



AFRL-RH-WP-TR-2020-0036

DEVELOPMENT OF JOINT-SERVICE CRITERION INSTRUMENTS FOR ENLISTED JOBS

**Laura A. Ford
Martin C. Yu
Christopher R. Graves
Christopher R. Huber
Teresa L. Russell
Michael P. Wilmot
Brenda Ellis**

**Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 700
Alexandria, VA 22314-1578**

May 2020

Interim Report

DISTRIBUTION A. Approved for public release; distribution unlimited.

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

NOTICE AND SIGNATURE PAGE

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

This report was cleared for public release by the 88th 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-2020-0036 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signed//

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

//signed//

TIMOTHY S. WEBB, Ph.D., DR-III
Chief, Collaborative Interfaces and
Teaming Branch
Warfighter Interface Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

//signed//

LOUISE A. CARTER, Ph.D., DR-IV
Chief, Warfighter Interface Division
Airman Systems Directorate
711th 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.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) 06-05-2020		2. REPORT TYPE Interim		3. DATES COVERED (From - To) 10 January 2018 – 1 May 2020	
4. TITLE AND SUBTITLE Development of Joint-Service Criterion Instruments for Enlisted Jobs				5a. CONTRACT NUMBER FA8650-14-D-6500, Task Order 0007	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER 62202F	
6. AUTHOR(S) Laura A. Ford, Martin C. Yu, Christopher R. Graves, Christopher R. Huber, Teresa L. Russell, Michael P. Wilmot, and Brenda Ellis				5d. PROJECT NUMBER 5329	
				5e. TASK NUMBER 09	
				5f. WORK UNIT NUMBER HOSA	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Human Resources Research Organization (HumRRO) 66 Canal Center Plaza, Suite 700 Alexandria, VA 22314-1578				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Material Command Air Force Research Laboratory 711 th Human Performance Wing Airman Systems Directorate Warfighter Interface Division Collaborative Interfaces and Teaming Branch Wright-Patterson AFB, OH 45433				10. SPONSORING/MONITORING AGENCY ACRONYM(S) 711 HPW/RHCC	
				11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S) AFRL-RH-WP-TR-2020-0036	
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution A. Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES Report contains color. 88ABW-2020-1934; Cleared 31 May 2020					
14. ABSTRACT The Department of Defense (DoD) uses the Armed Services Vocational Aptitude Battery (ASVAB) and other predictor measures to select over 100,000 new military recruits every year and place them in military occupations. 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 across the five Armed Services (Army, Air Force, Marine Corps, Navy, and Coast Guard). The purpose of this project was to develop a unified set of criterion measures that can be used by all Services. In the first phase of the project, Allen et al., (2020) developed taxonomies of job performance, attitudes, and organizational outcomes for first term enlisted personnel; constructed a database of criterion measures used by the Services; and linked criterion measures to the performance domain constructs. This report describes the second phase of the project: the development of unified criterion measures, including performance appraisal rating forms, self-assessment tools, a situational judgment test, and a methodology for measuring attrition consistently across Services.					
15. SUBJECT TERMS Criterion instruments, enlisted, joint-service, cross-service, attrition, survey, performance rating scale, situational judgment test					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Thomas R. Carretta
Unclassified	Unclassified	Unclassified	SAR	83	19b. TELEPHONE NUMBER (Include Area Code)

TABLE OF CONTENTS

LIST OF FIGURES	iii
LIST OF TABBLES	iii
ACKNOWLEDGEMENTS	iv
1.0 EXECUTIVE SUMMARY	1
1.1 Research Requirement.....	1
1.2 Procedures	1
1.3 Findings	2
1.3.1 Procedures to Align Outcome Variables across Services	2
1.3.2 Performance Rating Scales to Address Job Performance Constructs	2
1.3.3 Self-Report Assessments	2
1.3.4 SJT	2
1.3.5 DoD-Wide Criterion Measurement Research Plan	3
1.4 Utilization and Dissemination of Findings.....	3
2.0 INTRODUCTION	4
2.1 Background	4
2.2 Research Objectives	4
3.0 DEVELOPMENT OF PROCEDURES FOR ALIGNING ADMINISTRATIVE VARIABLES.....	6
3.1 Attrition	6
3.1.1 Collecting and Screening Raw Attrition Variables.	6
3.1.2 Understanding Inter-Service Separation Codes	7
3.1.3 Categorizing ISCs into Four Types of Attrition	8
3.1.4 Constructing Attrition Types Variables	11
3.1.5. Constructing Overall and Cumulative Attrition Variables.....	11
3.1.6 Coding Attrition “Gates”.....	12
3.2 Training	13
3.2.1 Raw Training Variables.....	13
3.2.2 Constructing Training Outcome Variables.....	14
3.3 Physical Fitness and Disciplinary Incidents	16
3.3.1 Physical Fitness Raw Variables.	16
3.3.2 Constructed Physical Fitness Outcome Variables.....	17
3.3.3 Disciplinary Incidents Raw Variables.....	18
3.3.4 Constructed Disciplinary Incident Outcome Variables.....	19
3.4 Summary	19
4.0 DEVELOPMENT OF PERFORMANCE RATINGS SCALES.....	21
4.1 Identify Performance Areas.....	21

4.2	Select Scaling Techniques	22
4.3	Develop Scale Content	23
4.4	Develop Instructions.....	23
4.5	Develop Familiarity Rating Scales.....	24
4.6	Produce EOT and IU Forms	24
4.7	Rater Training.....	25
4.8	Tailoring for Service-Specific Administration.....	26
5.0	DEVELOPMENT OF SELF-REPORT ASSESSMENTS.....	27
5.1	Assessment Development Process	27
5.2	EOT Assessment	28
5.3	IU Assessment	29
5.4	Exit Survey	31
6.0	DEVELOPMENT OF THE SITUATIONAL JUDGMENT TEST	36
6.1	Test Specifications	36
6.2	Identifying Potential Items	37
6.3	Preliminary Item Editing.....	39
6.4	Military SME Review	39
6.5	Post-Workshop Review	40
7.0	SUMMARY AND RECOMMENDATIONS	43
7.1	Summary	43
7.2	Recommendations for Evaluating New Measures: Criterion Measurement Research Plan	44
7.3	Conclusion.....	45
8.0	REFERENCES	46
9.0	LIST OF ACRONYMS, ABBREVIATIONS AND SYMBOLS	46
	APPENDIX A. Final Procedures for Aligning Administrative Variables.....	51
	APPENDIX B. Training for Performance Rating Scales	56
	APPENDIX C. DoD-Wide Criterion Measurement Research Plan	63

LIST OF FIGURES

Figure 1.	Familiarity Scales.....	24
Figure 2.	Example SJT Item.....	36
Figure 3.	SJT Effectiveness Rating Scale.	42

