of its values and goals, and a behavioral desire to put forth effort to support it.	0	0	8	0
Withdrawal Cognitions/ Intentions	Thinking about or intending to quit one's job	0	0	12	0
Person-Environment Fit	Congruence between the individual's abilities, needs, and expectations and characteristics of the organization, job or group.	0	0	5	0

**Job Performance Constructs.** Recall from Section 3.0 that the job performance taxonomy is hierarchical. At the highest level, there are four performance categories. The middle level has 12 performance dimensions, and 33 subdimensions are nested within the performance dimensions. Staff researchers participating in the rating exercise completed mapping at the subdimension level of the taxonomy (i.e., the finest level of granularity). Due to this nesting structure, we computed two statistics to summarize the mapping of instruments onto the higher-order performance categories and dimensions (see results in Table 8).

- **C%**, Coverage Percent, is the percentage of criterion instruments that mapped to at least one subdimension. For example, 100% Coverage for A. Technical Performance Constructs for the Performance Rating Scales indicates that all of the instruments in that category mapped to at least one of the subdimensions.
- **S%**, Saturation Percent, indicates how thoroughly, on average, each criterion instrument addressed the performance category or dimension. S% (for performance dimensions) is

the percent of subdimensions mapped to criterion instruments. For example, 58% for A.1. Task Performance indicates that, on average across Performance Rating Scales criteria, that percentage of subdimensions is assessed. S% (for performance categories) is the average of the performance dimension saturations for that performance category. For example, 49% for A. Technical Performance Constructs is the average of the saturation values for the four dimensions in this performance category (i.e., A.1. Task Performance; A.2. Communication; A.3. Decision Making, Problem Solving, and Innovation; and A.4. Safety and Security Conscientiousness).

Closer inspection of Table 8 reveals that three of the four categories (i.e., Technical Performance, Organizational Citizenship & Peer Leadership, and Physical Performance) were typically mapped to rating scales. Psychosocial Well-being was less frequently mapped. In 2011, a special issue of *American Psychologist* was devoted to well-being in military occupations (e.g., Cornum, Matthews, & Seligman, 2011). Lack of attention to well-being issues was one concern threaded throughout the special issue.

Considering the measurement methods, the performance categories were more frequently mapped to performance rating scales than other measurement methods. The performance instruments we reviewed tended to measure Technical Performance (e.g., job knowledge tests) and Physical Performance (e.g., the Services' physical performance tests). The performance instruments tended to be narrowly focused, aiming only at a specific subdimension. Therefore, their saturation indices were typically low. As shown by the small percentages for attitudinal surveys, some attitudinal surveys also contained self-reported performance items that mapped to the performance categories.

**Table 8. Summary Statistics for Instruments Mapped to Job Performance Dimensions**

Performance Category / Dimension	No. Dimensions / Subdimensions	Performance Rating Scales (k=13)		Performance Tests (k=13)		Attitudinal Surveys (k=20)		Administrative Data (k=28)	
		C%	S%	C%	S%	C%	S%	C%	S%
A. Technical Performance Constructs	4	100	49	50	11	10	2	0	0
A.1. Task Performance	2	85	58	36	18	0	0	0	0
A.2. Communication	3	69	38	7	5	5	3	0	0
A.3. Decision Making, Problem Solving, and Innovation	1	77	77	21	21	5	5	0	0
A.4. Safety and Security Conscientiousness	2	38	23	0	0	0	0	0	0
B. Organizational Citizenship and Peer Leadership	4	100	50	21	10	24	8	0	0
B.1.Planning and Structuring Work	4	92	44	21	14	24	12	0	0
B.2.Conscientious Initiative	4	92	56	14	7	24	8	0	0
B.3.Support for Peers	6	69	49	21	14	24	6	0	0
B.4.Organizational Support	4	77	50	7	5	19	6	0	0
C. Psychosocial Well-Being	2	69	36	14	7	14	6	0	0
C.1.Adapting to Stressful Situations	1	62	62	14	14	5	5	0	0
C.2.Counterproductive Work Behavior	4	23	10	0	0	10	7	0	0
D. Physical Performance	2	85	50	50	29	10	5	0	0
D.1. Physical Fitness	1	69	69	50	50	10	10	0	0
D.2. Physical Endurance	1	31	31	7	7	0	0	0	0

*Note.* C%, percent coverage, is the percentage of criterion measures that mapped to at least one subdimension. For example, 100% Coverage indicates that all of the criterion measures mapped to at least one subdimension.

S%, percent saturation (for performance dimensions), is the average percentage of subdimensions mapped to criterion instruments.

S%, percent saturation (for performance categories), is the average of the performance dimension saturations for that performance category.

The performance rating scales also addressed most of the performance dimensions. Performance dimensions that tended *not* to be well-covered by any of the measurement methods were Safety and Security Consciousness, Physical Endurance, and Counterproductive Work Behavior. It is possible that Physical Endurance is not a common requirement across all enlisted military occupations and hence was measured less frequently than Physical Fitness.

Finally, the subdimensions results appear in Appendix G. Most subdimensions were linked to specific measures except for the following:

- Nonverbal communication
- Safety and security consciousness subdimensions
- Training subdimensions
  - Learning/Training Self-Management
  - Classroom Learning
- Accepting differences
- Motivating others
- Counterproductive work behavior subdimensions
- Physical Endurance

Some dimensions may be more important for specific occupations and therefore may not appear in instruments designed for Service-wide use. We already mentioned this as a possible explanation for the low coverage percentages for endurance. The taxonomy includes two subdimensions (i.e., learning/training self-management and classroom learning) that are intended to capture behaviors that are only relevant during training courses. Those subdimensions have lower frequencies, as would be expected, because they mapped only to criterion instruments designed for use during basic or advanced training.

Consequently, it appears that some important/critical dimensions simply have not typically been included in many previous criterion measures. For example, Counterproductive Work Behaviors and Safety/Security Consciousness were both found to be highly important and critical across Services yet are underrepresented in the analysis of previously used instruments. This suggests cross-Service batteries of criterion measures should be supplemented with measures of these dimensions/subdimensions.

## **4.4 Finalizing the Database**

### **4.4.1. Procedures**

Once the above activities—online data entry tool development, criterion measure processing, criterion measure mapping—were complete, the metadata elements for each criterion measure were entered into the online tool. We took the opportunity as part of this data entry task to also conduct a final quality assurance check before the data were entered. Specifically, the two researchers performing the data entry (a) identified differences in how individual researchers entered data and applied rules to ensure uniformity across criteria and (b) re-checked the taxonomic mapping for the exploratory criteria (based on the outcomes of the activities described in Section 4.3). Final minor adjustments were made to the metadata elements in the online tool based on these quality assurance checks.



## 4.4.2. Results

Combining both the operational and exploratory instruments, a total of 226 criterion measures were entered into the final database. Results were output and formatted in an Excel document that summarizes all available information on these instruments, illustrated in Figure 2. A data entry guide was developed for future use by researchers who may want to add criterion measures to the tool in the future. The guide includes general guidelines, such as how to navigate the tool, as well as guidelines for specific items.

No.	Criterion Instrument	Content Definition	Subcales/Dimensions	Hyperlink to DOD Criterion Survey (copy and paste the link into your browser)
1	Appraisal of Cross-Cultural Competence	"the knowledge, skills, abilities, and other characteristics that enable learning and adapting to unfamiliar cultures (Abbe, Guick, & Herman, 2008)."	Perspective taking, Organizational awareness, Cultural knowledge, Communication skills, Interpersonal skills, Adaptability	<a href="https://apps.humvro.org/platform/?accessCode=ts0X50E">https://apps.humvro.org/platform/?accessCode=ts0X50E</a>
2	Work Cynicism	"Situational cynicism has a developmental component. Situational cynicism is amenable to change."	Not Applicable	<a href="https://apps.humvro.org/platform/?accessCode=tsa8b8v">https://apps.humvro.org/platform/?accessCode=tsa8b8v</a>
3	Organizational commitment (DEOCS)	"organizational commitment was assessed with five items that are consistently included in the DEOCS to assess the construct"	Not Applicable	<a href="https://apps.humvro.org/platform/?accessCode=6V0MzBa">https://apps.humvro.org/platform/?accessCode=6V0MzBa</a>
4	Exit from Training Survey	A survey designed to measure P-O fit, reasons for leaving, and Navy/training experiences from people who exit the Navy (either in RTC or A-school)	Separation situation, type, and location of separation, Navy life compared with expectations, Reasons for leaving the Navy, Modified Ways of Coping Checklist (WCCL), Navy P-O Fit Scale (Mottner, White, & Alderton, 2002), RTC Experiences A School Experiences, Experiences in the Fleet, Social Support while in Training, Organizational commitment - values similarity (Meyer & Allen, 1987), Self-assessed performance improvement	<a href="https://apps.humvro.org/platform/?accessCode=udraw0a3">https://apps.humvro.org/platform/?accessCode=udraw0a3</a>
5	Self-perceived military fit	"degree of Self-Perceived Military Fit was measured through two items from Selection Part 1: motivation for military service and self-reported suitability for military service."	motivation, suitability	<a href="https://apps.humvro.org/platform/?accessCode=t2QPeli">https://apps.humvro.org/platform/?accessCode=t2QPeli</a>
6	Defence Physical Fitness Test	"On the first day of training, the physical fitness of trainees was measured by a broad array of tests."	Cooper test, Push-ups, Sit-ups, Body fat, Body mass index	<a href="https://apps.humvro.org/platform/?accessCode=i9982WEJ">https://apps.humvro.org/platform/?accessCode=i9982WEJ</a>
7	Supervisory ratings	"overall supervisor rating of performance"	Not Applicable	<a href="https://apps.humvro.org/platform/?accessCode=zrpDLV9J">https://apps.humvro.org/platform/?accessCode=zrpDLV9J</a>
8	In-Unit Army Life Questionnaire (IU ALQ)	The ALQ measures Soldiers' self-reported attitudes and experience in the Army.	Affective Commitment, Career Intentions, Reenlistment Intentions, Attrition Cognitions, MOS Fit, Army Fit, MOS Satisfaction, Disciplinary Incidents, Physical Fitness (APFT score), Promotion Rate, Resilience/Citizenship / Leadership Behavior, Counterproductive Soldier Behavior, Motivation to Lead	<a href="https://apps.humvro.org/platform/?accessCode=3USx7X3">https://apps.humvro.org/platform/?accessCode=3USx7X3</a>
9	Drug use	"Illicit drug use was measured as the nonmedical use of any of 10 drug classes during the past 30 days: marijuana/hashish, cocaine, LSD, PCP, MDMA, other hallucinogens, methamphetamine, heroin or other opiates, GHB/GBL, and inhalants."	Not Applicable	<a href="https://apps.humvro.org/platform/?accessCode=j3H8t90">https://apps.humvro.org/platform/?accessCode=j3H8t90</a>
10	Affective and normative commitment	"Affective and normative commitments were measured using Meyer, Allen, and Smith (1993) 12-item measurements."	Affective, Normative	<a href="https://apps.humvro.org/platform/?accessCode=USKLtdV">https://apps.humvro.org/platform/?accessCode=USKLtdV</a>
11	Turnover intentions	"Turnover intentions were measured with the item 'I am lately considering looking for another job outside the Royal Netherlands Army'. Response options associated with this item were 'no', 'yes, within the armed forces', 'yes, outside the armed forces', and 'yes, both within and outside the armed forces.'"	Not Applicable	<a href="https://apps.humvro.org/platform/?accessCode=XhBH7pM">https://apps.humvro.org/platform/?accessCode=XhBH7pM</a>

Figure 2. Screenshot from the Criterion Instrument Database Output.

## 4.5 Estimating Psychometric Quality and Feasibility of Measurement Methods

Once the data entry tool had been populated, output from the tool was used to estimate<sup>5</sup> the psychometric quality and feasibility of the criterion instruments to inform our recommendations described in Section 5.0.

### 4.5.1 Procedures

Each criterion measure sourced from Section 4.2 was evaluated on three factors:

1. Generalizability (importance and criticality) of construct(s) assessed by the criterion measure

<sup>5</sup> We use the term "estimate" deliberately. The actual psychometric properties of each instrument were included as metadata elements (see Appendix E for details). However, information from the source material to populate those elements was not available for all instruments, and often different metrics were reported for different instruments (e.g., different estimates of reliability). Rather than rely on this incomplete information, we estimated the psychometric properties of the measurement *approach* (as opposed to the specific measure) so all instruments could be included in the analysis.

2. Psychometric quality of measurement method(s)<sup>6</sup> used by the criterion measure
3. Ease of administering the measurement method(s) used by the criterion measure.

**Generalizability.** Generalizability was determined by the importance and criticality of what was being measured by each instrument. We determined importance and criticality for each job performance, attitudinal, and organizational outcome subdimension using the procedures described in Section 3.0.

**Psychometric quality and ease of administration.** To evaluate the psychometric quality and ease of administration of each measurement method captured by the metadata elements, an extensive literature review was prepared that summarized the characteristics of different types of measurement methods on eight dimensions. For each measurement method, the write-up included a review of the relevant extant literature summarizing, for each measurement method, (a) comprehensiveness/deficiency, (b) susceptibility to contamination, (c) reliability, (d) discriminability, (e) validation uses, (f) cost of measure development, (g) ease and quality of administration, (h) ease and quality of data management, and (i) ease/cost of maintenance.<sup>7</sup> The dimensions are described in more detail below. Sections without relevant extant literature were supplemented with input based on experience gleaned from years of assessment development and military research experience.

We then set up a rating task, where we asked eight military measurement experts to review the write-up described above and make ratings to evaluate 22 measurement methods on the following evaluation dimensions.

#### Psychometric Quality

- *Susceptibility to contamination* – The extent to which score variance is attributable to job irrelevant determinants. (Scale 1-5: 1 - High susceptibility, 5 – Low susceptibility)
- *Reliability* – The extent to which scores produced by the measurement method are consistent over time. (Scale 1-5: 1 – Low reliability, 5 – High reliability)
- *Discriminability* – The extent to which the measurement method distinguishes between good and poor performers. (Scale 1-5: 1 – Low discriminability, 5 – High discriminability)
- *Validation Uses* – The extent to which the criterion measurement methods have proven useful for assessing the validity of cognitive, personality, interest, or physical ability constructs. (Scale 1-5: 1 – Low validation uses, 5 – High validation uses).

#### Administration Ease

- *Ease and Cost of Measure Development* – The cost associated with developing new measurement tools (Scale 1-5: 1 – High cost, 5 – Low cost)
- *Ease and Quality of Administration* - The extent to which high-quality data can be collected efficiently. (Scale 1-5: 1 – Low ease/quality, 5 – High ease/quality)

---

<sup>6</sup> Some criterion measures used more than one measurement method (e.g., self-reported attitudes and self-reported performance items).

<sup>7</sup> This write-up was provided to the project sponsors as a separate deliverable.

- *Ease and Quality of Data Management* – The extent to which data can be stored and managed easily, securely, and accurately. (Scale 1-5: 1 – Low ease/quality, 5 – High ease/quality)
- *Ease and Cost of Maintenance* – The cost associated with updating, revising, or maintaining the measurement tools and databases over time. (Scale 1-5: 1 – High cost, 5 – Low cost)

#### 4.5.2 Results

Rater ICCs (C,8) for the psychometric quality and application ease ratings were very high ranging from .75 to .95 across the evaluation dimensions with a median of .91. Appendix H provides the mean ratings for each of the 22 measurement methods on the evaluation dimensions. To summarize the data, we computed grand means across the psychometric quality and application ease evaluation dimensions. Those grand means appear in Table 9.

**Table 9. Mean Psychometric Quality and Application Ease Ratings**

Measurement Method	Types	Psychometric Quality	Application Ease
Performance Rating Scales: Supervisor	Ratings	3.48	4.16
Performance Rating Scales: Peer	Ratings	3.08	4.16
Performance Rating Scales: Self	Ratings	2.78	4.41
Performance Rating Scales: Multisource	Ratings	3.63	3.84
Work Samples/Hands-on	Performance	3.23	2.41
Simulations/Assessment Centers	Performance	3.13	2.41
Oral Interview	Performance	3.45	3.50
Situational Judgment Tests	Performance	3.45	3.22
Job Knowledge Tests	Performance	4.15	3.63
Self-report survey: Attitudes	Survey	3.73	4.56
Self-report survey: Objective performance/Personnel Data	Survey	3.03	4.56
Attrition	Admin	2.82	3.81
Reenlistment	Admin	2.69	3.97
Operational supervisor ratings	Admin	2.40	4.47
Promotion rate	Admin	2.83	4.38
Training school grades or pass/fail	Admin	3.13	3.91
Production indices	Admin	2.47	3.56
Personnel file records: performance scores (e.g., physical fitness, rifle qualification)	Admin	3.21	3.97
Personnel file records: Merit-based awards and recognition	Admin	2.57	4.16
Personnel file records: Negative outcomes (Articles 15, Counseling)	Admin	2.58	4.13
Personnel file records: Promotion points/scores	Admin	2.85	4.19

*Note.* The scale ranged from 1 (Low) to 5 (High).  $n = 8$  experts with extensive military research experience.

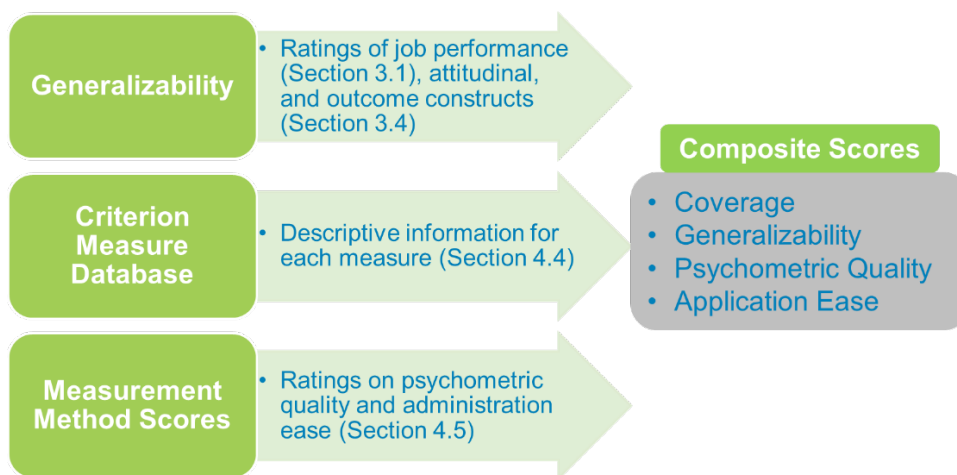
#### 4.6 Criterion Measure Scores

In order to identify the specific criterion measures with the most promise for cross-Service validation, we combined all the ratings described above and mapped them onto each instrument

using the appropriate metadata elements. As illustrated in Figure 3, composite scores were created for each instrument in the criterion measure database in three steps.