LIST OF TABLES

Table 1.	Raw Attrition Variables by Service	7
Table 2.	Types of Attrition Across ISCs.....	9
Table 3.	Raw Training Variables by Service	13
Table 4.	Crosswalk of Training Status Codes.....	15
Table 5.	Raw Physical Fitness Variables by Service	17
Table 6.	Raw Disciplinary Incidents Variables by Service	18
Table 7.	Job Performance Taxonomy	22
Table 8.	Differences in Example Behaviors Between Forms by Performance Area	25
Table 9.	Content of the Initial EOT Assessment	28
Table 10.	Content of the IU Assessment.....	30
Table 11.	Content of the Exit Survey.....	33
Table 12.	Existing SJT Sources for Item Development.....	37
Table 13.	Performance Dimensions Targeted by SJT.....	38
Table 14.	Initial Coverage of Target Performance Dimensions by Potential SJT Items	38
Table 15.	Characteristics of the SJT Review Workshop Participants.....	40
Table 16.	Summary of Item Generalizability Ratings	41
Table 17.	Summary of Performance, Organizational Outcome, and Attitude Constructs Targeted by Cross-Service Criterion Measures	43
Table 18.	Proposed Projects and Purposes for Evaluating Criterion Measures.....	45

ACKNOWLEDGEMENTS

We want to thank Dr. Thomas Carretta (Air Force Research Laboratory[AFRL]) and Dr. Cristina Kirkendall (United States Army Research Institute for the Behavioral and Social Sciences) for their oversight and guidance throughout this project. We also 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, Captain Alex Ryan, Ms. Mary Ellen (Ellie) Stone, Dr. Sofiya Velgach, and Mr. Johnny Weissmuller.

1.0 EXECUTIVE SUMMARY

1.1 Research Requirement

The 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 across the five Armed Services (U.S. Army, Air Force, Marine Corps, Navy, and Coast Guard). 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 measure first term enlisted performance and other important outcomes. A preliminary phase of this project laid the foundation for the development of new criterion measures (Allen et al., 2020). This report describes the second phase of the project: the development of unified criterion measures.

1.2 Procedures

The objectives of the current effort were to:

- establish a common metric for similar, but not the same, criterion instruments used across Services (e.g., attrition, training),
- adapt existing criterion instruments for cross-service use (e.g., self-assessments, exit surveys),
- develop a Situational Judgment Test (SJT) for measuring important criterion in the joint-service environment,
- prepare a research plan for evaluating the psychometric characteristics of the new set of criterion measures.

Existing criterion measures were used as a starting point for the development of several new measures. Military subject matter experts (SME) in each Service reviewed draft criterion instruments and provided feedback and suggestions to tailor test items, ratings scales, and/or surveys to their specific Service branch.

A necessary component of accomplishing the above objectives was garnering input from representatives of each Service (Air Force, Army, Coast Guard, Marine Corps, and Navy). Thus, 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 (a) feedback through discussions and emails, and (b) access to additional SMEs.

1.3 Findings

1.3.1 Procedures to Align Outcome Variables across Services

After reviewing methods the Services use to measure and track outcome variables, we developed procedures to standardize (or align) attrition, training outcomes, physical fitness scores, and counts of disciplinary incidents across the Services. Based on our review of the currently available databases, the new procedures for constructing attrition and training outcome variables should be feasible for immediate implementation. Similarly, the information needed to construct physical fitness scores appears to be available in military personnel databases, despite the relatively infrequent use of physical fitness scores in prior research. Of the outcome variables we reviewed, disciplinary incidents were the most sparsely documented. The possibility of accessing the Defense Incident Based Reporting System (DIBRS) to obtain disciplinary incident data should be investigated before any associated recommendations can be implemented.

1.3.2 Performance Rating Scales (PRS) to Address Job Performance Constructs

Cross-Service job PRS were developed to address the first term enlisted job performance constructs identified in the first phase of this project (Allen et al., 2020). Two forms of PRS were developed: (a) one to be administered at the end-of-training (EOT) and (b) the other to be administered during the 1st term of Service. Both forms are suitable for use by supervisors or peers. The PRS were designed to measure all the constructs of the job performance taxonomy including: (a) technical proficiency, (b) organizational citizenship and peer leadership, (c) psychosocial well-being, and (d) physical performance.

1.3.3 Self-Report Assessments

Self-report assessments were developed to extend coverage of the criterion domain. Three versions of self-report measures were developed:

- EOT assessment – to measure dimensions of attitudinal constructs, performance constructs and organizational outcome constructs (physical and technical performance, reprimands, and awards);
- In-unit (IU) assessment – to measure dimensions of attitudinal, performance, and outcome constructs as with the EOT assessment;
- Exit survey – to provide a more fine-tuned measure of attrition, focusing on detailed reasons for attrition, withdrawal cognitions, counterproductive work behavior (CWB), and work satisfaction.

All three assessment types were designed for use across Services, and each include the minimal number of items required to measure the targeted constructs. In application, researchers will be able to supplement the core set of items with Service-specific items in support of their unique research needs.

1.3.4 SJT

A cross-service item pool for an SJT was developed to address the constructs in the job performance taxonomy that lacked coverage from other measures. The constructs targeted were (a) technical performance (specifically decision making, problem solving, and innovation) and (b) organizational citizenship and peer leadership (to include planning and structuring work, conscientious initiative, support for peers, and organizational support).

The SJT was designed to assess test-takers' judgments about various problems they might encounter as E-3s and E-4s. Each SJT item consists of two parts: (a) a description of a hypothetical scenario and (b) several possible responses to that scenario. Examinees are asked to evaluate the effectiveness of different responses by comparing them to one another (e.g., indicating the best and worst responses) or by separately rating the effectiveness of each response.

This initial phase involved constructing an item pool by identifying existing SJT items and modifying them for cross-service use. As discussed in the DoD-Wide Criterion Measurement Research Plan, additional research will be needed to finalize and validate the SJT.

1.3.5 DoD-Wide Criterion Measurement Research Plan

Based on input from military personnel researchers, we developed a plan for evaluating the psychometric characteristics of the new set of cross-service measures. The research plan describes a strategy for conducting the data collections that will be required for thorough evaluation of the criterion measures. The plan describes the challenges and possible solutions for:

- Obtaining data collection resources,
- Using technology for data collection,
- Enhancing data quality,
- Obtaining Institutional Review Board (IRB) approval,
- Evaluating the accuracy of attrition data,
- Maximizing usefulness of data collected,
- Streamlining training and in-unit data collection,
- Planning EOT and IU data collection.

In addition, the research plan describes three proposed projects for evaluating specific types of criterion measures:

- Attrition – to evaluate procedures for constructing attrition variables, the integrity of attrition data;
- EOT – to evaluate performance appraisal ratings and a self-assessment survey administered at the end of technical training;
- IU – to evaluate performance appraisal ratings, a situational judgment test, and a self-assessment survey administered to E-3 and E-4 Service members having 18 to 36 months of experience.