1. **Mapping generalizability ratings to metadata elements.** We mapped the generalizability ratings to the associated metadata elements in the database. For example, if an instrument assessed Teamwork and Initiative, the importance and criticality ratings associated with those constructs (in this case,  $M_{importance} = 3.92$ ;  $M_{criticality} = 3.62$  for Teamwork and  $M_{importance} = 3.68$ ;  $M_{criticality} = 3.59$  for Initiative) were mapped to that measure. For each construct / measure combination, importance and criticality were averaged to create an overall generalizability score.
2. **Mapping method ratings to metadata elements.** We mapped the psychometric quality / application ease ratings to the associated metadata elements. For example, if Teamwork and Initiative were assessed using a Situational Judgment Test (SJT) method, the psychometric quality and application ease ratings associated with SJTs (in this case,  $M_{psychometric} = 3.45$ ;  $M_{app\_ease} = 3.22$ ) were mapped to that measure.
3. **Construct composite scores.** For each measure, we constructed overall composite scores indicating coverage (i.e., how much of the criterion space is covered by each instrument), generalizability (i.e., how important / critical is each measure to the Services), psychometric quality (i.e., what is the overall psychometric quality of the measurement method used by this measure), and ease of use (i.e., how easy is the measurement method to develop, administer, manage, and maintain).

For each of the latter three composites (generalizability, psychometric quality, ease of use) we created two composite variants—one assessing the average and the other assessing the maximum. In the SJT example above, the Generalizability composite would have an average variant of 3.70 (reflecting the overall average of ratings for the Teamwork and Initiative dimensions) and a maximum variant of 3.77 (reflecting the value for the highest rated dimension, Teamwork). For the psychometric quality and ease of use composites, the average and maximum variants would be the same since there is only one measurement method. Criterion measures that use more than one measurement method would have different average and maximum variants.



**Figure 3. Overview of Composite Development Approach.**

We used these composite scores to determine the instruments that provided the best dimension coverage, measurement ease, and psychometric quality. Instruments that had high scores across composites were more likely to be included in our recommendations below than those with lower scores. Instruments that provided the best balance across all of the indicators were also more likely to be included in our higher-tiered recommendations.

## **5.0 RECOMMENDATIONS**

Our recommendations fall into two categories (a) recommendations for using and maintaining the products developed in the current project and (b) recommendations for measurement of criterion constructs.

### **5.1 Use and Maintenance of Research Products**

#### **5.1.1 Use of the Criterion Constructs**

Section 3.0 of this report describes the identification of criterion constructs that are relevant to first term enlisted occupations and generalizable across the Services. These constructs can be used by the Services in several ways. For example, an organization tasked with revising an operational rating form or survey could map extant items against the joint-Service constructs to determine measurement areas that might be missing or that might be overemphasized with redundant items. Or training and development materials could be revised to include constructs to emphasize in training.

We used the judgments of experts to evaluate the relevance and generalizability of the job performance constructs. For those judgments we sought input from individuals having cross-occupation knowledge within each Service. That is, the ratings were top-down, not bottom-up from each occupation. We are highly confident in the ratings; they were highly reliable and logically consistent across Services. To buttress the validity of the constructs and their prioritization, however, we recommend that the Services continue to collect data on them. Future job analysis and performance measurement projects should also include the job performance constructs.

#### **5.1.2 Use and Maintenance of the Criterion Instrument Database**

Section 4.0 of this report outlines the development of a criterion instrument database. The following deliverables were provided in support of this activity:

- *Online data entry tool.* The online data entry tool can be used to enter and maintain information about the identified criteria over time. It provides military researchers with a centralized point for accessing metadata information on existing criterion measures for use in future projects. It also allows for additional criterion measures to be added. For example, should future research desire to conduct a similar study for later-term enlisted Service members or officers, those measures could be captured using this tool.
- *Information on over 200 criterion measures.* We populated the database with criterion measures used by individual Services to validate new predictor measures and report to senior stakeholders. These criterion measures provide a reference for developing new criterion measures in the future, and further aligning criterion measurement across Services. Where we have them, we have also provided a library of instruments and references (HumRRO, 2018) that can be used in future research.

- *Analysis of measurement methods.* We also conducted a detailed analysis of different measurement approaches, to include a literature review and ratings of different measurement approaches. This analysis could also be beneficial to future researchers looking to build new criterion measures for military populations, either for the population of interest in this study, or for a new military population.

In terms of maintenance, we recommend the Services input additional criterion measures outside the parameters of the current study into the online data entry tool, should there be cross-Service populations of interest. For example, the sourcing, selection, and development of cyber warriors is of increasing importance across DoD components and an active area of research. This data entry tool can support cross-Service initiatives by serving as a repository for measures related to validating new selection instruments for those roles. There may also be cross-Service interest in other populations, such as officers, non-commissioned officers, and warrant officers. The tool may also be beneficial as a repository of criterion instruments within Service, providing a way for Services to easily access information about previously used criterion measures when conducting research.

## 5.2 Criterion Measurement Recommendations Overview

We identified the following guidelines for criterion measurement. Criterion measures should be:

- *Relevant and Generalizable.* Cover important components of the criterion space, as determined by ratings of importance across Services.
- *Feasible.* Relatively easy to develop, administer, score, manage, and maintain by minimizing the burden on Service members wherever possible.
- *Psychometrically Sound.* Yield data that are reliable, sufficiently variable, and relatively free from contaminating variance.
- *Flexible.* Allow some flexibility across Services to enhance Service-specific use, while ensuring support of needed criterion-related validity inferences.
- *Future-oriented.* Focus on what Services could accomplish with small, moderate, or greater amounts of effort, rather than solely on current practices.
- *Utilitarian.* Make the most of the Services' current practices and procedures by considering variables/instruments that are available first and supplement with new measures where they are needed to fill gaps.

With these guidelines in mind, we developed three tiers of recommendations (see Figure 4). Tier 1 is the most basic plan, with Tiers 2 and 3 adding additional measures with greater levels of effort. Tier 1 measures limited facets of the first term enlisted criterion domain. Each new tier expands measurement of the criterion domain and measurement quality.



**Figure 4. Three-Tiered Recommendations.**

- **Tier 1** – Recommendation 1.0 Maximize use of administrative data.
  - Focuses on using only available administrative data to include data from the Service’s personnel surveys.
  - Involves (a) developing standardized attrition and outcome variables and (b) aligning institutional personnel surveys to provide attitudinal and outcome data.
- **Tier 2** – Recommendation 2.0 Improve measurement of attitudinal, outcome and performance constructs.
  - Short self-report tools are the best way of measuring attitudinal criteria and have been shown to yield reasonably accurate data on performance outcomes and counterproductive work behavior.
  - Involves identifying/developing three short self-assessment tools, one for each of the following points in time: (a) end-of-technical training, (b) in-unit, and (c) exit.
- **Tier 3** – Recommendation 3.0 Improve measurement of job performance.
  - The Services should measure:
    - Organizational Citizenship and Peer Leadership with job performance ratings, an SJT, or both and,
    - Technical performance with job performance ratings, a job knowledge test, or both.
  - Without additional measurement, beyond Tier 2, substantial portions of the criterion domain are not covered.

### **5.3 Tier 1 - Maximize Use of Administrative Data**

#### **5.3.1 Tier 1 Overview**

Tier 1 recommendations focus on the most common approach Services take to validating new pre-accession selection instruments—maximizing the use of institutional administrative data collected across Services. Attrition, defined as separation before the end of a Service member’s first term, is the most common administrative variable used in criterion-related validation studies. Four out of five Services in our research had conducted at least one predictive validation study that used attrition as a criterion variable. Other common criterion variables that rely on administrative data include (a) promotion rates and / or promotion points, (b) merit-based awards

and recognition, (c) negative outcomes (e.g., disciplinary incidents), (d) performance scores (e.g., physical fitness scores), and (e) training school outcomes.

Another reason for starting our recommendations with administrative criteria is ease of application. Our research examining different types of measurement methods (see Section 4.4 for details) found that administrative criteria tended to rate high on ease of application dimensions such as ease of measure development, ease of administration, and ease of maintenance. Thus, these recommendations represent the “low hanging fruit” for Services to align their validation research.

While the Services commonly rely on administrative data for criterion-related validation, there is currently some variation in how individual Services construct and interpret these variables. Our recommendations in Tier 1 suggest explicit steps for aligning these administrative criteria, with particular emphasis on (a) standardizing data that are already collected and (b) enriching the use of those data in validation research.

### **5.3.2 Tier 1 Recommendations**

*Recommendation 1.1: Standardize computations of attrition across Services and make full use of the separation data.*

*Recommendation 1.1.a. Align attrition to specific “gates” and computational decision rules.*

As described previously, nearly every Service uses attrition as a key outcome for validating new predictor measures. However, there is some variation regarding, for example, at which points in the first term of Service attrition is computed. There are also differences in how each Service computes attrition; for example, whether an overall measure of attrition includes reasons not likely to be linked to individual differences (e.g., attrition due to injury or personal reasons). We recommend that the Services standardize their attrition measurement in terms of (a) key points during first term enlisted Service that serve as “gates” for attrition measurement and (b) decision rules regarding the treatment of different reasons for separation and other key administrative variables.

To implement the first part of this recommendation, we recommend that the Services construct attrition variables at monthly intervals (and maintain dates of separation) throughout a Service member’s first term of Service, and use the following points in time for validation research:

- Boot camp/basic training
- End of initial technical training
- Annual milestones (one-, two-, three-year attrition, potentially more depending on contract length)
- Contract completion

The specific number of months for some of these gates will differ across Services and by technical specialty within Service. While these gates are modelled after attrition research conducted by the USMC (Charles & Moynihan, 2017), they are consistent with other Services’ conceptions of important points in the first term of Service.

To implement the second part of this recommendation, research should examine the separation codes used to construct attrition variables across the Services and align decision criteria. These data can be very messy and unreliable. For example, some codes are used as “defaults” even if



they do not accurately reflect the reason for an individual separation. Services deal with this messiness in different ways, leading to divergent methods of variable construction. This activity is elaborated on in Recommendation 1.1.b.

*Recommendation 1.1.b. Compute overall attrition and attrition “types.”*

Most Services rely on assessments of overall attrition at key points in time without differentiating between different reasons for that attrition. However, research examining such reasons demonstrate that this differentiation can be highly informative. For example, research conducted by the USAF demonstrates the types of insights that can be gained by examining different reasons for separation, as determined by separation codes. Hooper, Paullin, Putka, and Strickland (2008) found that types of attrition (e.g., medical, performance, moral character) were associated with separation at different points in Airmen’s first term of Service. In this study, medical separations were prevalent early in the first term of Service, while moral character reasons were more predominant later in the first term. They found Armed Forces Qualification Test (AFQT) category to be particularly predictive of performance-related reasons for attrition. This study illustrates that using separation codes to identify different categories of attrition can help to increase predictor / criterion correspondence and enrich the Services’ understanding of the effect of different predictors on key outcomes.

The performance taxonomy provides a guidepost for determining these types of attrition. For example, previous publications of DoD separation codes demonstrates alignment with different aspects of the performance taxonomy, such as psychosocial well-being (e.g., Misconduct [Drug Abuse]), physical fitness (e.g., Physical Standards), and general performance (e.g., Substandard Performance; see Army Regulation 635-5-1 for details).<sup>8</sup> Detailed examination and alignment of codes across the Services can create categories of attrition that are more descriptive and theoretically meaningful for criterion-related validation. For these reasons, we recommend that Services (a) identify conceptually meaningful (i.e., aligned to the performance taxonomy) clusters of separation codes and (b) construct attrition variables associated with those clusters at different “gates” identified in Recommendation 1.1.a.

*Recommendation 1.2: Closely examine performance-based administrative outcome measures and align their collection across the Services to the extent possible.*

We suggest prioritizing further examination of administrative variables that can be clearly linked to the performance model presented in Section 3.0.

*Recommendation 1.2.a. Explore aligning basic and technical training success across Services.*

Another common administrative variable examined across the Services is success in basic and technical training. In addition to separation from training (which would be included in the attrition variables referenced above), many Services have also examined training success, such as (a) success without wash-backs<sup>9</sup> and (b) graduation with honors. For example, in examining enlistment waiver policies, the USAF examined the effect of different waivers on (among other criteria) basic / technical training wash-backs and whether each Airman was an honor graduate in basic / technical training (Putka & Allen, 2008). The US Army makes use of initial military training performance, specifically restarts (must begin training again) and failures (failed one component of training), in its Tier One Performance Screen (TOPS) research (Knapp &

<sup>8</sup> [https://armypubs.army.mil/epubs/dr\\_pubs/dr\\_e/pdf/web/r635\\_5\\_1.pdf](https://armypubs.army.mil/epubs/dr_pubs/dr_e/pdf/web/r635_5_1.pdf)

<sup>9</sup> Other terms that are used include “recycles,” “restarts,” and “failures.”

Kirkendall, 2018; Knapp & Wolters, 2017). As a final example, the U.S. Navy included First-Pass-Pipeline-Success, defined as the “chance of passing all A-school training pipeline without any academic fail or academic setback training events,” in a recent validation study of a selection algorithm (Watson, 2016; slide 5). While the USMC and the US Coast Guard have not traditionally included these metrics in their validation research, it is possible that this information is also collected and maintained in a database where it is discoverable.

We recommend each Service examine the discoverability of its basic and technical training data as a proxy for the technical proficiency aspect of the job performance taxonomy. Services should capitalize on alignments in interpretation across Services. Unlike attrition, basic / technical training data tend to have limited information regarding *reasons* for outcomes; however, in all three of the Services mentioned above, researchers were able to distinguish between failures for academic versus non-academic reasons.

*Recommendation 1.2.b. Collect physical fitness from administrative records.*

To the extent possible, Services should collect physical performance from administrative records. Four out of the five Services (US Army, USAF, USN and USMC) require regular physical fitness exams as a Service requirement. While there is substantial variability in the importance and criticality of physical performance across the Services, all the Services<sup>10</sup> rated physical fitness as at least “important.” Thus, assessments of physical fitness should regularly be included in validation research. These include the following assessments for the four Services listed above:

- a. *Army Physical Fitness Test (APFT)* – The intent of the APFT is to provide an assessment of the Army's Physical Readiness Training (PRT) program. The APFT assesses base level of fitness for every Soldier by testing the muscular strength, endurance, and cardiovascular respiratory fitness of soldiers in the Army.
- b. *USAF Fitness Assessment (FA)* – An assessment of overall fitness of USAF personnel. It assesses aerobic and muscular fitness, flexibility, and optimal body composition.
- c. *Navy PRT* – The PRT evaluates aerobic capacity / cardio-respiratory endurance, muscular strength, and muscular endurance. It is part of a larger Physical Fitness Assessment that also includes medical screening and a Body Composition Assessment (BCA).
- d. *USMC Physical Fitness Test (PFT)* – The PFT is a collective measure of general fitness administered USMC-wide. It is designed to measure strength and stamina of the upper body, midsection, and lower body as well as efficiency of the cardiovascular and respiratory systems.

All four of the above assessments were rigorously developed, evaluated, and normed against Service populations, making them prime candidates for inclusion in validation research. Additionally, because successful completion of these assessments is a requirement for Service, scores are more likely to be included in administrative databases than other performance-based administrative outcome measures.

*Recommendation 1.2.c. Collect disciplinary incidents information from administrative records.*

To the extent possible, Services should also collect information regarding disciplinary incidents from administrative records. Disciplinary incidents are not typically included in Services’

---

<sup>10</sup>The Coast Guard did not participate in the ratings exercise.

validation research. For this reason, there are several outstanding questions, such as: Do the Services routinely collect this information in administrative databases? What is the quality of those data? How accessible are those data? To what extent is there consistency in the type of information collected across the Services? Despite this ambiguity, we include this recommendation because disciplinary incidents represent a potential proxy for a critical (and currently under-represented in validation research) portion of the job performance space—Psychosocial Well-Being. Thus, we recommend the Services investigate their procedures for collection and maintenance of these data.

*Recommendation 1.2.d. Explore methods of creating a performance outcome composite across Services.*

Another common approach to validating pre-employment assessments is to develop an overall performance metric using data available from administrative records. These variables often include some combination of awards (e.g., medals), training accomplishments, and time-in-Service. We recommend using research conducted by the USAF as a starting point for developing a performance outcome composite. Although the methods used by the USAF cannot be used directly in other Services due to differences in policy and available data, they have successfully developed and validated a couple of such aggregated measures that could be used to design a larger study for developing general overall performance criteria that can be applied across Services. In particular, the USAF's research into the MM-RS; Alley, Pacheco, & Birkelbach, 2007; Halper, Goodman, & Alley, 2010) variable—a score in months reflecting the amount of time spent at different skill levels—and its *Weighted Airman Promotion System (WAPS)* (Schiefer, Robbert, Crown, Manacapilli, & Wong, 2008; the Air Force is also currently conducting a study examining the validity of all WAPS components) are an effective place to begin for further development of one or more cross-Service performance outcome variables.

Note that these variables are not easy to construct. In the examples above, both variables are standardized within cohorts such as occupational specialty due to differences in the amount and difficulty of training in different specialty areas. These performance outcome composite variables also often require substantial SME input to develop and maintain. However, we believe in this case that the potential benefits to having a shared performance metric across Services outweighs the challenges. That said, the challenges inherent in the use of administrative data to create general performance variables suggests a feasibility study should be conducted before undertaking a large-scale effort.

*Recommendation 1.2.e. Conduct research to (a) examine the feasibility of Recommendations 1.2.a. through 1.2.d. and, if appropriate, (b) validate any new criterion variables created in 1.2.a. through 1.2.d.*

While this is incorporated into the above recommendations, we believe it is worth pointing out that we anticipate there will be wide variation in the availability and quality of data across the Services to construct the above variables. Thus, we recommend first conducting a feasibility study to examine the quality and availability of data to construct the above variables. Assuming the availability and quality is not prohibitive for further exploration across multiple Services, a follow-up study could be conducted to develop and validate the shared criterion variables.

### 5.3.3 Tier 1 Summary

Tier 1 recommendations involve collecting and, where appropriate, creating standardized variables based on data already collected by the Services. We believe implementation of these Tier 1 recommendations would yield substantial benefits to the Services in future validation research. Specifically, these measures would:

- Provide routine access to relevant, generalizable data addressing several parts of the criterion domain.
- Provide metrics that are meaningful to Service stakeholders for reporting purposes.
- Reduce the burden on the Services to routinely conduct local validation studies as administrative criteria will be more readily available than other types of criteria.
- Since administrative criteria are the most commonly used across the Services, implementation of these Tier 1 recommendations would yield the most benefit in aligning validation research across the DoD.

As seen in Figure 5.2, if fully implemented, the Tier 1 recommendations would address the following portions of the performance model: Technical Proficiency, Physical Performance, Attrition, Reprimands, Performance Records, and Merit-based Awards.

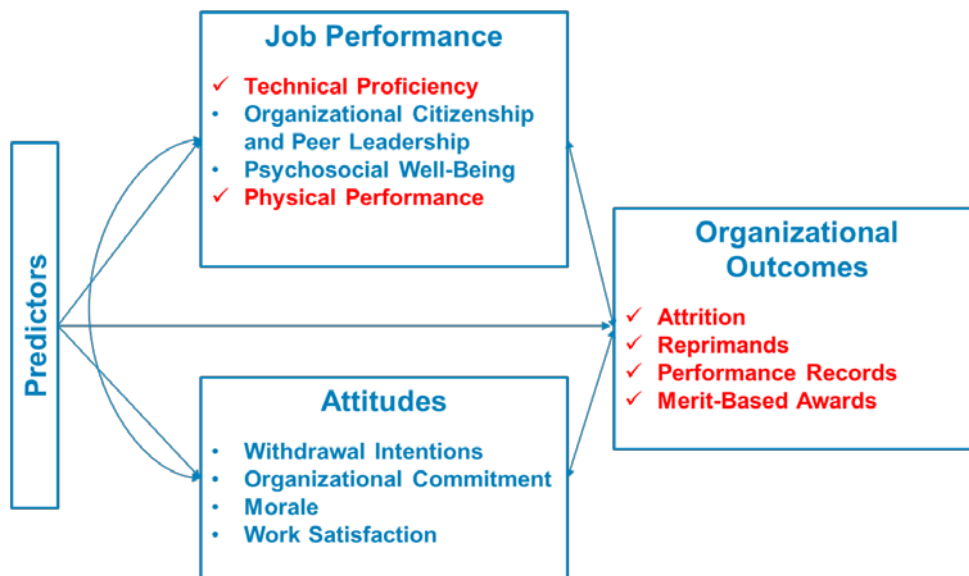


Figure 5. Criterion Domain Constructs Measured in Tier 1.