1.4 Utilization and Dissemination of Findings

The new unified set of criterion procedures and measures can be used in joint-service projects to evaluate the criterion-related validity of the ASVAB and other predictor measures. The research plan condenses ideas of many military personnel researchers and provides a useful starting point for planning future joint-service validation efforts.

2.0 INTRODUCTION

The purpose of this effort was to identify criterion constructs and measures for joint-service enlisted occupations. This report documents the second of two phases for the *Development of Criterion Measures for Evaluating Accession and Classification Testing* project. The first phase is documented in a report entitled “Identification of Criterion Constructs and Measures for Joint-Service Enlisted Jobs” (Allen et al., 2020). The second phase of the project focused on the development of a set of criterion measures targeted to assess the job performance, attitudes, and organizational outcomes of first term enlisted personnel across the Services. These criteria are intended to provide military researchers the necessary tools to conduct validity studies for predictor measures used across the Services or within an individual Service.

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 (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 across the Services.¹ 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 was 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 subject matter experts 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,
- organize criterion instruments into a taxonomic structure of outcomes of interest (e.g., job performance), and
- develop recommendations for a unified set of test evaluation criteria that can be used by all Services.

The objectives of the second phase of the project were, based on phase one recommendations, to:

¹ The U.S. Army, Air Force, Marine Corps, and Navy are DoD branches. The U.S. Coast Guard branch is under the U.S. Department of Homeland Security.

- establish a common metric for similar, but not the same, criterion instruments used across Services (e.g., attrition, training),
- adapt existing criterion instruments for cross-service use (e.g., self-assessments, exit survey),
- develop a SJT for measuring some important criterion in the joint-service environment,
- prepare a research plan for evaluating the psychometric characteristics of the new set of criterion measures.

A necessary component of accomplishing the above objectives was garnering input from representatives of each Service (Air Force, Army, Coast Guard, Marine Corps, and Navy). Thus, 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 (a) feedback through discussions and emails, and (b) access to SMEs for reviews of the criterion measures as they were being developed.

3.0 DEVELOPMENT OF PROCEDURES FOR ALIGNING ADMINISTRATIVE VARIABLES

In this section, we provide procedures for aligning attrition, training, physical fitness, and disciplinary incident outcome variables across the Services.

3.1 Attrition

Most Services use attrition as an outcome for validating new predictor measures. However, differences exist across Services in terms of *when* attrition is computed during the first term of Service (e.g., total days since accession versus at monthly intervals), as well as *how* it is computed. Further, the Services vary in the degree to which they document the reason for separation, and research indicates that separation reasons can be highly informative. To resolve these important issues, we provide an integrative framework for aligning separation codes across the Services and organizing them into four types of attrition.

We examined data coding rubrics from the Services to establish practical joint-service procedures for:

- Collecting and screening raw attrition variables,
- Using Inter-service Separation Codes to code four types of attrition,
- Constructing overall and cumulative attrition variables,
- Coding attrition *gates* that enlisted personnel pass through (i.e., basic training, technical training, and full term of Service).

3.1.1 Collecting and Screening Raw Attrition Variables.

To construct various measures of attrition, it is necessary to obtain, at minimum, three raw variables.² Because accession and separation dates are the foundation for constructing attrition measures, accurate data are paramount. For obvious reasons, separation date should follow accession date; in cases where it does not, it may be necessary to check for data entry errors and/or re-enlistment. Concerning the latter, in relevant cases, it also may be necessary to obtain a fourth raw variable to construct attrition measures:

1. Accession date (i.e., date of entry; reflecting actual ship date and not just contract date)
2. Separation date (i.e., date of exit)
3. Separation reason
4. Re-enlistment date (i.e., date of “re-up”)

Table 1 shows a list of raw attrition variables by Service and, when applicable, their source database(s).

² Delayed Entry Program (DEP) attrition is outside the scope of this document, but it is also an important variable to consider.

Table 1. Raw Attrition Variables by Service

Service	Source Database(s)	Raw Variables
U.S. Army	Total Army Personnel Database (TAPDB)	– Accession Date – Separation Date – Separation Program Designator (SPD) – Inter-service Separation Codes (ISC)
U.S. Air Force	Defense Manpower Data Center (DMDC)	– Accession Date – Loss Date – ISC
U.S. Navy	N/A	– Military Personnel Manual (MILPERSMAN) Regulation Codes
U.S. Marine Corps	Marine Corps Total Force System (MCTFS)	– Accession Date – Separation Date – Planned SPD – Actual SPD

Note. N/A indicates that information on these variables was not readily available for this project but does not imply that these variables do not exist. Blue text indicates a hyperlink.

3.1.2 Understanding ISC

We obtained Service-specific separation codes for four Services: Army, Air Force, Marine Corps, and Navy. The Army uses the “SPD”³ and the Marine Corps also uses a detailed variant of this coding system. The Naval MILPERSMAN documents regulation codes, and the Air Force uses ISCs as well as SPDs. ISC are used by the DoD’s DMDC as an organizing framework to track separation reasons across all DoD Services. Therefore, we recommend the use of ISC as a single standardized system to consolidate the tracking of attrition reasons across Services.

As an example of this, Appendix A presents a summary table of 46 attrition related ISC and their associated descriptions. For each code, this table provides a crosswalk to corresponding SPD codes for the Army, Air Force, and Marine Corps, as well as MILPERSMAN regulation codes for the Navy. To create this crosswalk, one coder systematically reviewed the descriptions for all SPD and MILPERSMAN codes and sorted them into a corresponding ISC code. Afterward, two other coders⁴ independently reviewed these sorted codes, and made recommendations for re-sorting as necessary. Results presented in Appendix A reflect all coders’ consensual crosswalk of codes according to the common ISC framework. As the table shows, both SPD and MILPERSMAN codes can be efficiently integrated and recoded into this coding system.

Separation codes, whether they are Service-specific or cross-service, are the official administrative reason for separation. Although they tend to be a good reflection of a Service member’s reason for leaving, they are not perfect (Ramsberger & Babin, 2005). It is possible that a Service member may have more than one reason for separating, or that the separation code

³ The first letter of the three-digit SPD code indicates *type* of separation (B, D, F = Resignation; G, J = Involuntary discharge; K = Voluntary discharge; L = Involuntary release from active duty or transfer; M = Voluntary release from active duty or transfer; P = Dropped from the rolls of the military; Q = Civil office or Failure to accept regular appointment; R, S, V, W = Retirement; Y = Release from military control); the second and third letters indicate *reason* for separation.

⁴ All coders had Ph.D. degrees in Industrial/Organizational Psychology and several years of military research experience.

does not reflect the Service member's true reason for separating. That said, they tend to be consistent with separation information obtained via alternate sources, such as exit interviews (Sipes et al., 2000). Although improving the accuracy and reliability of separation codes (e.g., allowing for multiple separation codes) would likely involve DoD-wide policy changes, minor modifications to the ISC coding system during variable construction would at least enhance organization and output. Whenever base rates of ISC codes permit, codes should be disaggregated into meaningful subcategories for more accurate classification of separation reasons. However, this is still limited by the codes that have been operationally used in the past. For example, ISC 87 (Entry level performance and conduct) has been anecdotally reported to be used as a catch-all, and the lack of more specific codes under this category makes it difficult to further subcategorize. As shown in Appendix A, attrition subcategories are identified for some codes. For example, the subcategories for ISC 8 (Early release) are (a) reduction in force, (b) force shaping, and (c) general/miscellaneous reasons.

3.1.3 Categorizing ISCs into Four Types of Attrition

Building on this integrative effort, Table 2 offers an organizing framework for classifying ISCs into four categories:

- Substandard Performance,
- Substandard Fitness,
- Counterproductive Behaviors, and
- Unfavorable Circumstances.

To form these four categories, or types, we drew on job performance models showing that the criterion space is multidimensional (Campbell & Wiernik, 2015) and that different criterion dimensions have different antecedents. For example, cognitive ability is a stronger predictor of task performance, but temperament (i.e., personality) is a relatively stronger predictor of non-task performance (e.g., contextual performance, counterproductive work behaviors; cf. Gonzalez-Mulé, Mount, & Oh, 2014). Moreover, physical fitness has long been recognized to be a critical criterion in military jobs (Campbell et al., 1990) and is likewise strongly predicted by non-cognitive attributes (i.e., temperament, vocational interests). Finally, although human attributes are key contributors to military effectiveness, a sizable proportion of variance in the criterion space is influenced by external factors mostly outside of individual and organizational control. Accordingly, the four types of attrition here build on advances in modeling the criterion space and are consistent with dimensions in the hierarchical trainee and first-term performance taxonomy developed for this effort (Allen et al., 2020).

Table 2 summarizes 39 ISC attrition related codes within four attrition types. The first type, *Substandard Performance*, concerns separation codes that reflect training and/or job performance behaviors that fail to meet minimum qualifications for retention, indicate unsatisfactory performance, or demonstrate entry-level performance and conduct (ISC 85–87).

Table 2. Types of Attrition Across ISCs

ISC	ISC Description	Substandard Performance	Substandard Fitness	Counter-productive Behaviors	Unfavorable Circumstances
2	Early release: Insufficient retainability				X
10	Condition existing prior to service				X
11	Disability: Severance pay				X
12	Permanent disability retirement				X
13	Temporary disability retirement				X
14	Disability: No condition existing prior to service				X
16	Unqualified for active duty: Other				X
17	Failure to meet weight or body fat standards		X		
22	Dependency or hardship				X
60	Character or behavior disorder			X	
64	Alcoholism			X	
65	Discreditable incidents: Civilian or military			X	
67	Drugs			X	
71	Civil court conviction			X	
72	Security			X	
73	Court-Martial			X	
74	Fraudulent Entry			X	
75	Absent without Leave (AWOL) or desertion			X	
77	Sexual perversion			X	
78	Good of the service (in lieu of court-martial)			X	

Note. ISC = Inter-service Separation Code. Only ISC codes considered to be attrition are listed. Codes not subsumed by these types, including 1 (Expiration of term of service), 3–8 (Early release), 40 (Officer commissioning program), 52 (Retirement), and 100 (Immediate re-enlistment), should all be considered non-attrit.

Table 2. Types of Attrition Across ISCs (continued)

ISC	ISC Description	Substandard Performance	Substandard Fitness	Counter-productive Behaviors	Unfavorable Circumstances
80	Misconduct: Reason unknown			X	
82	Unsuitability: Reason unknown			X	
83	Pattern of minor disciplinary infractions			X	
84	Commission of a serious offense			X	
85	Failure to meet minimum qualifications for retention	X			
86	Unsatisfactory performance	X			
87	Entry level performance and conduct	X			
90	Secretarial authority				X
91	Erroneous enlistment or induction				X
92	Sole surviving family member				X
94	Pregnancy				X
95	Minority (underage)				X
96	Conscientious objector				X
97	Parenthood				X
98	Breach of contract				X
99	Other				X
101	Dropped from strength: Desertion			X	
102	Dropped from strength: Imprisonment			X	
105	Dropped from strength: Other			X	

Note. ISC = Inter-service Separation Code. Only ISC codes considered to be attrition are listed. Codes not subsumed by these types, including 1 (Expiration of term of service), 3–8 (Early release), 40 (Officer commissioning program), 52 (Retirement), and 100 (Immediate re-enlistment), should all be considered non-attrit.

It is most related to the “Technical Performance” category in the aforementioned performance taxonomy. Relatedly, the second attrition type, *Substandard Fitness*, involves separation for failure to meet weight or body fat standards, or otherwise maintain physical fitness in such a way that meets basic military requirements. It is most related to the “Physical Performance” category in the performance taxonomy. The third attrition type, *Counterproductive Behaviors*, refers to individual behaviors that reflect social or moral impairment that detract from military- relevant goals, and outcomes of misbehavior, that result in separation. It is conceptually related to the performance taxonomy subcategory of the same name. Example ISCs are 60 (Character or behavior disorder), 64 (Alcoholism), 73 (Court Martial), 83 (Pattern of minor disciplinary infractions), and 101 (Dropped from strength: Desertion). The final type, *Unfavorable Circumstances*, involves separation codes that, from the perspective of the military, reflect undesirable external conditions, circumstances, or consequences, as well as unforeseen changes to them (e.g., ISC 12 [Permanent disability], 22 [Hardship and dependency], and 98 [Breach of

contract]). Due to its focus on external factors, it is not related to the performance taxonomy. Finally, ISC codes that are not subsumed by these types, including expiration of term of Service (1), early release (3–8), transfer (40), retirement (42), and immediate re-enlistment (100), should all be considered non-attrib.

3.1.4 Constructing Attrition Types Variables

To construct the four attrition types found in Table 2, a separate 0 or 1 indicator variable can be created based on whether the separation reason falls into the respective category. For example, a variable for attrition due to counterproductive behaviors should be coded as 1 for any ISC indicated under the *Counter-productive Behaviors* column in Table 2 and coded as 0 for any ISC not indicated under the same column. As described in the prior paragraph, the seven ISCs that do not correspond to one of the four types should be considered non-attribution and coded as missing. Procedures are identical whether separation reason is Service-specific or extracted from DMDC.

3.1.5. Constructing Overall and Cumulative Attrition Variables

Temporal attrition variables are relatively straightforward and facilitate construction of overall and cumulative attrition measures. This information is readily available across all Services. First, to construct overall time-to-attribution variable (i.e., time in service), compute the difference between accession and separation dates using the total number of days. If the Service member has not separated (e.g., no separation code or separation date), code the value as missing. If the time is negative, it should be flagged for manual review as the separation date precedes an accession date. Dates should be checked for data entry errors (e.g., invalid accession date in the future) or possible re-enlistment. Any cases that cannot be rectified should be coded as missing.