## 5.4 Tier 2 Improve Measurement of Attitudes and Outcome Variables

### 5.4.1 Tier 2 Overview

While Tier 1 would provide accessible administrative measures of outcomes, administrative variables do have limitations. The primary objective of Tier 2 recommendations is to offset some of these limitations through self-report assessments.

The first set of limitations has to do with attrition data. As described previously, attrition data take years to mature, particularly if reenlistment is of primary concern. This means that if a new

predictor of reenlistment (or attrition at the end of the first term) were to be validated, it could take three to five years to obtain the final attrition data. Even then, base rates for actual attrition can be low, making analytic work difficult. To overcome those issues, the Services often obtain self-reported withdrawal cognitions, morale, satisfaction, and commitment to serve as a near-term proxy in concurrent validation designs. Another issue regarding attrition data has to do with reasons for leaving. Prediction of attrition can improve if reasons for it are reported in a standardized fashion, and the Services conduct exit surveys for that purpose.

The second limitation has to do with the quality of outcome data in administrative databases. Administrative data are gathered through many steps in a large bureaucracy, with many players contributing pieces. Error can enter the system at any point and data can be dated. Research suggests that individuals self-report performance outcomes with reasonable accuracy. For example, in Project A research, the Army found that self-report disciplinary actions were consistent with official Army records. With that in mind, the Services often ask Service members to provide concrete data on their own performance outcomes such as physical fitness test scores, rifle qualification scores, awards, and reprimands.

The third limitation is that outcome data do not cover important facets of the criterion domain very well. Based on our research, the most important, generalizable performance category is Psychosocial Well-Being (PWB) and the only administrative variable addressing it is reprimands. Psychosocial Well-Being subsumes two dimensions (a) adapting to stressful situations and (b) Counterproductive Work Behavior (CWB), both of which can be addressed in self-report assessments. Self-report of stress reactions in assessments is relatively common. Also, a recent meta-analysis of CWB ratings found that self-raters reported *more* CWB than supervisors or other raters, seemingly in contradiction with other meta-analytic results reporting greater leniency in self-ratings (Berry, Carpenter, & Barratt, 2012). It may be that supervisors and others do not observe or notice some of the CWB that individuals report.

In sum, three types of information can usefully be collected in self-report assessments: (a) attitudes related to leaving and reasons for leaving, (b) performance outcomes, and (c) PWB indicators.

#### **5.4.2 Tier 2 Recommendations**

*Recommendation 2.1: Create three short self-report assessments oriented toward three points in time: (a) end-of-training (EOT), (b) in-unit (IU), and (c) exit.*

The purpose of the EOT and IU assessments would be to provide attitudinal, performance outcome, and PWB data to be used as criteria in validation studies. The exit survey will have a narrow purpose. Its goal will be to provide a better measure of attrition by capturing reasons for attrition in a standardized way.

The assessments should measure intended constructs with as few items as possible. We expect that the Services will want to supplement the core set of joint-Service items with some Service-specific items of their own choosing in their research projects. The core items should include only items that (a) measure the intended constructs and (b) are generalizable across Services.

*Recommendation 2.1.a. Identify and standardize core self-report items for an EOT assessment.*

The EOT assessment will measure (a) withdrawal intentions and reasons for leaving/staying, (b) commitment, (c) morale, (d) training performance outcomes (e.g., wash-backs, physical

fitness scores), and (e) psychosocial well-being (i.e., adjustment to stress and CWB). While self-ratings of job performance have been shown to suffer from leniency error, there is a growing body of research that suggests that people do report CWB reasonably accurately (Berry et al., 2012). Therefore, we recommend including a small set of CWB items on the form.

Two surveys developed for use at EOT provide the best sources of items for the standardized form:

- The US Army's IMT ALQ
- The USN's Recruit Training Command (RTC) Graduate Survey and A-School Graduate Survey

Both instruments address a large proportion of the constructs to be measured and have been used with considerable success by the Services. Both instruments were designed to be used at the end of basic or technical training either "as is" or with very minor tweaks in wording to adjust to the two time points. The core set of EOT items should also be constructed such that they can be used at either time point.

*Recommendation 2.1.b. Identify and standardize core self-report items for an in-unit assessment.*

The IU assessment will measure (a) withdrawal intentions and reasons for leaving/staying, (b) commitment, (c) morale, (d) work satisfaction, (e) performance outcomes (e.g., reprimands, awards, PFT scores), and (f) PWB (i.e., adjustment to stress and CWB).

The best sources of items for the standardized form are:

- IU US Army Life Questionnaire (IU ALQ)
- Draft USAF Life Questionnaire

Both the IU ALQ and the draft USAF Life Questionnaire contain questions aimed at both attitudes and performance outcomes. The IU ALQ also includes self-ratings for some CWBs and reasons for leaving. It is the most comprehensive self-assessment we reviewed, covering over 83% of the attitudinal constructs and 50% of the performance outcome taxonomy.

For the core joint-Service items, we recommend pulling the IU ALQ items that are directly related to the taxonomic elements in the joint-Service taxonomy and filling any measurement gaps based on literature reviews and the criterion measure database. For example, the IU ALQ does not currently include items about reactions to stress and self-destructive behaviors such as alcohol abuse. Special purpose surveys such as the Kessler Psychological Distress Scale or the Work Cynicism Survey can be mined for potentially useful items.

*Recommendation 2.1.c. Identify and standardize core self-report items for an exit assessment that is aligned with attrition types from Recommendation 1.1.b.*

As mentioned in Recommendation 1.1.b, researchers seek improvements to attrition measurement by fine-tuning types of attrition. People leave the Services for different reasons. Some pursue higher-paying civilian careers; others want to spend more time with their families.

The primary purpose of the exit assessment will be to gather information to fine-tune attrition measurement. The reasons for leaving and any other variables on the exit survey should be aligned to attrition types from Recommendation 1.1.b. We recommend starting with that typology and identifying the reasons for leaving and any other variables that should be needed

for classification of attrition into the appropriate type. In turn, the Services exit survey and self-assessments can be mined for specific items.

*Recommendation 2.2:* Align existing, operational personnel surveys to provide self-report data that would be routinely accessible for validation research projects.

While there are likely to be a number of challenges to aligning the content of operational surveys and making those data accessible for validation research, doing so could essentially result in a paradigm shift for DoD, at least where collection of self-report criterion data is concerned. If this recommendation was realized, standardized data from large samples would be routinely available.

The following institutional surveys currently collect attitudinal data relevant to the criterion model:

- Defense Manpower Data Center (DMDC) Status of Forces Survey
  - A family of six operational surveys measuring satisfaction with aspects of being in the military, broad and detailed retention items, affective and continuance commitment, tempo, readiness, stress, deployments, morale. Items are divided across six forms that are administered cyclically over the course of two years.
- The US Army's Sample Survey of Military Personnel (SSMP)
  - An operational survey that assesses career, separation, and attitudes of the Army. Note that this survey is no longer operational.
- USN-wide Personnel Survey (NPS)
  - An operational survey of attitude and opinion among Navy personnel. The objective of the NPS is to measure Sailor satisfaction with Quality of Work Life (QWL) indicators such as job satisfaction, organizational commitment, leadership satisfaction, and workplace climate and their effects on outcome measures such as retention intention.
- Career Decisions Survey also known as the Military Retention Survey
  - An operational survey of reasons people stay in the USAF and is parallel to the *New Directions (Exit) Survey*. It is administered every two to three years to a sample of Colonel (O6) and below who do not have an established date of separation.
- New Directions Survey also known as Military Exit Survey
  - An operational survey evaluating reasons people are leaving the USAF. It is administered monthly to separating or retiring O6 and below.
- Enlisted Accessions Survey
  - An operational survey of new USMC's attitudes and perceptions about the USMC. Participants have completed Basic Training. This survey has the same items as the Enlisted Retention Survey.
- USMC Exit Survey
  - An operational survey delivered to USMC Service members prior to separation.
- Organizational Assessment Survey

- An operational survey to collect work environment perceptions and attitudes for the US Coast Guard workforce (i.e., Active, Reserve, and Civilian).

*Recommendation 2.2.a. Create a task force tasked to establish working relationships with existing survey organizations.*

The surveys developed in Recommendation 2.1 will provide the common item set to be included in operational surveys. The task force will need to collect information and negotiate regarding:

- *Identifiability.* To be useful for validation, survey data must be identifiable so that Service members' data can be matched to administrative and predictor data. Our read of the current surveys is that some collect data in an identifiable way. While we acknowledge that this will likely be a significant hurdle to implementing these recommendations, we believe the research benefits to the Services will be substantial enough to outweigh these concerns.
- *Other considerations.* The task force should gather information about any other aspects of the survey that are important for consideration such as the survey mission, population, sampling, and timing.
- *Procedures for changes to survey content.* Organizations may resist adding new content to an existing survey. Fortunately, we believe that most of the surveys would require minimal change to cover core items.

### **5.4.3 Tier 2 Summary**

Tier 2 recommendations involve creating standardized self-report assessments. Self-report assessments can be used to efficiently measure job attitudes and outcomes. They cover much of the criterion domain efficiently, often just requiring 20 or 30 minutes of the respondent's time. As illustrated in Figure 5.3, Tier 2 recommendations address the portions of the performance model that appear in red.





**Figure 6. Criterion Domain Constructs Measured in Tier 2.**

## 5.5 Tier 3 Improve Measurement of Job Performance

### 5.5.1 Tier 3 Overview

While self-report assessments and outcome data can touch on or partially measure job performance, they are insufficient measures of job performance. People tend to be lenient in evaluating their own performance (Heidemeier & Moser, 2009); thus, self-ratings are often inflated and lack variability. While outcome measures, such as previously described training school wash-backs/graduation and fitness test scores, can be very useful measures of an aspect of job performance, they typically measure only a narrow portion of the job performance domain. As shown in Table 5.1, the performance categories / dimensions of Organizational Citizenship & Peer Leadership and Technical Performance are not well-measured by criterion instruments recommended in Tiers 1 and 2. Referring to the psychometric quality and ease of application rating in Section 4.5, job performance ratings and objective format tests have the highest potential for achieving adequate psychometric quality in a feasible fashion. Therefore, our Tier 3 recommendations focus on capturing key parts of the performance space not well-measured in Tier 1 and Tier 2 using standardized assessments. These criterion measures would be administered as part of a Service-specific validation study.

**Table 10. Gaps in Job Performance Measurement Tiers 1 & 2.**

Job Performance Categories and Dimensions	Recommended Instruments from Tiers 1 & 2	
	End-of-Technical Training	In-Unit
<b>Psychosocial Well-Being</b>		
Counterproductive Work Behavior	<ul style="list-style-type: none"> <li>• Tier 1 Administrative Data (e.g., reprimands)</li> <li>• Tier 2 EOT Assessment</li> </ul>	<ul style="list-style-type: none"> <li>• Tier 1 Administrative Data (e.g., reprimands)</li> <li>• Tier 2 In-Unit Assessment</li> </ul>
Adapting to Stressful Situations	<ul style="list-style-type: none"> <li>• Tier 2 EOT Assessment</li> </ul>	<ul style="list-style-type: none"> <li>• Tier 2 In-Unit Assessment</li> </ul>
<b>Organizational Citizenship &amp; Peer Leadership</b>		
Organizational Support	--	--
Planning and Structuring Work	--	--
Support for Peers	--	--
Conscientious Initiative	--	--
<b>Technical Performance</b>		
Safety Consciousness	<ul style="list-style-type: none"> <li>• Tier 1 Administrative Data (e.g., Wash-backs)</li> </ul>	--
Task Performance	<ul style="list-style-type: none"> <li>• Tier 1 Administrative Data (e.g., Wash-backs)</li> </ul>	--
Decision Making, Problem Solving, and Innovation	--	--
Oral and Written Communication	--	--
<b>Physical Performance</b>		
Fitness	<ul style="list-style-type: none"> <li>• Tier 1 Administrative Data (e.g., physical fitness test scores)</li> <li>• Tier 2 EOT Assessment</li> </ul>	<ul style="list-style-type: none"> <li>• Tier 1 Administrative Data (e.g., physical fitness test scores)</li> <li>• Tier 2 In-Unit Assessment</li> </ul>

**5.5.2 Tier 3 Recommendations**

*Recommendation 3.1:* In Service-specific validation studies, (a) *Organizational Citizenship & Peer Leadership* should be measured with job performance ratings, an SJT, or both and (b) *Technical Performance* should be measured with job performance ratings, a job knowledge test, or both.

There are tradeoffs associated with the recommended measurement methods. Job performance rating scales measure the performance domain broadly and are relatively inexpensive to develop. However, their psychometric quality hinges on careful administration practices such as collecting multi-source ratings and training raters. Job knowledge tests typically have strong psychometric properties, but they only measure a narrow portion of the full job performance domain. Similarly, SJTs can be developed to measure the interpersonal aspects of a job reasonably well, but they are less conducive to measuring other aspects of job performance.

For these reasons, we present performance ratings and objective tests as optional recommendations to give the Services flexibility in choosing the most appropriate measurement approaches. Performance ratings and objective tests could potentially be implemented during validation research to cover the criterion domain constructs. However, it is important to bear in mind that the most critical Tier 3 recommendation is 3.1—some form of performance measurement is needed to adequately assess Organizational Citizenship & Peer Leadership and Technical Performance.

Notably, we are not recommending the development of simulations or hands-on performance tests because they are laborious to develop, administer, and maintain. Given the expense, they are probably more effectively used as criteria in studies focusing on a specific occupation. Administering such measures on a large scale would be onerous.

*Recommendation 3.2: Collect end-of-technical-training (EOTT) and in-unit job performance ratings.*

Job performance ratings have a distinct advantage over most other measurement methods. They can cover a broad swath of the performance domain efficiently (Knapp & Campbell, 1993; O’Leary & Pulakos, 2017). Performance rating forms are relatively inexpensive to develop, easy to administer, and inexpensive to maintain. The primary downside is that it is very difficult to obtain reliable performance rating data. The psychometric quality of performance ratings has more to do with how the data are collected and who makes the ratings than the format of the actual rating form (Pulakos, 2007). Three factors are important for improving the quality of performance rating data.

1. Raters should receive training to become familiar with the rating dimensions. Frame-of-reference training has been shown to increase rating accuracy (Borman, Grossman, Bryant, & Dorio, 2017).
2. Data quality improves if ratings are collected from multiple raters and multiple sources. Ratings from different sources are typically thought to provide different perspectives on true performance (Smither, London, & Reilly, 2005) and the most valid depiction of true performance is expected to come from multiple sources and raters.
3. Performance ratings should be collected for research only. Raters should be told that the ratings will not affect the Service member’s career. Operational performance ratings typically have little variance and are more susceptible to rating errors and biases (Knapp & Campbell, 1993).

Based on the above, we offer the following specific recommendations to Recommendation 3.2:

*Recommendation 3.2.a. Develop standardized Performance Rating Scales (PRS) to cover all or almost all generalizable job performance dimensions in the First Term Enlisted Job Performance Taxonomy.*

Earlier in our research, we examined data on 51 job performance rating forms. No single form covered all the important job performance dimensions in the first term performance model well, and no single form is ready for joint-Service use. Therefore, we recommend the development of rating forms for the generalizable job performance dimensions in the First Term Enlisted Job Performance Taxonomy.

We recommend that the new rating form use a relative rating format that asks raters to compare the ratee's performance to that of their peers. The relative format offers several advantages over other rating scales for use in a validation setting. First, they are shorter and require less reading than other formats, leaving more time for frame-of-reference training. Second, research suggests that job performance is not normally distributed in organizations (LeBreton, Burgess, Kaiser, Atchley, & James, 2003). There are fewer inadequate performers for example. Rating forms that use an absolute scale where the lower two to three points focus on poor performance essentially waste those lower point of the scale. Relative ratings get around this problem by asking how the ratee performs relative to their peers.

*Recommendation 3.2.b. Develop processes, instructions/training, and an on-line rating tool.*

We strongly recommend collecting multisource ratings, where possible. In our experience, it is difficult to collect EOTT PRS from instructors. Instructors often do not know students by name and cannot rate very many aspects of training performance. They also typically are responsible for large numbers of students, making the rating task quite burdensome. However, it is reasonable to collect both peer and supervisor ratings in-unit.

Regarding other processes, we recommend:

- *Online administration* – Ratings can be collected efficiently online; however, response rates are lower than for in-person data collection and rating quality may suffer if raters are not sufficiently instructed and trained.
- *Instructions/Training* – Use frame-of-reference training, which attempts to familiarize raters with the content of rating dimensions as it has been shown to increase rating accuracy (Borman et al., 2017). Development of standardized, online video-based training would further enhance the uniformity and ease of training application.
- *Research-only* – As described previously, ratings used for operational decision-making (e.g., performance management) typically have little variance and are more susceptible to rating errors and the biases than research-only performance ratings (Knapp & Campbell, 1993).

*Recommendation 3.3: Collect Service-wide job knowledge data in-unit to measure technical performance.*

There are two types of objective measurement of technical knowledge: (a) training school grades and (b) job knowledge testd. Job knowledge tests (JKT) are maximal performance measures, not typical performance measures. That is, they capture the test-taker's best performance, not day-to-day performance. This means that they tap "can-do" performance more than "will-do" performance over time (Borman et al., 2017). Training grades are also considered a "can-do" performance criterion although will-do motivational aspects might play a slightly higher role than JKTs. While scores on technical tests typically correlate with supervisor and peer ratings of technical performance (e.g., Allen, Knapp, & Owens, 2016), tests and ratings are very different measurement methods. Technical performance ratings made by supervisors and peers reflect day-to-day performance on the job over time while performance ratings on technical job dimensions are thought to be a function of both can-do and will-do performance

In theory, training school grades are one of the most desirable criterion measures because (a) training performance can be linked directly to the costs the Services incur for recruit training and (b) grades provide more variability than washback/recycles for use as validation criteria. In

practice, as described in Recommendation 3.2, the only widely available training data in administrative databases are (a) whether the student passed the course and (b) how many times the student had to retake or recycle through the course before passing it. Some schools produce a grade, but only record pass/fail in administrative databases. In those cases, researchers can obtain grades directly from the school (e.g., Trippe, Moriarty, Russell, Carretta, & Beatty, 2014). Schools that do not produce grades at all should be excluded from these analyses.

*Recommendation 3.3.a. Collect end-of-school grades for special-purpose validation studies.*

In the past, when it has been particularly important to get good technical performance criterion data for joint-Service ASVAB validation, the DMDC; (now the Defense Personnel Assessment Center [DPAC]) has hired contractors to work with instructors for the courses included in the project to obtain course grades. For example, in the Enhanced Computer Adaptive Test (ECAT) project, a contractor obtained performance criteria for 13 Navy schools, two Air Force courses, and three US Army schools (Kieckhafer et al., 1992; Wolfe, Alderton, Larson, Bloxom, & Wise, 1997). To do this, the contractor collected data on quizzes, homework assignments, and laboratory assignments for samples of 700 students in each school. They developed dimensions of achievement in each school and computed scores for each student. Researchers computed criterion-related validity estimates for each school/course. Similarly, in DPAC's ASVAB Specs Project (formally, Item Evaluation for the Science and Technical Test Specifications Project) the contractor gathered data from nine Army and four USAF training courses (Oppler, Russell, Rosse, Keil, Meiman, & Welsh, 1997). The contractor contacted the course instructors, requested grades, and worked closely with instructors to obtain them.

Those studies focused specifically on evaluating the validity of (a) new cognitive and psychomotor predictors and (b) alternative test specifications for the ASVAB. Since the ASVAB is expected to predict technical performance, it was necessary to obtain strong technical performance criterion measures. While collecting and perhaps augmenting the training grades for criterion data is laborious, this is an approach that Services may want to consider in the future for specific validation efforts, particularly validation of new measures that are expected to improve prediction of technical performance.

*Recommendation 3.3.b. Develop Service-wide JKTs.*

JKTs have some important advantages compared to other performance measurement methods. A JKT can sample more tasks (i.e., a larger portion of the technical domain) with greater efficiency than hands-on tests or simulations. These tests use objective formats and can easily be administered in group sessions either on paper or on-line. Paper versions can use scannable forms that are relatively easy to process. JKTs also typically yield strong psychometric reliability estimates (Knapp & Campbell, 1993). The most significant downside of job knowledge testing is that test development typically involves heavy involvement of SMEs to ensure the accuracy and relevance of test items. Usually this involves training a pool of SMEs in item writing, reviewing and editing draft items, and asking SMEs to review and edit items. It can be challenging to write large numbers of good test items.

There are a couple of important considerations when developing Service-wide JKTs. The first is whether the test will assess occupation-specific knowledge or Service-wide knowledge. Developing occupation-specific knowledge tests in each branch of the armed Services would be an overwhelming and resource-intensive effort. Therefore, we recommend the development of

one Service-wide knowledge test for each Service. Service-wide JKTs should be aimed at the level of knowledge expected of Service member's well into or nearing the end of their first term, E3 – E4 level (18 to 36 months of experience).

Service-wide JKTs have been developed and used with reasonable success. In the Army's Future-oriented Experimental Army Enlisted Personnel Selection and Classification (Select21) project, researchers developed the Army-wide JKT to measure first term Soldier knowledge/performance of common tasks (e.g., land navigation, first aid, survival). Internal consistency reliability for this measure was acceptable ( $\alpha = .73$ ) and yielded reasonable correlations with other criteria, supporting its construct validity (Knapp, Sager, & Tremble, 2005). Another example is the USAF's PFE. The PFE is an operational test of general knowledge about the USAF and it is used as a part of promotion decisions. The content outline includes a host of topics (e.g., Air Force Heritage, Organization, and doctrine, leadership, security). New PFE forms are created each year and they are not equated.

The content outlines for the US Army and USAF tests differ substantially and since data analyses would likely be conducted within Service first and then summarized across Services, the tests need not cover the same content. Further, it would be useful for the Services to agree on a broad definition of content for the Services to use in developing their tests. An example definition is "general knowledge expected of all first term enlisted personnel at an agreed upon milestone (such as 18 – 36 months of experience)." Finally, Service-specific JKTs should measure two elements of Technical Performance—Safety Consciousness and Task Performance. Both were found to be important and critical across Services (refer back to Table 3.3) and are conducive to measurement using JKT formats.

*Recommendation 3.3.c. Focus Service-specific JKTs on IU technical performance.*

Another important consideration is whether JKT is done EOTT, IU, or both. IU testing is the higher priority because there are so few measures of in-unit technical performance. Ideally, for a Service-specific validation study, some effort would go into obtaining useful EOTT course grades and administrative training data. In-unit general technical proficiency could then be measured using Service-wide JKTs.

*Recommendation 3.4: Develop a DoD-wide SJT.*

SJTs, also referred to as low-fidelity simulations, are most commonly used as predictor measures. Accumulating evidence of their validity for predicting job performance with less adverse impact than cognitive ability measures makes SJTs desirable. Additionally, they appear to be less vulnerable to faking than self-report measures (Hooper, Cullen, & Sackett, 2006), and their job-relevance makes them more likely to be acceptable to users.

SJTs present a scenario (either textually or in video/animated clips) and ask the test taker to evaluate different courses of action that could be taken.<sup>11</sup> An SJT is a measurement method—a format for a test (Weekley, Ployhart, & Holz, 2006). What SJTs measure is a function of content choices made by developers. Even so, the method itself puts constraints on the types of constructs likely to be measured. That is, the characteristics of the SJT make it particularly suitable for measuring some constructs and unsuitable for measuring others. At the highest level,

---

<sup>11</sup> Many different SJT formats have been used in the past such as (a) pick the course of action you would take or (b) pick the best and worst actions. Formats that ask the test taker to evaluate different courses of action are generally more reliable than the other approaches and are expected to be less fakeable than formats that ask what the test taker would do.

SJTs simply measure judgment (Schmitt & Chan, 2006). Virtually all SJTs have a strong interpersonal component, and some SJTs rely more heavily on knowledge (and thus have a stronger positive relationship with cognitive ability) than others.

Because SJTs are typically multidimensional, internal consistency estimates are not high, nor are they expected to be (Schmitt & Chan, 2006). The most appropriate reliability estimate for SJTs is a test-retest estimate, but few organizations commit the resources needed to obtain it. In their meta-analysis of SJTs, McDaniel and colleagues (2001) found that internal consistency coefficients ranged from .43 to .94, with longer SJTs showing higher consistency. Others have found that the type of response instructions impacts reliability estimates, with “rate the effectiveness of each response” leading to higher internal consistency than “pick the best/worst” response format (Lievens, Peeters, & Schollaert, 2008). SJTs also yield reasonably high correlations with job performance ratings (average  $r = .34$  uncorrected) (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001).

*Recommendation 3.4.a. Develop an SJT to measure (a) judgment in Organizational Citizenship and Peer Leadership situations and (b) Technical Performance, with emphasis on Decision Making, Problem Solving, and Innovation. The SJT should be aimed at the level of skill expected of Service members well into or nearing the end of their first term, E3 – E4 level (18 to 36 months of experience).*

Over the years, the Army has developed four SJTs to serve as criterion measures. SJTs developed during Project A (Campbell & Knapp, 2010), NCO21 (Waugh & Russell, 2005), and *Expanding the Concept of Quality in Personnel* ([ECQUIP]; Peterson et al., 1999) projects were targeted toward non-commissioned officer (NCO) performance. One SJT developed in the *Select21* project (Knapp, Tremble, Russell, & Sellman, 2008) was developed to serve as a criterion measure for first-term enlisted performance. All four SJTs were paper-and-pencil versions. The two most recent SJTs from *Select21* and NCO21 would provide the best material for development of a new SJT.

The NCO21 SJT was a 40-item test with five items targeted to measure each of eight NCO skills: (a) directing, monitoring, and supervising individual subordinates, (b) training others, (c) team leadership, (d) concern for soldiers’ quality of life, (e) cultural tolerance, (f) motivating, leading, and supporting individual subordinates, (g) relating to and supporting peers, and (h) problem-solving and decision-making (Waugh & Russell, 2005). All NCOs completed the same 40-item form, but only 24 items were scored for each NCO rank (E5 and E6). NCOs were asked to pick the most effective and least effective response options. Total score reliabilities for the 24-item form ranged from .68 for E6s to .76 for E4s. While these reliabilities are not high for an objective format test, they are in line with what is typically observed for an SJT using this item format (pick best and pick worst). For E5s, it correlated .32 with supervisor ratings of performance, .28 with ratings of Senior NCO Potential, and .36 with an overall Effectiveness rating (Knapp, McCloy, & Heffner, 2004). Validities for E6s were lower than those for E5s: it correlated .25 with supervisor ratings, .18 with ratings of Senior NCO Potential, and .16 with an overall effectiveness rating.

The Air Force Personnel Center/Strategic Research and Assessment Branch (AFPC/DSYX) also has an NCO-level prototype SJT that was developed for use with E-7 (master sergeant) candidates to assess people/team competencies (Sullivan, Burgoyne, McCloy, & Whetzel, 2018). The prototype SJT consists of 25 items, each of which contains a scenario and four possible

actions (i.e., response options). The respondent selects one option as the most effective action and another option as the least effective action. While pilot testing indicated that internal consistency was very low (Waugh, Sullivan, & McCloy, 2019), it would provide a useful starting point for identifying relevant scenarios for the new SJT.

*Recommendation 3.4.b. Develop cross-Service SJT.*

We include this recommendation for a couple of reasons. First, SJTs are time-consuming and expensive to develop, and having one version across all Services would be an efficient use of resources. Second, the suggested criterion constructs (see Recommendation 3.4.a) lend themselves to scenarios that can generalize across Services.

Thus, our recommendation would be to develop SJT materials that contain scenarios and response options that are applicable across Services. For example, a scenario might involve dealing with a peer who is having personal problems, with options that include common actions that could be taken. Another option would be to have scenarios that have Service-specific settings but involve activities that generalize clearly across settings. Close work with SMEs from each Service would be needed to ensure consistency in interpretation across Services.

All the above said, we recognize that there may be Service-specific variation in ratings of effectiveness for different response options, due to factors such as culture. For this reason, while we recommend the same SJT across all Services, it may be necessary to develop Service-specific scoring keys to account for this variation.

*Recommendation 3.4.c. Focus the criterion SJT on IU performance during the first term of enlistment.*

Like Recommendation 3.3.c., we recommend this new SJT focus on IU performance because there are few measures of first term IU performance of *Organizational Citizenship & Peer Leadership* or *Decision Making, Problem Solving, and Innovation*. Additionally, first-term Service members gain more discretion after training, suggesting that these constructs are more likely to manifest themselves IU than in training.

*Recommendation 3.4.d. Enrich the SJT using online media (e.g., animation, video).*

Clearly computer-based animated or video-based SJTs are more expensive to develop than their paper and pencil counterparts. However, despite the added expense, animated and video SJTs have become increasingly popular. Research has found that they are (a) less contaminated by general mental ability (a 39% reduction in one study), (b) demonstrate higher convergent validity with established measures of similar constructs, (c) perceived more positively by test-takers, and (d) exhibit lower subgroup differences than text-based SJTs (Chan & Schmitt, 1997; Lievens & Sackett, 2006). In addition to cost, the inclusion of video also increases development time. Thus, depending on current priorities, it may make sense to start by developing a text-based SJT, then add video over time.

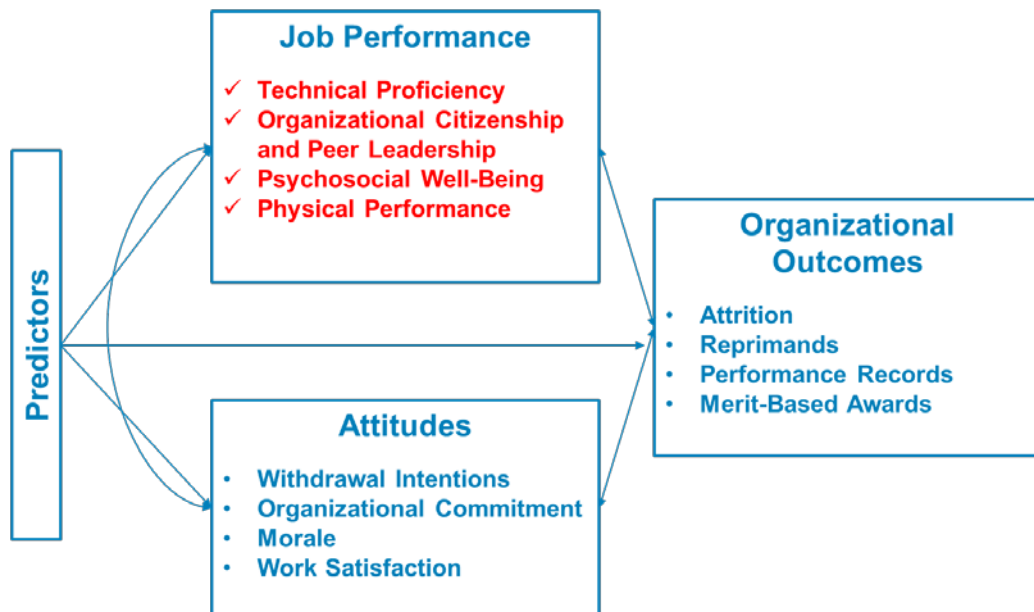
### **5.5.3 Tier 3 Summary**

The Tier 3 recommendations involve development of criterion measures to capture key parts of the criterion space not well-captured in Tiers 1 and 2. Development of these shared measures would significantly reduce the time and resources needed to conduct Service-specific validation studies by having off-the-shelf instruments ready to collect needed criterion data. Inclusion of these measures would capture the components of the criterion space illustrated in Figure 5.4.



We believe implementation of these Tier 3 recommendations would yield substantial benefits to the Services in future validation research. Specifically, these measures would:

- Assess key performance constructs not often included in Service-specific validation studies.
- Standardize the measurement of these performance constructs across Services
- Provide measures that are potentially useful for other purposes (e.g., training, evaluation)



**Figure 7. Criterion Domain Constructs Measured in Tier 3.**

## 6.0 SUMMARY AND CONCLUSIONS

The DoD uses the ASVAB to select new military recruits every year and place them in military occupations. Other predictor measures such as the TAPAS, interest inventories, and specialized tests to supplement the ASVAB are used to make personnel selection and assignment decisions. To ensure predictor measures are valid, the DoD and individual Services conduct rigorous, large-scale research projects to evaluate predictor measures against criterion metrics such as training/job performance or retention. However, criterion metrics are mostly Service-specific and sometimes occupation-specific, making it difficult to examine outcomes DoD-wide.

This report describes recommendations for standardizing criterion measurement across the Services in order to (a) facilitate robust comparisons of results within and across the Services and (b) strengthen DoD's conclusions about the validity and utility of the ASVAB and other predictors. Taxonomies of job performance, attitudes, and organizational outcomes were developed for first term enlisted Service personnel. A database of criterion measures used by the Services was developed and the criterion measures were linked to the performance domain constructs. Recommendations were made to develop a unified set of test evaluation criteria that can be used by all Services. The recommendations for development of measures are summarized in three tiers:

- Tier 1 – Maximize the use of administrative data.
  - Focuses on using only available administrative data to include data from the Service's personnel surveys.
  - Involves (a) developing standardized attrition and outcome variables and (b) aligning institutional personnel surveys to provide attitudinal and outcome data.
- Tier 2 – Improve measurement of attitudinal, outcome and performance constructs.
  - Short self-report tools are the best way of measuring attitudinal criteria and have been shown to yield reasonably accurate data on performance outcomes and Counterproductive Work Behavior.
  - Involves identifying/developing three short self-assessment tools, one for each of the following points in time: (a) end-of-technical training, (b) in-unit, and (c) exit.
- Tier 3 – Improve measurement of job performance.
  - The Services should measure:
    - Organizational Citizenship and Peer Leadership with job performance ratings, an SJT, or both and,
    - Technical performance with job performance ratings, a job knowledge test, or both.
  - Without additional measurement, beyond Tier 2, substantial portions of the criterion domain are not covered.

These recommendations were presented to the CMAP for consideration. Once the CMAP had provided feedback, development of a new set of joint-Service criterion measures was begun. The criterion measures were to cover as much of the job performance, attitudinal, and organizational outcome domains as possible, within the time and budget constraints of the project.

Tier 1 recommendations were addressed by developing statistical approaches for aligning variable construction across the Services. This entailed collecting and analyzing cross-Service

measures of attrition and training performance outcomes based on data in administrative databases (as described in Recommendations 1.1 and 1.2).

Tier 2 and 3 recommendations were addressed by adapting numerous criterion measures for cross-Service use. We developed (a) cross-Service self-assessment tools targeting end-of-training, in-unit, and exit milestones (as described in Recommendation 2.1) and (b) cross-Service job performance rating scales designed for use by peers and supervisors (as described in Recommendation 3.1). In addition, to more thoroughly measure constructs not measured by other tools (e.g., *Organizational Citizenship & Peer Leadership, Decision making, Problem Solving, and Innovation*), we developed a cross-Service in-unit SJT targeted toward the end of the first term E3-E4 (as described in Recommendation 3.4).

The development process for the new criterion measures are described in a separate report (Ford, Yu, Graves, Huber, Russell, & Wilmot, 2020).

## 7.0 REFERENCES

- Allen, M. T., Knapp, D. J., & Owen, K. S. (2016). *Validating future force performance measures (Army Class): Concluding analyses* (Technical Report 1355). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Alley, W. E., Pacheco, L. J., Birkelbach, D. B., Schwartz, K. L., & Weissmuller, J. J. (2007). *Modeling individual performance criteria in the Air Force* (AFCAPS-FR-2010-0015). Air Force Personnel Center.
- Arthur, W., Jr., Bell, S. T., Villado, A. J., & Doverspike, D. (2006). The use of person-organization fit in employment decision making: An assessment of its criterion-related validity. *Journal of Applied Psychology, 91*, 786-801.
- Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation, *Journal of Applied Psychology, 90*(6), 1185-1203.
- Berry, C. M., Carpenter, N. C., & Barratt, C. L. (2012). Do other-reports of counterproductive work behavior provide an incremental contribution over self-reports? A meta-analytic comparison. *Journal of Applied Psychology, 97*(3), 613-636.
- Borman, W. C., Grossman, M. R., Bryant, R. H., & Dorio, J. (2017). The measurement of task performance as criteria in selection research. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed.). Routledge.
- Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. Borman (Eds.), *Personnel selection in organizations* (p. 71-98). Jossey-Bass.
- Cable, D. M., & Edwards, J. R. (2004). Complementary and supplementary fit: A theoretical and empirical integration. *Journal of Applied Psychology, 89*, 822-834.
- Cable, D. M., & Judge, T. A. (1996). Person-organization fit, job choice decisions, and organizational entry. *Organizational Behavior and Human Decision Processes, 67*(3), 294-311.
- Campbell, J. P. (2012). Behavior, performance, and effectiveness in the 21<sup>st</sup> century. In S. Kozlowski (Ed.), *The Oxford handbook of organizational psychology: Volume 1*. Oxford University Press.
- Campbell, J. P., Hanson, M. A., & Oppler, S. H. (2001). Modeling performance in a population of jobs. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification*. Lawrence Erlbaum Inc.
- Campbell, J. P., & Knapp, D. J. (Eds.). (2001). *Exploring the limits in personnel selection and classification*. Lawrence Erlbaum Inc.
- Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations*. Jossey-Bass Publishers.
- Campbell, J. P., & Wiernik, B. M. (2015). The modeling and assessment of work performance. *Annual review of organizational psychology and organizational behavior, 2*: 47-74. <https://doi.org/10.1146/annurev-orgpsych-032414-111427>

- Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, 82, 143-159.
- Charles, E., & Moynihan, G. T. (2017, July). *USMC TAPAS Pilot Study Update*. A presentation made at July, 2017 meeting of the Defense Advisory Committee for Military Personnel Testing.
- Cornum, R., Matthews, M. D., & Seligman, M. E. P. (2011). Comprehensive soldier fitness: Building resilience in a challenging institutional context. *American Psychologist*, 66(1), 4-9.
- Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology*, 90(6), 1241-55.
- Dawis, R. V., & Lofquist, L. H. (1984). *A psychological theory of work adjustment: An individual-differences model and its applications*. The University of Minnesota.
- Department of Defense (2018). *Summary of the 2018 National Defense Strategy of The United States of America: Sharpening the American military's competitive edge*. Author.
- Dorsey, D. W., Cortina, J. M., Allen, M. T., Waters, S. D., Green, J. P., & Luchman, J. (2017). Adaptive and citizenship-related behaviors at work. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed.). Routledge.
- Edwards, J. R. (1996). An examination of competing versions of the person-environment fit approach to stress. *Academy of Management Journal*, 39, 292-239.
- Ford, L. A., Yu, M. C., Graves, C. R., Huber, C. R., Russell, T. L., & Wilmot, M. P. (2020). *Development of joint-service criterion instruments for enlisted jobs*, AFRL-RH-WP-TR-2020-xxxx. Wright-Patterson AFB, OH: 711 Hman Performance Wing, Warfighter Interface Division, Collaborative Interfaces and Teaming Branch.
- Forrester, J., O'Hanlon, M., & Zenko, M. (2001). Measuring U.S. military readiness. *National Security Studies Quarterly*, VII (2), 99-120.
- Garrison, D. R. (1997). Self-directed learning: Toward a comprehensive model. In *Adult Education Quarterly*, Fall 97, 48(1), 18-33.
- Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of race on organizational experiences, job performance, evaluations, and career outcomes. *Academy of Management Journal*, 33(1), 64-86.
- Goffin, R. D., Woycheshin, D. E., Hoffman, B. J., & George, K. (2013). The dimensionality of contextual and citizenship performance in military recruits: Support for nine dimensions using self-, peer, and supervisor ratings. *Military Psychology*, 25, 478-488.
- Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of race on organizational experiences, job performance, evaluations, and career outcomes. *Academy of Management Journal*, 33(1), 64-86.

- Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: positive behavior in uncertain and interdependent contexts. *Academy of Management Journal*, *50*(2), 327-347.
- Halper, M. L., Goodman, T. M., & Alley, W. E. (2010). *Months of mission-ready service*. Operational Technologies Corporation.
- Heidemeier, H., & Moser, K. (2009). Self–other agreement in job performance ratings: A meta-analytic test of a process model. *Journal of Applied Psychology*, *94*(2), 353-370.
- Hom, P. (2011). Organizational exit. In S. Zedeck (Ed.), *Handbook of industrial & organizational psychology, Vol. 2. Selecting and developing members for the organization* (p. 325-375). American Psychological Association.  
<https://psycnet.apa.org/doi/10.1037/12170-011>
- Hom, P., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). One hundred years of employee turnover theory and research. *Journal of Applied Psychology*, *102*(3), 530-545.  
<https://doi.org/10.1037/apl0000103>
- Hooper, A. C., Cullen, M. J., & Sackett, P. R. (2006). Operational threats to the use of SJTs: Faking, coaching, and retesting issues. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement and application*. Erlbaum.
- Hooper, A., Paullin, C., Putka, D. J., & Strickland, W. J. (2008). *An empirical analysis of reasons for attrition among first term airmen in the USAF* (FR-08-32). Human Resources Research Organization.
- HumRRO (2018, October). *Task Order FA8650-17-F-6828 (TO 7 – Task 4): Memorandum of Record*. Human Resources Research Organization.
- Ingerick, M., Allen, M., Weaver, E., Caramagno, J., & Hooper, A. (2006). *Retention incentives to mitigate deployment effects on soldier retention* (FFR-08-111). Human Resources Research Organization.
- Judge, T. A., & Kammeyer-Mueller, J. D. (2012). Job attitudes. *Annual Review of Psychology*, *63*, 341-367.
- Judge, T. A., Cable, D. M., Boudreau, J. W., & Bretz, R. D. (1994). *An empirical investigation of the predictors of executive career success* (CAHRS Working Paper Series). Cornell University ILR School.
- Judge, T. A., Weiss, H. M., Kammeyer-Mueller, J. D., & Husin, C. L. (2017). Job attitudes, job satisfaction, and job affect: A century of continuity and of change, *Journal of Applied Psychology*, *102*(3), 356-374.
- Judge, T. A., & Kammeyer-Mueller, J. D., (2012). Job attitudes. *Annual Review of Psychology*, *63*, 341-367. doi:10.1146/annurev-psych-120710-100511
- Kieckhafer, W. F., Ward, D. G., Kusulas, J. W., Cole, D. R., Rupp, L. M., & May, M. H. (1992). *Criterion development for 18 technical training schools in the Navy, Army, and Air Force* (Contract N66001-90-D-9502, DO 7J08). Navy Personnel Research and Development Center.

- Klafehn, J., Anderson, L., Taylor, W., Ingerick, M., & Ford, L. (2018, February). *Criterion development for evaluation of the cross-cultural Competence Assessment System* (Draft). Unpublished document.
- Knapp, D. J., & Campbell, R. C. (Eds.). (2006). *Army enlisted personnel competency assessment program: Phase II report* (Technical Report 1174). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & Campbell, R. C. (2004). *Army enlisted personnel competency assessment program Phase I (Volume I): Needs analysis* (Technical Report 1151). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & Campbell, J. P. (1993). *Building a joint-service classification research roadmap: Criterion-related issues* (AL/HR-TP-1993-0028, AD-A269 735). Human Resources Directorate Manpower and Personnel Research Division.
- Knapp, D. J., & Kirkendall, C. D. (2017). *Tier One Performance Screen initial operational test and evaluation: 2015-2016 Biennial Report* (Technical Report 1367). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., Sager, C. E. & Tremble, T. R. (Eds.) (2005). *Development of experimental Army enlisted selection and classification tests and job performance criteria* (Technical Report 1168). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & Tremble, T. R. (2007). *Concurrent validation of experimental Army enlisted personnel selection and classification measures* (Technical Report 1205). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., Tremble, T. R., Russell, T. L., & Sellman, W. S. (2008). *Future-oriented Army enlisted personnel selection and classification project (Select21) summary report* (Technical Report 1224). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., McCloy, R. A., & Heffner, T. S. (2004). *Validation of measures designed to maximize 21<sup>st</sup>-century Army NCO performance* (Technical Report 1145). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & Wolters, H. M. K. (2017). *Tier One Performance Screen initial operational test and evaluation: 2014 annual report* (Technical Report 1358). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Lance, C. F., Teachout, M. S., & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology*, 77, 437-452.
- LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, 6(1), 80-129. <https://doi.org/10.1177/1094428102239427>
- Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, 91(5), 1181-1188.

- Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, 37(4), 426-441.
- Macey, W. H., & Schneider, B. (2008). The meaning of employee engagement. *Industrial and Organizational Psychology*, 1, 3-30.
- Matthews, M. D., Eid, J., Johnsen, B. H., & Boe, O. C. (2011). A comparison of expert ratings and self-assessments of situation awareness during a combat fatigue course. *Military Psychology*, 23, 125-136.
- Mayberry, P. W. (1990). *Validation of the ASVAB against infantry job performance* (CRM 90-182). Center for Naval Analyses. <http://www.dtic.mil/dtic/tr/fulltext/u2/a235406.pdf>
- McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, 86, 730-740.
- Meyer, J. P., & Allen, N. J. (1991). A three-component conceptualization of organizational commitment. *Human Resource Management Review*, 1, 61-89.
- Meyer, J. P., Kam, C., Goldenberg, I., & Bremner, N. L. (2013). Organizational commitment in the military: Application of a profile approach. *Military Psychology*, 25(4), 381-401.
- Moriarty, K. O., & Knapp, D. J. (Eds.). (2007). *Army enlisted personnel competency assessment program: Phase III pilot tests* (Technical Report 1198). U.S. Army Research Institute for the Behavioral and Social Sciences.
- O'Hanlon, M. E. (2017, 12 October). *Is the Pentagon headed for a military readiness crisis?* Retrieved from <https://www.brookings.edu/blog/order-from-chaos/2017/10/12/is-the-pentagon-headed-for-a-military-readiness-crisis/>
- O'Leary, R. S., & Pulakos, E. D. (2017). Defining and measuring results of workplace behavior. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed.). Routledge.
- Oppler, S. H., Russell, T. L., Rosse, R. L., Keil, C. T., Meiman, E. P., & Welsh, J. R. (1997). *Item evaluation for the Armed Services Vocational Aptitude Battery (ASVAB) science and technical test specifications: Final Report* (DMDC Technical Report 97-024). Defense Manpower Data Center Personnel Testing Division.
- Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington Books D.C. Heath and Company.
- O'Shea, P. G., Goodwin, G. F., Driskell, J. E., Salas, E., & Ardison, S. (2009). The many faces of commitment: Facet-level links to performance in military contexts. *Military Psychology*, 21(1), 5-23.
- Peterson, N. G., Anderson, L. E., Crafts, J. L., Smith, D. A., Motowidlo, S. J., Rosse, R. L., Waugh, G. W., McCloy, R., Reynolds, D. H., & Dela Rosa, M. R. (1999). *Expanding the concept of quality in personnel: Final report*. U.S. Army Research Institute for the Behavioral and Social Sciences.



- Pulakos, E. D. (2007). Performance measurement. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance, *Journal of Applied Psychology*, 85, 612-624.
- Putka, D. J., & Allen, M. (2008). *An empirical evaluation of the United States Air Force's enlistment waiver policy* (FR-08-21). Human Resources Research Organization.
- Rotundo, M., & Spector, P. E. (2017). New perspectives on counterproductive work behavior including withdrawal. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed.). Routledge.
- Russell, T., Rosenthal, D., Paullin, C., & Putka, D. (2006). *Addressing active learning concepts in the Navy Applicant Management Information System (NAMIS)*. Unpublished report.
- Russell, T. L., Sparks, T. E., Campbell, J. P., Handy, K., Ramsberger, P., & Grand, J. A. (2017). Situating ethical behavior in the nomological network of job performance. *Journal of Business and Psychology*, 32(3), 253-271.
- Sager, C. E., Russell, T. L., Campbell, R. C., & Ford, L. A. (2005). *Future soldiers: Analysis of entry-level performance requirements and their predictors* (Technical Report 1169). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Schiefer, M., Robbert, A. A., Crown, J. S., Manacapilli, T., & Wong, C. (2008). *The weighted airman promotion system: Standardizing test scores*. RAND Corporation.
- Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement and application*. Erlbaum.
- Sellman, W. S., Russell, T. L., & Strickland, W. J. (2010). Selection and classification in the U.S. military. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection*, (2nd ed., pp. 697-721). Routledge, Taylor and Francis Group.
- Shuffler, M. L., Pavlas, D., & Salas, E. (2012). Teams in the Military: A review and emerging challenges. In J. H. Laurence & M. D. Matthews (Eds.), *The Oxford handbook of military psychology*. Oxford University Press.
- Sims, W. J., & Hiatt, C. (2001). *Marine Corps selection and classification* (CAB D00003683.A1/Final). Centers for Naval Analysis.
- Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology*, 58, 33-66.
- Spector, P. E., Bauer, J. A., & Fox, S. (2010). Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior: Do we know what we think we know? *Journal of Applied Psychology*, 95(4), 781-790.  
<https://dx.doi.org/10.1037/a0019477>

- Spector, P. E., Fox, S., Penney, L. M., Brursema, K., Goh, A., & Kessler, S. (2006). The dimensionality of counterproductivity: Are all counterproductive behaviors created equal? *Journal of Vocational Behavior*, 68, 446-460.
- Sullivan, T. S., Burgoyne, T. C., McCloy, R. A., & Whetzel, D. L. (2018). *Review of Situational Judgment Test (SJT) prototype development process and materials* (Technical Report 2018 No. 061). Human Resources Research Organization.
- Trippe, D. M., Moriarty, K. O., Russell, T. L., Carretta, T. R., & Beatty, A. S. (2014). Development of a cyber/information technology knowledge test for military enlisted technical training qualification. *Military Psychology*, 26, 182-198.
- Wasko, L., Owens, K. S., Campbell, R., & Russell, T. (2012). Development of the combat/deployment performance rating scales. In D. J. Knapp, K. S. Owens, & M. T. Allen (Eds.), *Validating future force performance measures (Army Class): In-unit performance longitudinal validation* (Technical Report 1314). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Wathen, W. G. (2014). *Identifying factors that predict promotion time to E-4 and reenlistment eligibility for U.S. Marine Corps Field Radio Operators*. Monterey, CA: Naval Postgraduate School. Downloaded from [https://calhoun.nps.edu/bitstream/handle/10945/44686/14Dec\\_Wathen\\_William.pdf?sequence=1&isAllowed=y](https://calhoun.nps.edu/bitstream/handle/10945/44686/14Dec_Wathen_William.pdf?sequence=1&isAllowed=y)
- Watson, S. (2016, August). *Rating Identification Engine (RIDE) Algorithm*. A presentation prepared by Dr. Steve Watson, Director, Navy Selection and Classification Office (N-132G)
- Waugh, G. W., & Russell, T. L. (2005). Predictor situational judgment test. In D. J. Knapp, C. E. Sager, & T. R. Tremble (Eds.), *Development of experimental Army enlisted selection and classification tests and job performance criteria* (Technical Report 1168). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Waugh, G. W., Sullivan, T. S., & McCloy, R.A. (2019). *Psychometric evaluation of a situational judgment test (SJT) prototype for the Weighted Airman Promotion System* (Technical Report 2019 No. 005). Human Resources Research Organization.
- Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement and application*. Erlbaum.
- Weiss, D., Dawis, R., England, G., & Lofquist, L. (1967). *Instrumentation for the theory of work adjustment*. University of Minnesota.
- Wigdor, A. K., & Green, B. F. (1991). *Performance assessment for the workplace: Volume I*. National Academy Press. <https://doi.org/10.17226/1862>
- Wolfe, J. H., Alderton, D. L., Larson, G. E., Bloxom, B. M., & Wise, L. L. (1997). Expanding the content of CAT-ASVAB: New tests and their validity. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation*. American Psychological Association.

Zaccaro, S. J., Laport, K., & Jose, I. (2012) The attributes of successful leaders: A performance requirements approach. *The Oxford handbook of leadership. Oxford Handbooks Online.* <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780195398793.001.0001/oxfordhb-9780195398793-e-1>

## **8.0 LIST OF ACRONYMS**

<b>AFPT</b>	Army Physical Fitness Test
<b>AFQT</b>	Armed Forced Qualification Test
<b>AF-WIN</b>	AF Work Interest Navigator
<b>ASVAB</b>	Armed Services Vocational Aptitude Battery
<b>BCA</b>	Body Composition Assessment
<b>CMAP</b>	Criterion Measures Advisory Panel
<b>CWB</b>	Counterproductive Work Behavior
<b>DEP</b>	Delayed Entry Program
<b>DMDC</b>	Defense Manpower Data Center
<b>DoD</b>	Department of Defense
<b>DPAC</b>	Defense Personnel Assessment Center
<b>DTIC</b>	Defense Technical Information Center
<b>ECAT</b>	Enhanced Computer Adaptive Test
<b>ECQUIP</b>	Expanding the Concept of Quality in Personnel
<b>EOT</b>	End of Training
<b>EOTT</b>	End of Technical Training
<b>FA</b>	Fitness Assessment
<b>ICC</b>	Interclass Correlation
<b>IMT ALQ</b>	Initial Military Training Army Life Questionnaire
<b>IMTA</b>	International Military Testing Association
<b>I-O</b>	Industrial and Organizational
<b>IU AIQ</b>	IU Army Life Questionnaire
<b>IU</b>	In-Unit
<b>JKT</b>	Job Knowledge Test
<b>JOIN</b>	Job Opportunities in the Navy
<b>JPM</b>	Joint Performance Measurement
<b>MAPWG</b>	Manpower Assessment Policy Working Group
<b>MM-RS</b>	Months of Mission Ready Service
<b>NCO</b>	Non-Commissioned Officer
<b>NPS</b>	Navy-side Personnel Survey
<b>Perform21</b>	Performance Measures for the 21st Century

<b>PFE</b>	Promotion Fitness Exam
<b>PFT</b>	Physical Fitness
<b>PRO Mark</b>	Proficiency Marks
<b>PRS</b>	Performance Rating Scales
<b>PRT</b>	Physical Readiness Test
<b>PWB</b>	Psychosocial Well-Being
<b>QMM</b>	Qualified Man Months
<b>QWL</b>	Quality of Work Life
<b>RTC</b>	Recruit Training Command
<b>SJT</b>	Situational Judgement Test
<b>SKT</b>	Skill/Knowledge Test
<b>SME</b>	Subject Matter Expert
<b>SSMP</b>	Sample Survey of Military Personnel
<b>TAPAS</b>	Tailored Adaptive Personality Assessment System
<b>TOPS</b>	Tier One Performance Screen
<b>USAF</b>	United States Air Force
<b>USMC</b>	United States Marine Corp
<b>USN</b>	United States Navy
<b>WAPS</b>	Weighted Airman Performance System
<b>WWL</b>	Quality of Work Life

## APPENDIX A - Retranslation Survey

### MEMORANDUM

To: CMAP and Retranslation Participants  
From: Teresa Russell, Matt Allen, and Laura Ford  
Re: Draft 1<sup>st</sup> Tour and Training Job Performance Taxonomy  
Date: 1 May 2018

Thank you for agreeing to participate in this retranslation exercise to evaluate taxonomic structures for Training and 1<sup>st</sup> Tour Military Job Performance. The taxonomy has two purposes. First and foremost it will provide a structure for describing the content of criterion instruments. Second, in later stages of this project, the taxonomy will be useful in guiding the development of new criterion instruments.

In the long run, the taxonomic structure can be tested empirically. In the meantime, we are asking researchers with performance measurement experience to categorize 33 specific dimensions identified in military and civilian literature into three different higher-order structures:

- Can-do/Technical Performance vs. Will-do/Contextual Performance
- Four Broad Categories (the Big Four)
- Ten Broad Dimensions (the Top Ten)

### **Development of the Draft Taxonomy**

To develop the draft dimensions, we gathered and integrated literature describing other taxonomies. Some of the taxonomies contained many constructs, others focused on a particular domain of constructs, and others were military specific. The Campbell Model is the most extensively researched and documented of the taxonomies and it emerged from military work. Therefore, we used it as a scaffold, sorting dimensions into it and adding dimensions where the fit was poor. It is important to note that we have borrowed heavily from other authors in writing or selecting dimension definitions. Primary sources will be credited in write-ups of the dimensions, but we are not citing them in this exercise.

### **Instructions**

The judgments you make, in combination with those of other researchers, will be used to evaluate taxonomic structures for 33 job performance dimensions.

Please follow these steps:

#### **Step 1. Prepare**

Open “retranslation exercise.xls” and rename it with your initials at the end. Read the 33 performance dimensions and their definitions. Note that we are presenting them in random order.

#### **Step 2. Make Can-do/Technical vs. Will-do/Contextual Performance Judgements**

Go to the Can-do/Will-do column in the Rating Tab.