Secondly, to construct overall cumulative attrition on monthly intervals (i.e., attrition in monthly increments up to 48 months⁵), for each monthly interval, code 0 or 1 for whether the service member had attrited by that point based on their time-to-attribution variable. For example, if a service member separated after 12 months of service, attrition for the one through 11-month intervals would be coded 0, but attrition at 12 months would be coded 1. However, if the service member had separated, but his or her separation code is one of the codes considered non-attrib as described previously (e.g., expiration of term of service), recode any value of 1 to 0.

In longitudinal data, censored data occur when values on a temporal variable are unknown. For example, a service member with 12 months of service would not have any attrition information available for intervals after 12 months. All censored data should be treated as missing. Using the prior example, all by-month variables after 12 months would be coded as missing. As a further example, if a service member has 12 months in-service, but has not separated, s/he would receive codes of 0 up to 12 months and then missing thereafter.

Finally, an overall attrition variable should be constructed as the most recent non-missing cumulative attrition value. This would be either 1 for attrit or 0 for non-attrib. If no non-missing value is available, then overall attrition should be coded as missing.

⁵ While 48 months should capture the vast majority of active duty service obligations (ADSO), some ADSOs can be up to 60 or 72 months long and increments could be increased accordingly whenever the data or research question requires doing so.

These attrition variables can be further specified into additional variables for each attrition type. For example, if there is interest in examining attrition due to substandard fitness only, then additional versions of the above variables can be constructed such that 1 indicates whether a service member had attrited for substandard fitness reasons and 0 for non-attrited or attrited for reasons not related to substandard fitness.

3.1.6 Coding Attrition “Gates”.

There are three important *gates* that most service members pass through in their service term: (a) basic training, (b) technical training, and (c) full term of service. Research supports the importance of examining these gates in conjunction with attrition as much of first-term enlisted attrition occurs early in service, especially during basic and technical/specialty training (Gibson et al., 2014; Knapik et al., 2004). Additionally, different attrition types predominate at different gates. For example, Hooper et al., (2008) found that in the Air Force, most attrition in the first few months was due to moral character or medical reasons. On the other hand, performance-based attrition started off low, but rapidly increased and reached a peak at six months, consistent with the timeline of basic and technical training. Similarly, an Army study (Project First Term) found that attrition for performance and medical/physical reasons predominated during the first six months of service and diminished afterwards, at which point moral character attrition became more prevalent (Strickland, 2005).

Although the by-month attrition variables described previously can be used as proxies for these gates (e.g., basic training should be completed by three months of service), we do not recommend using them for this purpose when the goal is to consolidate comparisons across services. Considerable variability exists across services, as well as between specialties within each service. Basic training can typically range from eight weeks for the Navy to 13 weeks for the Marine Corps, and technical/specialty training can range from one month to one year or more depending on specialty. Further, a term of service typically ranges between two to six years. Therefore, while these gates can be approximated based on time in service, doing so will likely incur a good deal of measurement error.

Instead, we recommend capturing these gates directly using additional data sources. The general procedure for constructing gate variables is for 0 to indicate that a service member has passed the gate, and 1 to indicate that a service member had separated before passing the gate. For the “completion of term of service” gate, a variable should be constructed that is coded 0 if the service member’s separation code is indicated as a completion or expiration of term of service, 1 if the separation code is indicated as a different reason, and missing if the service member has not yet separated.

Construction of training failure variables are described in the next section. For attrition due to training failures, there may be a time lag between when the service member fails his or her training course(s) and when the service member separates. Thus, for the basic and technical/specialty training gates, variables should be constructed for each that are coded as follows:

- 0 = Graduated from training
- 1 = Failed training and separated for a performance-related reason
- 2 = Failed training and separated for a reason not related to performance

- NA = Neither graduated from training (regardless of failure or not completed) nor separated.

3.2 Training

Success in basic and technical training is a common administrative variable examined across Services. Beyond separation reasons, which can track service members who separated directly due to training failures, many Services also include training success, such as (a) success without recycles⁶ and (b) graduation with honors, in their validation research. Examples include research on waivers on training recycles and graduation with honors in the U.S. Air Force (Putka & Allen, 2008), and evaluation of training restarts and failures in the U.S. Army's Tier One Performance Screen research (Knapp & Kirkendall, 2018; Knapp & Wolters, 2017).

Using training outcome variables allows for the technical proficiency aspect of the job performance taxonomy to be evaluated. Further, the interpretation of technical proficiency results can be improved by aligning training performance variables across Services.

3.2.1 Raw Training Variables

Table 3 lists the databases, by service, where training outcome data currently reside.

Table 3. Raw Training Variables by Service

Service	Source Database(s)	Raw Variables
U.S. Army	Army Training Requirements and Resource System (ATRRS)	- Training course status/reason codes (e.g., graduate, fail, recycle)
U.S. Air Force	Basic Training Management System (BTMS) Technical Training Management System (TTMS)	- Training course status/reason codes - Field school grade (qualifying, awarding, and grade) - BTMS Course 1 test score and grade - Air Force Specialty Code (AFSC) Technical Training course grade
U.S. Navy	Corporate Enterprise Training Activity Resource Systems (CeTARS)	- Person Event (PEVT) Code - Field school grade
U.S. Marine Corps	MCTFS	- Academic test grade

It is difficult to draw meaningful comparisons across Services because of the diversity in training content and assessment. Accordingly, to align training outcomes, the clearest points of comparison are overall performance in basic and initial technical (or specialty) training. The reason is that these training events occur during similar time periods for each service member's career trajectory, and they serve similar training goals across Services. For basic and technical training, the raw outcome variables that are most commonly available for making cross-service comparisons are (a) service members' training grades and (b) training status codes (e.g., graduation, recycle/wash-back, and fail/washout).

⁶ Recycles occurs when a trainee repeats a part of training. It may also be referred to as setbacks or washbacks.

3.2.2 Constructing Training Outcome Variables

Whenever cross-service comparisons are of interest, construction of training outcome variables should focus primarily on using training status codes. Although training grades would provide more fine-grained information about service members' training performance, they should be reserved for intra-service analyses because of differences in training programs and performance documentation across Services. However, training grades do have utility in constructing training status outcome variables.

Constructing training status outcome variables depends on the availability of training status and/or training grade variables. Three hypothetical cases merit consideration.

Case 1: Only training statuses are available, not training grades. This first case primarily applies to the Army⁷. In this first case, Service-specific status codes should be recoded into three broad, dichotomous indicators of whether a service member (a) graduated, (b) recycled/washed-back, or (c) failed/washed out. Table 4 presents information about which codes map to these three statuses across Services. Codes should be made separately for basic and technical training. For each status, separate status-specific variables should be coded as 1 = indicated and 0 = not indicated. Given sufficient fidelity in the status codes shown in Table 4, the recycle and fail indicator variables can be further partitioned into more specific variables indicating whether a service member recycled or failed for academic and/or non-academic reasons. Similarly, depending on the reason of interest, indicators would also be coded as 1 = indicated and 0 = not indicated. For example, in constructing an indicator variable for “recycled for academic reasons,” service members who did not recycle or who recycled for non-academic reasons would both be coded as 0.

⁷ Training grades are supposedly recoded in the Digital Training Management System (DTMS) but according to the Army Research Institute for the Behavioral and Social Sciences (ARI), access to that data is not currently available for research purposes.

Table 4. Crosswalk of Training Status Codes

Training Status	Army ATRRS	Air Force BTMS/TTMS	Navy CeTARS PEVT
Graduate	Single indicator	98, 99	288
Recycle or setback, academic	Single indicator; does not differentiate between academic and non- academic	Single indicator; does not differentiate between academic and non-academic	48, 50, 51, 52, 53, 56, 58, 231, 324, 855, 964
Recycle or setback, non- academic			73, 230, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 305, 308, 965
Failure, academic	A, 8	92, 93, 94, 95, 96	78, 81, 84, 103, 106, 135, 138, 141, 142, 309, 371
Failure, non-academic	3, 6, 4, M, Y, V, N, H, U, ?, =, +, \$, -, #, %, @, <	97	145, 148, 149, 155, 158, 167, 168, 169, 171, 172, 173, 175, 176, 177, 178, 181, 182, 183, 184, 188, 189, 190, 191, 192, 196, 197, 198, 200, 202, 203, 204, 207, 211, 212, 213, 214, 222, 311, 320, 366, 367, 385, 941, 969, 970, 971, 972, 973, 986, 987, 988, 994

Note. ATRRS = Army Training Requirements and Resource System. BTMS = Basic Training Management System. TTMS = Technical Training Management System. CeTARS PEVT = Corporate Enterprise Training Activity Resource Systems Person Event Code. The Marine Corps does not appear in this table as training status codes were not identified.

Additional variables should be constructed that count the number of times each service member had recycled prior to graduation. This includes service members who are ultimately indicated as having graduated; for those who are ultimately indicated as having failed out, it also includes a count of the number of times each service member had recycled prior to failing out. The preceding variables should be constructed as follows:

- Number of times recycled prior to graduation
 - If the graduation indicator variable = 0, code as missing
 - If the graduation indicator variable = 1, count the number of recycles indicated in the service member's training records.
- Number of times recycled prior to failing out
 - If the fail indicator variable = 0, code as missing
 - If the fail indicator variable = 1, count the number of recycles indicated in the service member's training records.

Case 2: Only training grades are available, not training statuses. This second case applies to the Marine Corps. Training grades have been identified in the MCTFS database, but training status codes have not. However, this limits the fidelity of training outcome variables relative to those described for Case 1. In this case, training grades should be dichotomized into whether service members passed or did not pass their basic and technical training, based on their Service-specific training criteria. If multiple training events or attempts are tracked, then it would also be possible to count the number of times a service member failed before ultimately graduating or failing out.

Case 3: Both training statuses and grades are available. This final case applies to the Army, Air Force, and Navy. Their databases contain both training statuses and training grades. In this case, the training outcome variables should be constructed using the methods described in Cases 1 and 2. The purpose of Case 3 is to provide a check on the consistency or accuracy of the constructed training outcome variables. If the graduation or fail indicators between those constructed in the first and second methods do not agree, then the data should be manually examined. Given the variety of data issues that could conceivably occur, developing a full list of them and their suggested resolutions is not realistic. However, to provide two salient examples, if separate training records or databases were merged, service member training records should be verified to ensure that the merge was accurate. A second issue concerns whether the training statuses and grades describe the same training event or course, or instead pertain to different ones.

3.3 Physical Fitness and Disciplinary Incidents

Administrative records containing indicators of physical fitness and disciplinary incidents can serve as direct measures of performance. However, they have typically not been included in most Services' past validation efforts. As a result, information on them is less accessible than attrition and training performance in the databases currently used across the Services. However, both physical fitness and disciplinary incident outcome variables provide useful information about individual service member performance. In particular, disciplinary incidents capture a highly important and difficult to measure aspect of performance—psychosocial well-being (Allen et al., 2020).

3.3.1 Physical Fitness Raw Variables.

The Services currently administer the following physical fitness tests:

- *Army Physical Fitness Test (APFT)*⁸ – The APFT provides 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.
- *U.S. Air Force Fitness Assessment (FA)* – The FA provides an assessment of the overall fitness of Air Force personnel. It assesses aerobic and muscular fitness, flexibility, and optimal body composition.
- *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).
- *USMC Physical Fitness Test (PFT)* – The PFT is a collective measure of general fitness administered Marine Corps-wide. It is designed to measure the strength and stamina of the upper body, midsection, and lower body, as well as the efficiency of the cardiovascular and respiratory systems.

⁸ As of the writing of this report, in 2020 the Army is rolling out a transition from the APFT to the Army Combat Fitness Test (ACFT) which is scored on a different metric. Construction of these outcome variables for the ACFT should be evaluated once the data becomes available for research purposes.

The raw scores and results for these assessments are found in the databases shown in Table 5.

Table 5. Raw Physical Fitness Variables by Service

Service	Source Database(s)	Raw Variables
U.S. Army	TAPDB	- APFT Result (Pass/Fail)
U.S. Air Force	Air Force Fitness Management System (AFMS II)	- Composite Fitness Score - Fitness Level Score Category - Fitness Level Score Pass/Fail
U.S. Navy	Navy Standard Integrated Personnel System (NSIPS)	- PRT Score
U.S. Marine Corps	MCTFS	- Combat Fitness Test Score - Combat Fitness Test Classification - Body Fat Composition within standard - Physical Fitness Test Score - Physical Fitness Test Classification

Note. N/A indicates that information on these variables was not readily available but does not imply that these variables do not exist.

3.3.2 Constructed Physical Fitness Outcome Variables.

Because physical fitness requirements differ across Services, scores need to be placed on a comparable metric. Two methods should be used: standardizing scores and creating pass/fail indicators.

Method 1: Standardization. The first method is standardizing physical fitness test scores within Services. If scores are extracted as part of a standalone datafile, they should be standardized for each individual in the datafile. On the other hand, if the scores are part of a continually updated longitudinal datafile, scores should be standardized across each individual in a given year, so as not to recompute scores for everyone with every update of the datafile. Standard scores should be computed as z-scores,

$$z_i = \frac{x_i - \bar{x}}{s}$$

where z_i is the z-score for an individual, x_i is the raw score for an individual, $\bar{x}$ is the mean score for the sample, and s is the score standard deviation for the sample.