Review the following definitions of Can-do/Task vs. Will-do/Contextual Performance:

*Can-do/Task performance* – performance of activities that contribute to the organization’s technical core. Task activities usually vary between different jobs in the same organization. Task performance is usually predicted by knowledge, skills, and abilities. Task performance is role-prescribed, i.e., formally recognized as part of the job. Can-do performance is typically measured using maximal performance instruments.

*Will-do/Contextual performance* – performance of activities that support the organizational, social, and psychological environment, e.g., organizational citizenship behaviors. Contextual activities are important across jobs. Motivational and personality characteristics are key determinants of contextual performance. Contextual activities may not be role-prescribed. Will-do performance is typically measured using typical performance instruments.

Borman and Motowidlo (1993) talk about specific performance dimensions as being saturated with either task or contextual elements. Dimensions can vary somewhat in terms of how reliant performance is on task or contextual elements. So, for example, some dimensions might be saturated with both task and contextual elements. Managerial dimensions that involve both planning and dealing with people might be the best example of dually-saturated dimensions.

Based on your experience and knowledge of job performance prediction, we would like you to judge the Can-do/Technical vs. Will-do/Contextual saturation of each of the 33 specific dimensions using the following scale:

- 2 = Can-do/task
- 1 = Can-do/task with some will-do saturation
- 0 = Equally can-do/task and will-do/contextual
- +1 = Will-do/contextual with some can-do/task saturation
- +2 = Will-do/contextual

Use the pull-down menu to record your judgement.

Please **read each dimension title and the full dimension carefully**. Make note of any concerns about the dimension definition in the Comment column.

Please **make your Can-do/Will-do judgments for all 33 dimensions before moving to the next judgment**. But as you make your Big Four judgments below, do feel free to adjust your can-do/will-do judgments if you feel upon further reflection that changes are warranted.

### Step 3. Make Big Four Judgements

Go to the Big Four column in the Rating Tab.

Review the following definitions of the Big Four dimensions:

Technical Performance	Performing core job tasks including the full range of cognitive tasks (e.g., analyzing intelligence data), skilled tasks (e.g., driving a truck), and interpersonal tasks (e.g.,
-----------------------	--

providing administrative personnel support that require skill or knowledge.

Counterproductive Work Behavior	Exhibiting deviant behaviors directed at the organization or at individuals.
Citizenship & Peer Leadership	Demonstrating conscientious initiative, support for peers and the organization, and responsible self-management.
Physical Performance	Performing physical tasks, sometimes over long periods of time in difficult conditions.

Indicate which of the Big Four dimensions best reflects each specific dimension, i.e., categorize the 33 specific dimensions into the Big Four.

Use the pull-down menu to record your judgement.

Please **make your Big Four judgements for all 33 dimensions before moving to the next judgment.** But as you make your Top 10 judgments below, do feel free to adjust your previous judgments if you feel upon further reflection that changes are warranted.

#### Step 4. Make Top 10 Judgements

Go to the Top 10 column in the Rating Tab.

Review the following definitions of the Top 10 dimensions:

1. Technical Performance  
Performing core job tasks including the full range of cognitive tasks (e.g., analyzing intelligence data), skilled tasks (e.g., driving a truck), and interpersonal tasks (e.g., providing administrative personnel support that require skill or knowledge).
2. Communication  
Exchanging information between a sender and a receiver irrespective of the medium.
3. Conscientious Initiative  
Working extra hours, voluntarily taking on additional tasks, pursuing own development, going beyond prescribed responsibilities, or working under extreme or adverse conditions.
4. Counterproductive Work Behavior  
Exhibiting deviant behaviors directed at the organization or at individuals.
5. Support for Peers  
Showing consideration and support for peers by helping them, cooperating, showing respect, tolerating differences, modelling good behavior, and motivating others.



6. Initiating Structure	Establishing a course of action when in a leadership role or when working as a team member.
7. Individual Work Responsibility	Making decisions responsibly, managing personal and work commitments, and acting with honesty and integrity.
8. Organizational Support	Supporting the organization by following rules, demonstrating loyalty, and representing the organization.
9. Health and Safety Consciousness	Being attentive to and taking steps to ensure the health and safety of self and others.
10. Physical Performance	Performing physical tasks, sometimes over long periods of time in difficult conditions.

Indicate which of the Top 10 dimensions best reflects each specific dimension, i.e., categorize the 33 specific dimensions into the Top 10.

Use the pull-down menu to record your judgement. **Please note that only the first 8 dimensions show in the pull-down menu. You have to scroll to see the last two dimensions.**

Feel free to sort the spreadsheet on different columns to evaluate your judgements. Please return your spreadsheet to item number ordering before sending it in.

Step 5. Provide Comments

Your spreadsheet provides a comment column. Please let us know if you found awkward wording or were troubled by some aspect of a dimension definition.

Step 6. Turn in Your Spreadsheet

Please email your completed spreadsheet.

Thanks so much! Your ratings in combination with those of other raters will be used to identify a taxonomic structure for use in this project.

## APPENDIX B - Retranslation Results

### Table B1. Summary of Retranslation Results

Dimension	2D (N=17) <sup>1</sup>		4D (N=17) <sup>2</sup>				10D (N=16) <sup>3</sup>									
	MN	SD	TP	CWB	CPL	PP	TP	COM	CI	CWB	SP	IS	IWR	OS	HS	PP
A. Technical Performance																
A.1. Task Proficiency																
General Proficiency	-1.53	1.07	17	0	0	0	16	0	0	0	0	0	0	0	0	0
Job-Specific Proficiency	-1.59	.62	17	0	0	0	16	0	0	0	0	0	0	0	0	0
A.2. Decision Making, Problem Solving, and Innovation (Judgment)																
Decision Making, Problem Solving, and Innovation	-1.18	.88	17	0	0	0	9	0	0	0	0	2	5	0	0	0
A.3. Communication																
Written Communication	-1.29	1.05	17	0	0	0	1	15	0	0	0	0	0	0	0	0
Oral Communication	-1.12	.99	16	0	1	0	0	16	0	0	0	0	0	0	0	0
Nonverbal Communication	-.41	1.00	15	0	2	0	0	16	0	0	0	0	0	0	0	0
A.4. Safety Consciousness																
Safety Consciousness during Mission Operations	-.47	1.23	13	0	4	0	2	0	2	0	0	0	0	0	12	0
Safety Consciousness in Everyday Work	.24	1.20	9	0	8	0	0	0	0	0	0	0	1	0	15	0
B. Organizational Citizenship & Peer Leadership																
B.1. Planning and Structuring Work																
Providing Structure	-.59	1.06	6	0	11	0	1	0	0	0	1	13	0	1	0	0
Teamwork	-.06	1.25	8	1	8	0	3	0	0	0	5	8	0	0	0	0
Managing Responsibilities	.29	1.26	5	1	11	0	1	0	1	0	0	0	14	0	0	0
Learning Self-Management	.35	1.32	7	0	10	0	1	0	7	0	0	0	8	0	0	0

<sup>1</sup>Can/do - will/do rating scale: -2 = Can-do/task, -1 = Can-do/task with some will-do saturation, 0 = Equally can-do/task and will-do/contextual, +1 = Will-do/contextual with some can-do/task saturation, +2 = Will-do/contextual

<sup>2</sup>TP = Technical Performance, CWB = Counterproductive Work Performance, CPL = Citizenship and Peer Leadership Performance, and PP = Physical Performance

<sup>3</sup>TP = Technical Performance, COM = Communication, CI = Conscientious Initiative, CWB = Counterproductive Work Performance, SP = Support for Peers, IS = Initiating Structure, IWR = Individual Work Responsibility, OS = Occupational Support, and PP = Physical Performance

<sup>4</sup>Formerly Tolerating Differences - edited to more accurately reflect the dimension.

<sup>5</sup>Formerly Complying with Organizational Rules and Procedures - edited to more accurately reflect the full dimension.

**Table B1. Summary of Retranslation Results (Continued)**

Dimension	2D (N=17) <sup>1</sup>		4D (N=17) <sup>2</sup>				10D (N=16) <sup>3</sup>									
	MN	SD	TP	CWB	CPL	PP	TP	COM	CI	CWB	SP	IS	IWR	OS	HS	PP
<b>B.2. Conscientious Initiative</b>																
Active Learning	.88	1.17	3	0	14	0	0	0	15	0	0	0	1	0	0	0
Self-Development	1.35	.61	4	0	13	0	1	0	15	0	0	0	0	0	0	0
Persistence	1.18	1.19	3	0	14	0	0	0	15	0	0	0	1	0	0	0
Initiative	1.41	.87	0	0	17	0	0	0	14	0	0	1	1	0	0	0
<b>B.3 Support for Peers</b>																
Helping Peers	.94	.83	2	0	15	0	0	0	0	0	15	1	0	0	0	0
Cooperating	1.47	.51	0	0	17	0	0	0	0	0	15	0	0	1	0	0
Courtesy & Respect	1.53	.87	0	0	17	0	0	1	1	0	13	0	1	0	0	0
Accepting Differences <sup>4</sup>	1.76	.44	0	0	17	0	0	0	1	0	15	0	0	0	0	0
Motivating	1.76	.56	1	0	16	0	0	0	0	0	16	0	0	0	0	0
Serving as a Model	1.76	.56	1	0	16	0	0	0	2	0	10	1	1	2	0	0
<b>B.4 Organizational Support</b>																
Military Presence	1.65	.70	1	0	16	0	1	0	0	0	0	1	2	12	0	0
Selfless Service	1.82	.73	0	0	17	0	0	0	1	0	3	0	0	12	0	0
Support for the Organization <sup>5</sup>	.76	1.30	2	0	15	0	0	0	0	0	0	0	1	15	0	0
Integrity/Moral Courage	1.35	1.06	0	0	17	0	0	0	0	0	0	1	15	0	0	0
<b>C. Psychosocial Well-Being</b>																
<b>C.1. Well-being</b>																
Maintaining own Well-Being	1.38	.62	1	0	14	2	0	0	1	0	0	0	3	0	12	0
<b>C.2. Counterproductive Work Behavior</b>																
Loafing and Tardiness	1.53	1.07	0	17	0	0	1	0	0	15	0	0	0	0	0	0
Abusing Substances and Other Self-Destructive Behavior	1.75	.58	0	15	2	0	0	0	0	14	0	0	1	0	1	0
Bullying, Harassing, or Hurting Others	1.82	.53	0	17	0	0	0	0	0	14	1	0	0	0	0	0
Delinquency	1.76	.56	0	17	0	0	0	0	0	15	0	0	0	1	0	0
<b>D. Physical Performance</b>																
Physical Endurance	-.25	1.53	0	0	0	17	0	0	0	0	0	0	0	0	0	16
Physical Fitness	-.35	1.37	0	0	1	16	0	0	0	0	0	0	0	2	3	11

<sup>1</sup>Can/do - will/do rating scale: -2 = Can-do/task, -1 = Can-do/task with some will-do saturation, 0 = Equally can-do/task and will-do/contextual, +1 = Will-do/contextual with some can-do/task saturation, +2 = Will-do/contextual

<sup>2</sup>TP = Technical Performance, CWB = Counterproductive Work Performance, CPL = Citizenship and Peer Leadership Performance, and PP = Physical Performance

<sup>3</sup>TP = Technical Performance, COM = Communication, CI = Conscientious Initiative, CWB = Counterproductive Work Performance, SP = Support for Peers, IS = Initiating Structure, IWR = Individual Work Responsibility, OS = Occupational Support, and PP = Physical Performance

<sup>4</sup>Formerly Tolerating Differences - edited to more accurately reflect the dimension.

<sup>5</sup>Formerly Complying with Organizational Rules and Procedures - edited to more accurately reflect the full dimension

## APPENDIX C - Final Performance Construct Definitions

### Table C.1. Hierarchical Trainee and 1st Term Performance Taxonomy Definitions

Performance Category	Sub-Category	Specific Dimension	Definition
A. Technical Performance			Performing job tasks proficiently; communicating clearly; making sound decisions; and being alert to safety and security concerns.
	A.1. Task Performance		Being able to perform job-specific and Service-wide tasks proficiently
		Job-Specific Proficiency	Being able to perform job-specific tasks at the appropriate skill level.
		General Proficiency	Being able to perform Service-wide tasks at the appropriate skill level (e.g., navigation in the Army and Marine Corps).
	A.2. Decision Making, Problem Solving, and Innovation		Making sound, timely decisions, even under pressure; analyzing situations and innovating solutions to problems; resolving conflicts; adapting plans and decisions as situations change.
		Decision Making, Problem Solving, and Innovation	Making sound, timely decisions, even under pressure; analyzing situations and innovating solutions to problems; resolving conflicts; adapting plans and decisions as situations change.
	A.3. Communication		Conveying oral and written information clearly; using appropriate nonverbal communication.
		Oral Communication	Conveying information in a clear, understandable, organized manner when speaking.
		Written Communication	Conveying information in a clear, understandable, organized manner when writing.
		Nonverbal Communication	Using alternative, culturally appropriate methods to interpret and convey meaning when common language is not shared.
	A.4. Safety and Security Consciousness		Following routine safety and security guidelines; and being alert to safety and security threats in non-routine situations.
		Safety and Security Consciousness in Everyday Work	Following safety and security guidelines and instructions, noticing and alerting others to potential hazards in day-to-day work.
		Safety and Security Consciousness during Mission Operations	Being alert to enemy and environmental threats and taking actions that do not place self or others at unwarranted risk.

*Note.* Dimension definitions draw heavily from published literature, particularly Campbell and Wiernik (2015), Dorsey et al. (2017), Russell et al. (2006), Sager et al. (2005), Shuffler, Pavlas, & Salas (2012), Spector, Bauer, & Fox (2010), Wasko et al. (2012), and Waugh & Russell (2005).

**Table C.1. Hierarchical Trainee and 1st Term Performance Taxonomy Definitions (Continued)**

<b>Performance Category</b>	<b>Sub-Category</b>	<b>Specific Dimension</b>	<b>Definition</b>	
B. Organizational Citizenship & Peer Leadership	B.1. Planning and Structuring Work		Planning and structuring own work, and when in a leadership role, the work of others; taking initiative and persisting in work or training despite difficult conditions; supporting, helping, motivating, and respecting peers; and showing commitment to the organization, the team, and moral/ethical principles.	
			Leading peer when given a leadership role; working with team members to plan work; planning and organizing own responsibilities and studying.	
		Providing Structure	Leading peers when given a leadership role, giving clear instructions, distributing tasks, and gaining others' cooperation.	
		Teamwork	Working with other team members to interpret the mission, set and prioritize team goals, and monitor team performance.	
		Self-Management	Managing own responsibilities (e.g., work assignments, gear, equipment, personal finances, family, and personal well-being), and appearing on duty prepared for work. Setting personal work objectives.	
		Learning/Training Self-Management	Planning, organizing, and using study time effectively (e.g., setting aside specific times to study; completing assignments on time).	
		B.2. Conscientious Initiative		Taking initiative; persisting with extra effort despite obstacles; taking steps to enhance own knowledge and skill.
			Classroom Learning	Being actively engaged in own learning by searching for and obtaining information, taking notes in class, highlighting relevant material, practicing new skills, and participating/contributing during classes.
			Self-Development	Developing or adapting own knowledge and skills by taking courses on own time, volunteering for training and development opportunities offered within the organization; and trying to learn new knowledge and skills on the job from others or through new job assignments.
			Persistence	Persisting with extra effort despite difficult conditions and setbacks, accomplishing goals that are more difficult and challenging than normal completing work on time despite unusually short deadlines, and performing at a level of excellence that is significantly beyond normal expectations.
			Initiative	Taking the initiative to do all that is necessary to accomplish team or organizational objectives encountered, finding additional work to perform when own duties are completed, and volunteering for work assignments.

**Table C.1. Hierarchical Trainee and 1st Term Performance Taxonomy Definitions (Continued)**

<b>Performance Category</b>	<b>Sub-Category</b>	<b>Specific Dimension</b>	<b>Definition</b>
	B.3. Support for Peers		Helping and motivating peers; cooperating with others; being respectful and considerate; accepting individual differences; and modeling core values.
		Helping Peers	Helping others by offering suggestions about their work, showing them how to accomplish difficult tasks, teaching them useful knowledge or skills, directly performing some of their tasks, and providing emotional support for personal problems.
		Cooperating	Cooperating with others by accepting their suggestions, following their lead, being open-minded and adapting to others' ways, and informing others of events or requirements that are likely to affect them.
		Courtesy & Respect	Showing consideration, courtesy, and tact in relations with others.
		Accepting Differences	Showing interest in and respect for people of other backgrounds or cultures by regularly engaging with them in a manner considerate of their norms.
		Motivating	Motivating others by applauding their achievements and successes, cheering them on in times of adversity, showing confidence in their ability to succeed, helping them overcome setbacks, and modelling leadership behavior.
		Serving as a Model	Modeling core values by acting unselfishly, enduring hardships without complaint, treating others well, behaving ethically, and showing confidence and enthusiasm.
	B.4. Organizational Support		Complying with organizational rules; demonstrating selfless service; presenting a positive image of the Service; and demonstrating honesty and integrity.
		Military Presence	Presenting a positive and professional image of self and the military even when off duty, maintaining proper military appearance.
		Selfless Service	Committing to the greater good of the team or group putting organizational welfare ahead of individual goals.
		Support for the Organization	Complying with organizational rules and procedures, encouraging others to comply with organizational rules and procedures, and suggesting procedural, administrative, or organizational improvements.
		Integrity/Moral Courage	Demonstrating honesty and integrity in job-related matters, even when own self-interests might be jeopardized, taking steps to protect the security of military equipment/supplies, and voluntarily reporting thefts, misconduct, and any other violations of military order and discipline.

**Table C.1. Hierarchical Trainee and 1st Term Performance Taxonomy Definitions (Continued)**

Performance Category	Sub-Category	Specific Dimension	Definition
C. Psychosocial Well-Being			Maintaining emotional control in stressful situations; and <b>not</b> engaging in counterproductive work behaviors.
	C.1. Adapting to Stressful Situations		Maintaining emotional control in stressful situations; noticing/monitoring own signs of stress from combat, work and home life and taking positive steps in managing stress reactions.
		Adapting to Stressful Situations	Maintaining emotional control in stressful situations; noticing/monitoring own signs of stress from combat, work and home life and taking positive steps in managing stress reactions.
	C.2. Counterproductive Work Behavior		<b>Not</b> engaging in delinquent behaviors or behaviors that affect the productivity of the organization (e.g., loafing, tardiness); <b>not</b> bullying, harassing, or hurting others; and <b>not</b> engaging in self-destructive behaviors.
		Loafing and Tardiness	Arriving late for work or not showing up; spending work time on personal activities (e.g., surfing the web).
		Abusing Substances and Other Self-Destructive Behavior	Engaging in self-destructive behavior (e.g., alcohol or drug abuse).
		Bullying, Harassing, or Hurting Others	Engaging in deviant behavior directed at others (e.g., physical attacks, verbal abuse, harassment).
		Delinquency	Engaging in deviant behaviors directed at the organization (e.g., theft, sabotage).
D. Physical Performance			Meeting fitness standards and sustaining physical performance over time.
	D.1. Physical Endurance		Sustaining physical performance over long periods of time despite lack of sleep and difficult conditions. Adapting to environmental challenges (e.g., weather, terrain).
		Physical Endurance	Sustaining physical performance over long periods of time despite lack of sleep and difficult conditions. Adapting to environmental challenges (e.g., weather, terrain).
	D.2. Physical Fitness		Meeting military standards for weight, physical fitness, and strength, maintaining own health.
		Physical Fitness	Meeting military standards for weight, physical fitness, and strength, maintaining own health.