To promote comparability, these standardized scores should only be computed for Servicemembers who complete the regular physical fitness test. In the case that a Servicemember is tested on an alternate physical fitness test (e.g., on profile and recovering from an injury that precludes him or her from being able to complete a particular test), his or her score on the alternate test should be excluded from the calculation.

Method 2: Pass/Fail. The second method is creating a pass/fail indicator variable on physical fitness for each service member, based on the Services' respective physical fitness requirements. Raw physical fitness variables available in Service-specific databases already contain categorical indicators of fitness level, but with varying degrees of specificity (e.g., Pass/Fail for the Army

versus Excellent/Satisfactory/Unsatisfactory for the Air Force). Thus, to facilitate alignment between Services, they should be consolidated into the following codes:

- P = Passed physical fitness requirements
- F = Did not pass physical fitness requirements
- W = Obtained a waiver for physical fitness requirements

Each method is straightforward to construct, but their use depends on the analysis and research question of interest. Standardization of physical fitness test scores would have more fidelity and preserve rank-ordering of individuals within each Service. On the other hand, the pass/fail variable would provide an indicator of physical fitness performance relative to each Service's specific requirements.

3.3.3 Disciplinary Incidents Raw Variables.

Although disciplinary incidents can occur for any of the Services, documentation of disciplinary incidents variables could only be obtained from Marine Corps (see Table 6) for purposes of this project. The Marine Corps records pay deductions and fines due to courts martial and non-judicial punishments, as well as punitive rank reductions, in the MCTFS.

Table 6. Raw Disciplinary Incidents Variables by Service

Service	Source Database(s)	Raw Variables
Army	N/A	N/A
Air Force	N/A	N/A
Navy	N/A	N/A
Marine Corps	MCTFS	- Court Martial Pay Deduction - Court Martial/Non-Judicial Punishment Fine - Punitive Rank Reduction

Note. N/A indicates that information on these variables was not readily available but does not imply that these variables do not exist.

Although there is a lack of documentation regarding how disciplinary incidents are tracked by each Service, there is a DoD-wide database that tracks disciplinary incidents: The DIBRS. Given that all disciplinary incidents can be tracked in DIBRS, if feasible, we recommend that all raw disciplinary incident data should be obtained from this common, standardized source. In doing so, the primary raw variables of interest are:

- Date of Offense,
- Date of Adjudication,
- Offense Identifier (see Tables 7 and 8 of DoD Manual (DoDM) 7730.47-M-V2, *Defense Incident-Based Reporting System (DIBRS): Supporting Codes*, for codes),
- Offense Result (attempted or completed),
- Involvement (principal, accessory, conspiracy, or solicit),
- Action Taken (no action, administrative, non-judicial, or judicial).

However, it is currently unclear whether DIBRS can be accessed for research purposes. We first recommend that, if possible, procedures for accessing DIBRS should be established. The following recommendations for constructing disciplinary incident outcome variables depend on the possibility of extracting raw data from DIBRS.

3.3.4 Constructed Disciplinary Incident Outcome Variables.

For each service member, disciplinary incident outcome variables can be constructed from their raw DIBRS variables. Because base rates for individual Offense Identifiers are likely to be low, the Offense Identifier variable should first be recoded into broader categories of offenses against person, property, or society (see pages 8-49 [Tables 7 and 8] of [DoDM 7730.47-M-V2](#)).

Following that, because it is possible for multiple disciplinary incident records to exist for a single service member, this information should be most efficiently aggregated by constructing separate numeric variables to count the number of times a service member has been indicated for each of the following raw DIBRS variables:

- Overall offenses, which should be subcategorized into offenses against person, property or society;
- Offense attempted;
- Offense completed;
- Involved as principal;
- Involved as accessory, conspiracy, or solicit;
- Action taken: None,
- Action taken: Administrative or non-judicial,
- Action taken: Judicial.

Although the basic recommendation here is to create event count variables for disciplinary incidents, it is possible that the base rates for these variables could be too low for analysis purposes. As they are counts, the values for these variables can range from 0 to any number greater than 0; however, the most common values are likely to be 0 or 1. If dichotomous versions of these variables would be more appropriate given the base rates or the research question of interest, count values greater than 1 could be recoded to 1. In such a case, the resulting values represented would only be either 0 or 1 in each variable.

3.4 Summary

Currently, considerable variability exists in the availability and construction of administrative outcome variables across Services, which makes it difficult to examine outcomes across the DoD. Standardization of criterion measurement would enable robust comparisons of results within and across the Services, as well as strengthen DoD conclusions about the validity and utility of their existing assessments. The recommendations presented in this report provide organizing frameworks for attrition and training status codes, as well as a unified set of variable construction procedures for attrition, training, physical fitness, and disciplinary incident outcome variables. Summaries of these recommendations are presented in Appendix B. These constructed variables provide a broad range of information about performance and non-performance outcomes for conducting validation research across the Services.

Based on our review of the currently available databases, the recommendations for constructing attrition and training outcome variables should be feasible for immediate implementation. In fact, many have been used in several past research projects. For physical fitness, the information necessary to construct outcome variables for research purposes should be available in the currently identified databases. However, these have not been used in past research to the extent that attrition and training variables have for reasons including a greater interest in attrition and training outcomes, and unreliability in physical fitness variables. Of the outcome variables reviewed, disciplinary incidents are the most sparsely documented. The possibility of accessing DIBRS to obtain disciplinary incident data should be established before any associated recommendations can be implemented.

4.0 DEVELOPMENT OF PRSs

This section describes the development of cross-service PRS to address the constructs that constitute the job performance taxonomy. Two forms of PRS were developed: (a) one to be administered at the EOT and (b) the other to be administered during the 1st term of service. Both forms are suitable for use by supervisors or peers. The PRS were designed to measure all the constructs of the job performance taxonomy (see Table 4.1), including: (a) technical proficiency, (b) organizational citizenship and peer leadership, (c) psychosocial well-being, and (d) physical performance. Development of the PRS included six steps: (a) identifying the constructs to measure; (b) selecting scaling techniques (c) developing scale content, that is, descriptions of the performance to be rated; (d) developing instructions for scale users; (e) developing familiarity rating scales to support the assessment of rating accuracy; and (f) producing EOT and IU forms.

In addition, a storyboard for rater frame-of-reference training was developed to be administered along with the PRS. Frame-of-reference training has been shown to increase accuracy in performance ratings (Borman et al., 2017).

The PRS forms and training were tailored for cross-service use by SMEs. The Services provided one or more SMEs for PRS review: Army (1), Navy (2), Air Force (1), Marine Corps (1), and Coast Guard (2). Tailoring ensured all technical content, language, and terminology were appropriate for each Service while maintaining the maximum degree of consistency in construct measurement across Service-specific forms.