## APPENDIX D - Attitudinal and Organizational Outcome Construct Definitions

**Table D.1. Attitudinal Criterion Domain Taxonomy**

Construct	Definition	Facets
Work Satisfaction	An individual's satisfaction with work	<ul style="list-style-type: none"> <li>- Whole job satisfaction</li> <li>- Job facet satisfaction</li> <li>- Career satisfaction</li> </ul>
Morale	A holistic judgment of one's own morale	
Organizational Commitment	An individual's psychological bond with the organization, as represented by an affective attachment to the organization, internalization of its values and goals, and a behavioral desire to put forth effort to support it.	<ul style="list-style-type: none"> <li>- Affective</li> <li>- Continuance</li> <li>- Normative</li> </ul>
Withdrawal	Thinking about or intending to quit	
Cognitions/Intentions	one's job	<ul style="list-style-type: none"> <li>- Attrition cognitions</li> <li>- Short-term active duty career continuance intentions</li> <li>- Long-term active duty career continuance intentions</li> <li>- Post-active duty plans</li> <li>- Deployment-attributed change in career intentions</li> </ul>
Person-Environment Fit (PE Fit)	Congruence between the individual's abilities, needs, and expectations and characteristics of the organization, job or group.	<ul style="list-style-type: none"> <li>- Person-Job, Needs-supplies fit</li> <li>- Person-job, Demands-abilities fit</li> <li>- Person-organization fit</li> <li>- Person-team fit</li> </ul>

*Note.* Based primarily on Allen, Knapp, & Owens (2016), Arthur, Bell, Villado, & Doverspike (2006), Cable & Edwards (2004), Cable & Judge (1997), Dawis & Lofquist (1984), Edwards (1996), Greenhaus, Parasuraman & Wormley (1990), Hom (2011), Hom, Lee, Shaw, & Hausknecht (2017), Judge, Cable, Boudreau, & Bretz (1994), Judge & Kammeyer-Mueller (2012), Judge, Weiss, Kammeyer-Mueller, & Husin (2017), Meyer & Allen (1991), Meyer, Kam, Goldenberg, & Bremner (2013), Weiss, Dawis, Lofquist, & England (1967).



**Table D.2. Organizational Outcome Taxonomy**

<b>Outcome Construct</b>	<b>Facet</b>	<b>Example Indicators</b>
Attrition	Delayed entry program (DEP)	- DEP attrition
	Boot Camp	- Attrition from boot camp
	Advanced Training	- Attrition from advanced training
	In-unit	- Attrition in-unit (premature attrition)
	Re-enlistment	- Re-enlistment for 2nd term; propensity to re-enlist
Reprimands	Reprimands	- Articles 15/ reprimands
Experience	Tenure	- Time in grade/rank - Time in uniform/Length of service
	Rank	- Rank
Initiative	Awards	- Merit-based awards and commendations
Performance	Advanced Training	- Training school grades - Pass/Fail - Rank in class - Training recycles/Wash-backs
	In-unit	- Supervisor performance ratings/ Enlisted Performance Ratings (EPR in USAF)/ Proficiency marks (PRO marks in USMC) - Job knowledge test scores (e.g., USAF Skill/Knowledge Test [SKT]; USAF Promotion Fitness Examination [PFE])
	Skill Upgrading	- Skill level attainment (e.g., USAF skill level badges)
	Promotion Potential	- Promotion exam scores
	Physical	- Current physical fitness
	Qualifications	- Rifle/pistol qualification score - Other qualifications (swim, brown belt, Ranger)
	Re-enlistment Eligibility	- Computed Tier Score (re-enlistment eligibility composite based on a number of qualifications)
Productivity	Skilled Tenure	- Qualified man months (QMM - number of months in service at qualified level based on skills test)
	Skilled Tenure	- Months mission ready service (months of service at the highest skill level)
	Quantity of Performance	- Productive capacity (rate of task performance)
Promotion	Rate	Promotion rate (a deviation score comparing to other service members with the same time in service and in the same job)
	Time	Promotion time to E-4

*Note.* Indicators were drawn primarily from Alley, Pacheco, Birkelbach, Schwartz & Weissmuller (2007), Campbell & Knapp (2001), Halper, Goodman, & Alley (2010), Ingerick, Allen, Weaver, Caramagno, & Hooper (2006), Knapp & Campbell, 1993, Mayberry (1990), Sims & Hiatt (2001), and Wathen (2014).

## APPENDIX E - Data Entry Survey Tool

### Section 1. Identification

Please enter your first and last name.

### Section 2. Content Area

#### Basic Information

##### Name of Measure

Please enter the name of the criterion instrument as it appears in the technical documentation. If the criterion instrument is referred to by different names, please describe the instrument as best as possible.

##### Content Definition

Please enter the original content definition of the criterion being measured as it appears in the technical documentation. For example, "The IMT ALQ was designed to measure Soldiers' self-reported attitudes, experiences, and training performance in the Army."

If there are subscales/dimensions, please list them here.

##### Taxonomic Dimensions

Please select the dimension(s) that best represents what the criterion instrument is measuring based on the following definitions. Though you may select more than one option, it is best to treat administrative / outcome criteria separately from criteria collected for research purposes.

**Administrative / Outcome Criteria** – Criteria that reflect an outcome rather than discrete dimensions, and are gathered from administrative records maintained by a DOD component. In general, several performance and motivational dimensions will contribute to one or more of these outcomes.

- Attrition** – Attrition includes voluntary or involuntary separation from a service, which may occur during a term of service or after (e.g., re-enlistment).
- Reprimands** – Records of formal disciplinary action against a service member.
- Experience** – Such as time in grade, rank, or specific experiences (e.g., deployments).
- Performance Records** – These may include training outcomes (e.g., school grades, class rank, pass/fail a course), performance evaluation scores, or qualifications (e.g., physical fitness qualification, weapon qualifications).
- Promotion** – Promotion outcomes include rate of promotion, time to be promoted, etc.
- Awards** – Merit-based awards and commendations.
- Productivity** – Includes composites of one or more elements above to attain metrics of service quality (e.g., number of qualified months of service, combining experience in tenure).
- Other** – Criteria derived from administrative records that don't fit into one or more of the above categories.

---

**Performance Dimensions** – These dimensions reflect direct measures of on-the-job performance, collected through the use of maximal (e.g., simulations, tests) or typical (e.g., supervisor ratings) performance measures.

- General Proficiency [Technical]** – Being able to perform service-wide tasks at the appropriate skill level (e.g., navigation in the Army and Marine Corps).
- Job-Specific Proficiency [Technical]** – Being able to perform job-specific tasks at the appropriate skill level.
- Decision Making, Problem Solving, and Innovation [Technical]** – Making sound, timely decisions, even under pressure; analyzing situations and innovating solutions to problems; resolving conflicts; adapting plans and decisions as situations change.
- Oral Communication [Technical]** – Conveying information in a clear, understandable, organized manner when speaking.
- Nonverbal Communication [Technical]** – Using alternative, culturally appropriate methods to interpret and convey meaning when common language is not shared.
- Written Communication [Technical]** – Conveying information in a clear, understandable, organized manner when writing.
- Safety and Security Consciousness in Everyday Work [Technical]** – Following safety guidelines and instructions, noticing and alerting others to potential hazards in day-to-day work.
- Safety and Security Consciousness during Mission Operations [Technical]** – Being alert to enemy and environmental threats, and taking actions that do not place self or others at unwarranted risk.
- Classroom Learning [Organizational Citizenship & Peer Leadership]** – Being actively engaged in own learning by searching for and obtaining information, taking notes in class, highlighting relevant material, practicing new skills, and participating/contributing during classes.
- Persistence [Organizational Citizenship & Peer Leadership]** – Persisting with extra effort despite difficult conditions and setbacks, accomplishing goals that are more difficult and challenging than normal, completing work on time despite unusually short deadlines, and performing at a level of excellence that is significantly beyond normal expectations.
- Self-Development [Organizational Citizenship & Peer Leadership]** – Developing or adapting own knowledge and skills by taking courses on own time, volunteering for training and development opportunities offered within the organization, and trying to learn new knowledge and skills on the job from others or through new job assignments.

- Initiative [Organizational Citizenship & Peer Leadership]** – Taking the initiative to do all that is necessary to accomplish team or organizational objectives encountered, finding additional work to perform when own duties are completed, and volunteering for work assignments.
- Learning/Training Self-Management [Organizational Citizenship & Peer Leadership]** – Planning, organizing, and using study time effectively (e.g., setting aside specific times to study; completing assignments on time).
- Integrity/Moral Courage [Organizational Citizenship & Peer Leadership]** – Demonstrating honesty and integrity in job-related matters, even when own self-interests might be jeopardized, taking steps to protect the security of military equipment/supplies, and voluntarily reporting thefts, misconduct, and any other violations of military order and discipline.
- Self-Management [Organizational Citizenship & Peer Leadership]** – Managing own work and personal responsibilities (e.g., work assignments, gear, equipment, personal finances, family, and personal well-being), and appearing on duty prepared for work. Setting personal work objectives.
- Cooperating [Organizational Citizenship & Peer Leadership]** – Cooperating with others by accepting their suggestions, following their lead, being open-minded and adapting to others' ways, and informing others of events or requirements that are likely to affect them.
- Courtesy & Respect [Organizational Citizenship & Peer Leadership]** – Showing consideration, courtesy, and tact in relations with others.
- Helping [Organizational Citizenship & Peer Leadership]** – Helping others by offering suggestions about their work, showing them how to perform difficult tasks, teaching them useful knowledge or skills, directly performing some of their tasks, and providing emotional support for personal problems.
- Motivating [Organizational Citizenship & Peer Leadership]** – Motivating peers by applauding their achievements and successes, cheering them on in times of adversity, showing confidence in their ability to succeed, and helping them overcome setbacks, and modelling leadership behavior.
- Serving as a Model [Organizational Citizenship & Peer Leadership]** – Modelling core values by acting unselfishly, enduring hardships without complaint, treating others well, behaving ethically, and showing confidence and enthusiasm.
- Accepting Differences [Organizational Citizenship & Peer Leadership]** – Showing interest in and respect for people of other backgrounds or cultures by regularly engaging with them in a manner considerate of their norms.
- Organizational Support [Organizational Citizenship & Peer Leadership]** – Complying with organizational rules and procedures, encouraging others to comply with organizational rules and procedures, and suggesting procedural, administrative, or organizational improvements.
- Military Presence [Organizational Citizenship & Peer Leadership]** – Presenting a positive and professional image of self and the military even when off duty, maintaining proper military appearance.
- Selfless Service [Organizational Citizenship & Peer Leadership]** – Committing to the greater good of the team or group, putting organizational welfare ahead of individual goals.
- Providing Structure [Organizational Citizenship & Peer Leadership]** – Leading peers when given a leadership role, giving clear instructions, distributing tasks, and gaining others' cooperation.
- Teamwork [Organizational Citizenship & Peer Leadership]** – Working with other team members to interpret the mission, set and prioritize team goals, and monitor team performance.
- Fitness [Physical Performance]** – Meeting military standards for weight, physical fitness, and strength, maintaining own health.
- Endurance [Physical Performance]** – Sustaining physical performance over long periods of time despite lack of sleep and difficult conditions. Adapting to environmental challenges (e.g., weather, terrain).

- Abusing Substances and Other Self-Destructive Behavior [Psychosocial Well-Being]** – Engaging in self-destructive behavior (e.g., alcohol or drug abuse).
  - Bullying, Harassing, or Hurting Others [Psychosocial Well-Being]** – Engaging in deviant behavior directed at others (e.g., physical attacks, verbal abuse, harassment).
  - Loafing and Tardiness [Psychosocial Well-Being]** – Arriving late for work or not showing up and spending work time on non-work (e.g., surfing the web).
  - Delinquency [Psychosocial Well-Being]** – Engaging in deviant behaviors directed at the organization (e.g., theft, sabotage).
  - Adapting to Stressful Situations [Psychosocial Well-Being]** – Maintaining emotional control in stressful situations; noticing/monitoring own signs of stress from combat, work and home life and taking positive steps in managing stress reactions.
  - Overall Performance** – Global evaluation of overall performance, specific dimensions not specified.
- 

**Attitudinal Dimensions** – Attitudinal dimensions reflect servicemember attitudes, motivations, values, and so forth and are typically collected through self-report measures.

- Work Satisfaction** – An individual's satisfaction with work and career.
- Morale** – A holistic judgment of one's own morale.
- Organizational Commitment** – An individual's psychological bond with the organization, as represented by an affective attachment to the organization, internalization of its values and goals, and a behavioral desire to put forth effort to support it.
- Withdrawal Cognitions/Intentions** – Thinking about or intending to quit one's job.
- Person-Environment Fit (PE Fit)** – Congruence between the individual's abilities, needs, and expectations and characteristics of the organization, job or group.
- Other** - Attitudinal criterion that does not fit into one of the above categories.

## Section 3. Measurement

### Measurement Method

Is this measure in a **verbal** or **written** format (e.g., interview, essay)?

- Yes  
 No

If yes, please select the applicable option(s) below:

- Interview – One-on-one  
 Interview – Panel  
 Short answer  
 Essay
- 

Does this measure involve **ratings** from at least one individual (e.g., supervisor)?

- Yes  
 No

If yes, please indicate the applicable option(s) below:

- Supervisor  
 Peer  
 360
- 

Is this measure a type of **simulation**, such as situational judgment tests (SJT), virtual role plays (VRP), and assessment centers?

- Yes  
 No

If yes, please indicate the applicable option(s) below:

- Low fidelity – distant proxy for the content being assessed and attempts to describe elements from the natural working environment such as reading or hearing details about the job (e.g., situational judgment test)  
 High fidelity – very similar to the content being assessed and may directly include realistic elements of the working environment such as being immersed in the job (e.g., virtual role play)  
 In-person (e.g., assessment center) – the individual being evaluated must come to a testing site to be assessed

---

Does this measure involve ratings about **attitudes** or **fit**? **Fit** refers to the congruence or compatibility between the employee and environment (e.g., organization, workgroup, job) in terms of values, attitudes, tasks, and other characteristics.

- Yes  
 No

If yes, please indicate the applicable option(s) below:

- Self-report  
 Other, such as peer-report (please specify)

---

Is this measure based on **administrative** information (e.g., training, attrition)?

- Yes  
 No

If yes, please indicate the applicable option(s) below:

- Training (e.g., grades, recycles)  
 Attrition  
 Discipline  
 Performance (e.g., marksmanship, physical fitness)  
 Exceptional performance (e.g., medals)  
 Other (please specify)

---

Does this measure assess **job performance** or **job knowledge** and is not covered by one of the above methods?

- Yes  
 No

If yes, please indicate the applicable option(s) below:

- Work sample  
 Knowledge test  
 Fitness test  
 Self-report (objective) - *quantitative value such as course grades/GPA*  
 Self-report (subjective) - *evaluative judgements such as performance ratings*

## Scaling

Please select the type(s) of scaling of the criterion measure.

- Dichotomous (e.g., go/no go; yes/no)
- Likert
- Multiple choice
- Forced choice (i.e., select only one of two statements)
- Matching
- Drag and Drop (e.g., categorizing statements)
- Serviceperson comparison (e.g., rank order, paired comparison, forced distribution)
- Frequency count (e.g., number of "yes" responses for training achievements, training failures, and disciplinary incidents)
- Other, such as open-ended response for physical fitness (please specify)

Please select the type(s) of scale anchor of the criterion measure.

- Agreement** – the degree to which an individual agrees with a statement (e.g., strongly disagree to strongly agree)
- Behavioral Anchor Rating Scale (BARS)** – scale points are descriptions of behaviors
- Frequency** – how often the behavior is performed (e.g., ranging from once a day to once a year)
- Relative** – compared with other servicepersons (i.e., among the weakest/best, ranging from bottom 20% to top 20%)
- Most/least effective** – the degree of effectiveness of what is being rated, without the aid of behavioral anchors
- Not applicable** – Criterion does not include scales with anchors (e.g., total scores, administrative records)
- Other (please specify)**



Please indicate the number of scale points for this criterion measure. If there are multiple scales, please indicate the number of points for different scales. If the criterion measure does not include scales (e.g., total scores, administrative records), please write "not applicable."

## Section 4. Data Accessibility

### Data Accessibility

#### Data Location

Please describe where the data is stored, if applicable. For example, "The data was collected experimentally via an online survey tool;" "Data is stored in [administrative database name];" "Data was collected by academic researchers."

#### Ownership

Please describe the entity that owns the data, if applicable. For example, "The Defense Manpower Data Center."

#### Discoverable

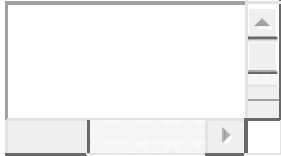
For example, "Data are accessible internal to a DOD service through administrative records or is maintained in an archival location. Criteria that were collected by researchers outside of DOD or were administered in studies where data may have been lost should be marked as not discoverable."

## Section 5. References

### References

#### Central Reference(s)

Please enter the full citation (preferably APA style) of the primary reference(s) that described this criterion measure, if applicable. These can include references that describe the development of the measure or evaluate its quality. Include the author name(s), title of text, technical report number (if applicable), page numbers (if applicable), etc.



#### Ancillary Reference(s)

Please enter the full citation (preferably APA style) of any secondary reference(s) (e.g., other references that used this criterion measure that may be relevant) that describe this criterion measure, if applicable.



#### Previous Versions

Please describe any previous versions of the criterion measure, if applicable. For example, "Original version was developed in 2005 as part of Select21 project; at that time it was called the Army Life survey (ALS; Van Iddekinge, Putka & Sager, 2005) and was later refined as part of the Army Class project (Moriarty, Campbell, Heffner & Knapp, 2009)."



## Section 6. Operation

### Target Population

#### Service

- U.S. Army
- U.S. Navy
- U.S. Air Force
- U.S. Marines
- U.S. Coast Guard
- Cross-service
- Other (please specify):

#### Classification

Please select the classification(s) of the population being targeted for this criterion measure.

- Trainee
- Junior enlisted
- Non-Commissioned Officer (NCO)
- Warrant Officer
- Officer
- Cross-classification
- Other (please specify):

#### Service Component

Please select the service component(s) of the population being targeted for this criterion measure.

- Active duty
- Reserve
- National guard
- Cross-component
- Other (please specify):

### Test Delivery

#### Mode of Administration

Please select how the criterion measure is (was) administered.

- Computer-based, proctored

- Computer-based, unproctored
- Paper-and-pencil
- In-person, physical
- In-person, interview
- Administrative records
- Unknown / Unspecified

## Number of Forms

Please enter the number of forms available for the criterion measured, if applicable.

## Linear vs. Adaptive

Please select whether the criterion measure consists of items assembled in a linear or adaptive fashion. **Linear** testing refers to traditional administration such that all examinees receive the same test questions in the same order. **Adaptive** testing refers to adapting the test to the examinee's level such that each subsequent item is selected based on the examinee's scored responses to the previous items.

- Linear
- Adaptive
- Not Applicable

## Item Presentation

Please select the manner in which items in the criterion measure are presented to examinees. **One-at-a-time** refers to questions being presented individually. **Grouped** refers to questions being presented in sets of multiple questions.

- One-at-a-time
- Grouped
- Unknown / Unspecified
- Not Applicable

## Assessment Length

Please indicate the number of items in the assessment and / or time limit, as appropriate for this assessment.

## Test Window Length

Please enter the time period in which the criterion measure can be taken per each test administration. For example, if a test is administered three times a year and each administration is open to examinees for one month, then enter, "one month." If the assessment is administered operationally, please put "Not applicable."

## Frequency of Testing

Please enter the number of times this criterion is administered, if applicable. For example, if a test is administered once per year, then enter, "once per year."

## Throughput

Please enter an approximate number of examinees that are assessed with this criterion measure in a given timeframe, if applicable. For example, 3,000-4,000 per year. Leave blank if unknown or not applicable.

## Scoring

### Mode of Scoring

Please select how the criterion measure is scored or graded. **Automated** and **hand-scored** are scoring measures in the format of an exam or test. **Trained observers** provide ratings of examinees on dimensions of behaviors or other characteristics.

- Automated
- Hand-scored
- Trained observers
- Not specified

### Current Use

- Operational – non-decisional** – Assessment is operational, but not used to make decisions that impact individual employees (e.g., engagement surveys, end-of-course evaluations)
- Operational – decisional / high stakes** – Assessment is used to make operational decisions (e.g., final evaluation where a passing score is needed to complete basic or individual technical training, performance appraisal scores)
- Experimental** (e.g., to support validation efforts)

### Data Cleaning

Please describe any rules for excluding data collected in the criterion measure, if applicable. For example, "Data were flagged as unusable if the respondent (a) omitted more than 10% of the

assessment items, (b) took fewer than five minutes to complete the entire assessment, or (c) chose an implausible response to the careless responding item.”

## Procedures

Please describe the procedure or steps taken to calculate examinees' scores. Example 1: "Attitudinal scales are scored by taking an average of component items." Example 2: "Count of 'yes' responses for training achievements, training failures, and disciplinary incidents."

## Score Reporting – Standardized Scores

Please select whether or not scores on the criterion measure are converted to a standardized scale (e.g., stanine, t-score) for reporting purposes.

- Yes
- No

## Score Reporting – Norm-referenced scores

Please select whether or not scores on the criterion measure are norm-referenced so that examinees can be compared to a population.

- Yes
- No

## Section 7. Psychometric Properties

### Reliability

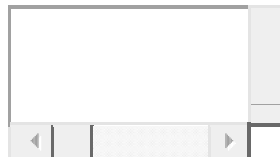
#### Coefficient Alpha

Please enter the Cronbach's alpha coefficient (internal consistency) for the criterion measure. If there is a range of coefficients, describe that information in your response. For example, ".75 to .86".



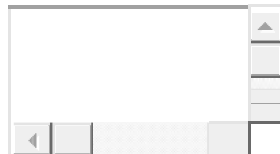
#### Interrater / Inter-observer Reliability

Please enter the interrater (or inter-observer) agreement, if applicable. For example, "ICC[C,k] = .70." Include technical details (e.g., adjusted to multiple raters or single rater) as appropriate.



#### Test-Retest Reliability

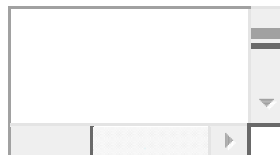
Please enter the test-retest reliability.



### Descriptive Statistics

Please enter information about the descriptive statistics for this criterion measure. If there are multiple scales, please list the appropriate statistics for each scale or identify the where each can be found (e.g., "Means for each scale can be found in Table 10 on Page 130 in Lee et al., 2010"). If the information is not available or not applicable, leave blank.

N



Mean



Median



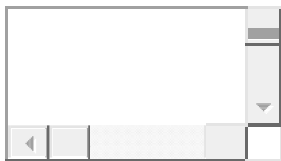
Standard Deviation (or other measure of spread if SD is not available)



Min



Max



## Descriptive Statistics

### Correlation

Please describe the source of any tables that display convergent / discriminant or other correlational validity evidence for the criterion measure. For example, "Tables B3, B5, B6 in Knapp & Wolters (2017)."



An empty rectangular text box with a light gray border. It features a vertical scrollbar on the right side and a horizontal scrollbar at the bottom, both with standard arrow and track icons.

### Demographic Subgroup Differences – Magnitude

Please describe the size of subgroup differences (small, medium, large) for the criterion measure, if applicable. Cohen's  $d = 0.2$  is **small**,  $d = 0.5$  is **medium**, and  $0.8$  is **large**.

An empty rectangular text box with a light gray border. It features a vertical scrollbar on the right side and a horizontal scrollbar at the bottom, both with standard arrow and track icons.

### Demographic Subgroup Differences – Notes

Please describe any additional information about subgroup differences for the criterion measure, if applicable. For example, include details about the particular subgroups that were compared, such as race (e.g., white-black), gender, and MOS categories (e.g., Combat Arms, Combat Support)

An empty rectangular text box with a light gray border. It features a vertical scrollbar on the right side and a horizontal scrollbar at the bottom, both with standard arrow and track icons.

## APPENDIX F - Criterion Instruments Included in Mapping Exercise

---

### Performance Instruments (K=13)

---

Army Combat Readiness Test (new, replaces APFT)  
Army Criterion Situational Judgment Test (CSJT)  
Army NCO Semi-Structured Interview  
Army NCO Situational Judgment Test  
Army NCO Situational Judgment Test-X (SJT-X)  
Army Physical Fitness Test (APFT)  
Army-Wide Job Knowledge Test (AW JKT)  
USAF Fitness Assessment  
USAF Promotion Fitness Exam  
USAF Specialty Knowledge Test  
USMC Combat Fitness Test  
USMC Physical Fitness Test  
USN Physical Readiness Test

---

### Attitudinal Surveys (K=20)

---

Army Experience and Activities Record (ExAct)  
Army Future Army Life Survey (FALS)  
Army Initial Military Training Army Life Questionnaire (IMT ALQ)  
Army In-Unit Army Life Questionnaire (IU ALQ)  
Army Life Survey (ALS)  
Army Personnel File Form (PFF)  
Army Sample Survey of Military Personnel  
Army Simulated Promotion Point Worksheet (SimPPW)  
DoD 2000 Military Exit Survey  
DoD Status of Forces Survey, August 2008 version  
USAF Air Force Life Questionnaire (draft)  
USAF Military Exit Survey (AKA New Directions Survey)  
USAF Military Retention Survey (AKA Career Decisions Survey)  
USCG Organizational Assessment Survey  
USMC Enlisted Accession Survey  
USMC Enlisted Retention Survey  
USMC Exit Survey  
USN Exit from Training Survey  
USN Navy-wide Personnel Survey (NPS)  
USN RTC Graduate Survey and A-School Graduate Survey

---

### Performance Rating Scales (K=13)

---

Army Class Combat Deployment Performance Rating Scales (CDPRS)  
Army Computer Adaptive Rating Scales (CARS)  
Army Initial Military Training Performance Rating Scales (IMT PRS)  
Army In-Unit Performance Rating Scales (IU PRS)  
Army NCO Expected Future Performance Rating Scales  
Army NCO Observed Performance Rating Scales  
Army Ranger Performance Rating Scales

Army TOPS IRB Enclosure D-1-c (PRS - Peer)  
Army TOPS IRB Enclosure D-1-d (PRS- Self Report)  
Army-Wide Current Observed Performance Rating Scales (AW COPRS)  
Army-Wide Future Expected Performance Rating Scales (AW FX)  
USAF Air Force-Wide Performance Rating Scales (draft)  
USMC Proficiency and Conduct Marks

---

**Administrative Data (K=28)**

---

Army Advanced Individual Training (AIT) Grade  
Army Attrition (Overall)  
Army Attrition by Separation Program Designator (SPD)  
Army Restarts / Failures (Overall) in Initial Military Training (IMT)  
USAF Cumulative 36-Month Attrition  
USAF Decorations  
USAF End-of-Class (EOC) Test Scores Basic Military Training (BMT)  
USAF Final Grade Technical Military Training (TMT)  
USAF Graduation Status BMT  
USAF Graduation Status TMT  
USAF Honor Graduate BMT  
USAF Honor Graduate TMT  
USAF Months of Mission-Ready Service  
USAF Time in Grade (TIG)  
USAF Time in Service (TIS)  
USAF Washbacks BMT  
USAF Washbacks TMT  
USAF Weighted Airman Promotion System (WAPS): Selection to E5-E7  
USMC Attrition Gates  
USMC Computed Tier Score  
USMC Time to promote to E4  
USMC Unsuitability attrition  
USN Advancement Rate  
USN Final Grade A-School  
USN Graduation Status from A-Schools  
USN Reenlistment Rate  
USN Setback A-School  
USN Training Assessment Framework

---

## APPENDIX G - Sub-Dimension Results for Measurement Mapping

### Table G.1. Number of Instruments Mapped to Sub-Dimensions

Sub-Dimension	Definition	Performance Rating Scales (K=13)	Performance Measures (K=13)	Attitudinal Measures (K=20)	Administra- tive Data (K=28)
<i>A. Technical Performance Constructs</i>					
A.1. Task Performance					
Job-Specific Proficiency	Being able to perform job-specific tasks at the appropriate skill level.	9	2	0	0
General Proficiency	Being able to perform service-wide tasks at the appropriate skill level (e.g., navigation in the Army and Marine Corps).	6	3	0	0
A.2. Communication					
Oral Communication	Conveying information in a clear, understandable, organized manner when speaking.	9	1	0	0
Written Communication	Conveying information in a clear, understandable, organized manner when writing.	6	0	0	0
Nonverbal Communication	Using alternative, culturally appropriate methods to interpret and convey meaning when common language is not shared.	0	1	0	0
A.3. Decision Making, Problem Solving, and Innovation					
Decision Making, Problem Solving, and Innovation	Making sound, timely decisions, even under pressure; analyzing situations and innovating solutions to problems; resolving conflicts; adapting plans and decisions as situations change.	10	3	0	0
A.4. Safety and Security					
Conscientiousness					
Safety and Security Consciousness in Everyday Work	Following safety and security guidelines and instructions, noticing and alerting others to potential hazards in day-to-day work.	4	0	0	0
Safety and Security Consciousness during Mission Operations	Being alert to enemy and environmental threats and taking actions that do not place self or others at <u>unwarranted risk.</u>	2	0	0	0

**Table G.1. Number of Instruments Mapped to Sub-Dimensions (Continued)**

<b>Sub-Dimension</b>	<b>Definition</b>	<b>Performance Rating Scales (K=13)</b>	<b>Performance Measures (K=13)</b>	<b>Attitudinal Measures (K=20)</b>	<b>Administrative Data (K=28)</b>
<b><i>B. Organizational Citizenship and Peer Leadership</i></b>					
<b>B.1.Planning and Structuring Work</b>					
Providing Structure	Leading peers when given a leadership role, giving clear instructions, distributing tasks, and gaining others' cooperation.	7	2	3	0
Teamwork	Working with other team members to interpret the mission, set and prioritize team goals, and monitor team performance.	5	3	3	0
Self-Management	Managing own responsibilities (e.g., work assignments, gear, equipment, personal finances, family, and personal well-being), and appearing on duty prepared for work. Setting personal work objectives.	8	2	0	0
Learning/Training Self-Management	Planning, organizing, and using study time effectively (e.g., setting aside specific times to study; completing assignments on time).	3	1	0	0
<b>B.2.Conscientious Initiative</b>					
Initiative	Taking the initiative to do all that is necessary to accomplish team or organizational objectives encountered, finding additional work to perform when own duties are completed, and volunteering for work assignments.	10	1	3	0
Persistence	Persisting with extra effort despite difficult conditions and setbacks, accomplishing goals that are more difficult and challenging than normal completing work on time despite unusually short deadlines, and performing at a level of excellence that is significantly beyond normal expectations.	10	1	0	0

**Table G.1. Number of Instruments Mapped to Sub-Dimensions (Continued)**

<b>Sub-Dimension</b>	<b>Definition</b>	<b>Performance Rating Scales (K=13)</b>	<b>Performance Measures (K=13)</b>	<b>Attitudinal Measures (K=20)</b>	<b>Administrative Data (K=28)</b>
Self-Development	Developing or adapting own knowledge and skills by taking courses on own time, volunteering for training and development opportunities offered within the organization and trying to learn new knowledge and skills on the job from others or through new job assignments.	6	2	1	0
Classroom Learning	Being actively engaged in own learning by searching for and obtaining information, taking notes in class, highlighting relevant material, practicing new skills, and participating/contributing during classes.	3	0	0	0
B.3.Support for Peers					
Helping	Helping others by offering suggestions about their work, showing them how to accomplish difficult tasks, teaching them useful knowledge or skills, directly performing some of their tasks, and providing emotional support for personal problems.	9	3	1	0
Cooperating	Cooperating with others by accepting their suggestions, following their lead, being open-minded and adapting to others' ways, and informing others of events or requirements that are likely to affect them.	8	2	0	0
Courtesy & Respect	Showing consideration, courtesy, and tact in relations with others.	8	3	2	0
Motivating	Motivating others by applauding their achievements and successes, cheering them on in times of adversity, showing confidence in their ability to succeed, helping them overcome setbacks, and modelling leadership behavior.	4	2	1	0
Serving as a Model	Modeling core values by acting unselfishly, enduring hardships without complaint, treating	6	0	0	0

**Table G.1. Number of Instruments Mapped to Sub-Dimensions (Continued)**

<b>Sub-Dimension</b>	<b>Definition</b>	<b>Performance Rating Scales (K=13)</b>	<b>Performance Measures (K=13)</b>	<b>Attitudinal Measures (K=20)</b>	<b>Administrative Data (K=28)</b>
	others well, behaving ethically, and showing confidence and enthusiasm.				
Accepting Differences	Showing interest in and respect for people of other backgrounds or cultures by regularly engaging with them in a manner considerate of their norms.	3	1	0	0
<b>B.4.Organizational Support</b>					
Organizational Support	Complying with organizational rules and procedures, encouraging others to comply with organizational rules and procedures, and suggesting procedural, administrative, or organizational improvements.	8	1	0	0
Selfless Service	Committing to the greater good of the team or group putting organizational welfare ahead of individual goals.	5	0	0	0
Military Presence	Presenting a positive and professional image of self and the military even when off duty, maintaining proper military appearance.	7	1	2	0
Integrity/Moral Courage	Demonstrating honesty and integrity in job-related matters, even when own self-interests might be jeopardized, taking steps to protect the security of military equipment/supplies, and voluntarily reporting thefts, misconduct, and any other violations of military order and discipline.	6	1	1	0
<b>C. Psychosocial Well-Being</b>					
<b>C.1.Adapting to Stressful Situations</b>					
Adapting to Stressful Situations	Maintaining emotional control in stressful situations; noticing/monitoring own signs of stress from combat, work and home life and taking positive steps in managing stress reactions.	8	2	1	0

**Table G.1. Number of Instruments Mapped to Sub-Dimensions (Continued)**

<b>Sub-Dimension</b>	<b>Definition</b>	<b>Performance Rating Scales (K=13)</b>	<b>Performance Measures (K=13)</b>	<b>Attitudinal Measures (K=20)</b>	<b>Administrative Data (K=28)</b>
<b>C.2. Counterproductive Work Behavior</b>					
Loafing and Tardiness	Arriving late for work or not showing up; spending work time on personal activities (e.g., surfing the web).	3	0	1	0
Delinquency	Engaging in deviant behaviors directed at the organization (e.g., theft, sabotage).	1	0	1	0
Bullying, Harassing, or Hurting Others	Engaging in deviant behavior directed at others (e.g., physical attacks, verbal abuse, harassment).	1	0	1	0
Abusing Substances and Other Self-Destructive Behavior	Engaging in self-destructive behavior (e.g., alcohol or drug abuse).	0	0	0	0
<b>D. Physical Performance</b>					
<b>D.1. Fitness</b>					
Fitness	Meeting military standards for weight, physical fitness, and strength, maintaining own health.	9	7	1	0
<b>D.2. Endurance</b>					
Endurance	Sustaining physical performance over long periods of time despite lack of sleep and difficult conditions. Adapting to environmental challenges (e.g., weather, terrain).	4	1	0	0
<b>Overall Performance</b>	A holistic judgment of overall performance.	8	0	0	0



## APPENDIX H - Measurement Method Full Ratings

**Table H.1. Mean Ratings For 22 Measurement Methods on 9 Evaluation Dimensions**

Measurement Method	Evaluation Dimensions							
	Susceptibility to contamination	Reliability	Discriminability	Validation Uses	Ease / Cost of Measure Development	Ease and Quality of Administration	Ease and Quality of Data Management	Ease / Cost of Maintenance
Performance Rating Scales: Supervisor	2.25	2.75	2.63	4.50	4.50	3.63	4.38	4.38
Performance Rating Scales: Peer	1.75	2.50	2.25	3.88	4.50	3.63	4.25	4.38
Performance Rating Scales: Self	1.88	3.00	1.63	2.88	4.50	4.25	4.63	4.38
Performance Rating Scales: Multisource	3.25	2.75	2.88	4.25	4.38	3.25	3.63	4.25
Work Samples/Hands-on	3.13	3.63	3.88	4.00	2.13	1.13	4.00	2.38
Simulations/Assessment Centers	3.25	3.00	3.75	3.38	1.00	2.75	3.88	1.88
Oral Interview	2.63	3.63	3.63	3.25	3.75	2.13	3.88	3.63
Situational Judgment Tests	4.13	3.25	4.00	3.13	2.13	4.00	4.38	2.50
Job Knowledge Tests	4.13	5.00	4.38	4.25	3.00	4.00	4.63	3.13
Self-report survey: Attitudes	3.50	4.63	3.14	3.75	4.50	4.50	4.88	4.63
Self-report survey: Objective performance/Personnel Data	3.25	3.29	2.75	3.75	4.63	4.50	4.88	4.25
Attrition	3.00	1.94	2.85	4.38	4.63	4.63	2.75	4.13
Reenlistment	3.00	1.94	2.71	3.75	4.63	4.63	3.25	4.25
Operational supervisor ratings	1.63	2.13	1.75	2.25	4.88	4.63	3.50	4.50
Promotion rate	2.38	3.34	2.85	2.75	4.75	4.88	3.38	4.50
Training school grades or pass/fail	3.38	3.13	3.00	3.63	4.75	4.63	3.25	4.00
Production indices	1.94	3.27	3.13	2.25	4.25	4.25	2.88	4.13
Personnel file records: performance scores (e.g., physical fitness, rifle qualification)	3.75	3.63	3.38	3.25	4.75	4.38	3.25	4.25
Personnel file records: Merit-based awards and recognition	2.88	2.56	2.63	2.75	4.88	4.50	3.38	4.13
Personnel file records: Negative outcomes (Articles 15, Counseling)	3.25	2.50	2.41	2.88	4.88	4.38	3.25	4.25
Personnel file records: Promotion points/score	2.75	2.88	3.14	3.13	4.75	4.63	3.50	4.25

*Notes.* N = 8. The scale ranges from 1-5, where 1 means the measurement method performs poorly on that evaluation dimension and 5 indicates that it performs well on that dimension. Means of 3.5 and higher appear in grey shading. Means of 2.5 or lower appear in a box.