4.1 Identify Performance Areas

The PRS developed were designed to measure the constructs in job performance taxonomy, shown in Table 4.1. Performance areas for the PRS were specified at the dimension level. One scale was developed for each category, except for the CWB, where we specified two CWB: (a) one for unproductive behaviors and (b) one for harmful behaviors. Dividing CWBs into two scales was performed to reduce the disparity in the types of behaviors for which ratings would be provided, which should, in turn, increase rating accuracy. The final PRS form included 12 scales.

Table 7. Job Performance Taxonomy

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

4.2 Select Scaling Techniques

In accordance with recommendations formed in the first phase of this project (Allen et al., 2020), PRS forms use relative and frequency scaling techniques. The use of relative scales versus absolute scales reduces the reading burden for users. Relative scales also mitigate the impact of range restriction caused by commonly observed non-normal distributions of performance in organizations (LeBreton et al., 2003; O’Boyle & Aguinias, 2012). While relative scaling is applied across constructs, frequency scaling is used to measure psychosocial well-being, a dimension of CWB, as this is the most typical technique employed in that measurement domain (Dalal, 2005). Specific questions posed to raters for the two types of scales were, “How well

does the service member perform in this area compared to other service members you have observed?” and “In the past 3 months, how often did you observe the service member engage in the behaviors?”

4.3 Develop Scale Content

Scale content derived directly from the definitions provided for each construct, dimension, and sub-dimension in the job performance taxonomy. Definitions were stated in terms of detailed performance behaviors, which facilitated the selection and refinement of example behaviors. To minimize the reading requirement for raters, the number of example behaviors was kept to the bare minimum needed to cover all aspects of the dimensions. This was accomplished by combining related behavioral statements, when possible. We also eliminated non-essential verbiage to reduce statement length. Scales in the final form included between one and six example behaviors. For instance, the following behaviors were developed for the IU PRS Showing Support for the Organization performance area:

- presents a positive and professional image of self and the military, even when off duty;
- commits to the greater good of the team or unit, putting organizational welfare ahead of individual goals;
- complies with, and encourages others to comply with, organizational rules and procedures;
- demonstrates honesty and integrity, even when own self-interests might be jeopardized;
- takes steps to protect the security of military equipment/supplies; and
- voluntarily reports violations of military order and discipline.

4.4 Develop Instructions

To maximize accurate ratings, instructions were developed to familiarize users with the content of the rating scales (Borman et al., 2017) and the types of scaling employed. Instructions begin with a brief description of the rating task, which is followed immediately by a statement explaining that ratings will be used for research (i.e., selection or classification tool validation) only and will not be revealed to the Service, unit, or persons rated. Instructions remind users that the ratings will be used for research purposes only and will not impact any service member’s career. This is intended to reduce the tendency of raters to portray all ratees similarly, whether that is average or high, as is often the case in rating exercises employed for operational decision-making purposes (Knapp & Campbell, 1993).

After rater training, which is described later in this section, instructions list the performance areas to be rated and provide a detailed description of the types of scales used in the form. Final guidance reminds the reader to:

1. Prepare: Read the performance area name and the example behaviors carefully so that you understand the aspects of performance that you are being asked to rate.
2. Consider Typical Performance: Consider the service member’s typical performance in each area over the time you have observed him/her.
3. Consider Strengths and Weaknesses: Your ratings should accurately reflect the service member’s strengths and weaknesses. It is unlikely that any service member performs at the same level in all areas.

4. Consider Differences Among Service members: Your ratings should accurately reflect the differences among service members. In each area, use the entire scale as required to differentiate between high, average, and low performing service members.
5. Be Honest and Accurate: Mark the option that most accurately reflects the service member's performance in each area. If you have not observed the service member's performance in that area, choose the "Not observed/Cannot rate" option.

4.5 Develop Familiarity Rating Scales

Performance rating scales are often accompanied by ratee familiarity measures to shed light on the accuracy of ratings provided. Two familiarity scales that are currently used in conjunction with the Army's In-Unit PRS are used in the present form. These scales are shown in Figure 4.1.

1. Rate your familiarity with each Service member you will be rating.

	Not enough to judge the Service member's performance	Enough to judge some aspects of the Service member's performance	Enough to judge many aspects of the Service member's performance	Enough to judge most aspects of the Service member's performance
Service member 1	☐	☐	☐	☐
Service member 2	☐	☐	☐	☐

2. During an average work week, what percentage of your time is spent working in direct contact with each Service member?

	Less than 25% of my working hours	25-50% of my working hours	51-75% of my working hours	More than 75% of my working hours
Service member 1	☐	☐	☐	☐
Service member 2	☐	☐	☐	☐

Figure 1. Familiarity Scales.

4.6 Produce EOT and IU Forms

Developing EOT and IU forms of the PRS involved selecting and modifying example behavior statements. Most differences between forms resulted from specifying that performance occurred either during training or under in-unit conditions. Additional behaviors were added to some IU scales to more comprehensively cover performance requirements. Separate forms were required for 8 of the 12 scales, including:

- Task Performance
- Maintaining Safety and Security
- Planning and Structuring Work
- Showing Persistence and Taking Initiative
- Supporting Peers
- Supporting the Organization
- Engaging in Unproductive Work Behavior

- Engaging in Behavior Harmful to the Organization or Others.

Table 8 shows the types of differences between example behaviors in the EOT and IU forms. In the table, differences are shown in bold text.

Table 8. Differences in Example Behaviors Between Forms by Performance Area

Performance Area	EOT Form	IU Form
Task Performance	– Performs job-specific and service-wide tasks proficiently.	– Performs job-specific and service-wide tasks proficiently. – Performs job-specific and service-wide tasks at the appropriate skill level.
Planning and Structuring Work	– Leads other trainees when given a leadership role...	– Leads peers when given a leadership role...
Showing Persistence and Taking Initiative	– Takes responsibility for own learning ▪ Takes notes and participates in class ▪ Practices tasks – Searches for relevant information.	– Develops or adapts own knowledge and skills: ▪ Takes courses on own time. ▪ Volunteers for training and development opportunities. – Tries to learn on the job from others or through new job assignments.
Engaging in Unproductive Work Behavior	– Arrive late for or miss formations, class, or training activities without permission.	– Arrive late for or miss formations or work activities without permission.

Note. Bolded text highlights the differences between the two forms.

4.7 Rater Training

Rater training, or frame-of-reference training, which provides a common conceptualization of performance prior to a rating task, has been shown to increase rating accuracy (Borman et al., 2017). A variant of this training was developed to: (a) familiarize raters with the scales and their content, (b) instruct raters on use of the two types of rating scales, and (c) explain ways to avoid common rating errors. The training is currently provided in a storyboard that serves as a design document for developing the training for use in an on-line administration environment.

The training begins with three stated learning objectives that include developing familiarity with:

- the areas of performance to be rated,
- the scales to be used to make ratings, and
- ways to avoid common rating errors.

As seen in the training storyboards in Appendix B, users can review the 12 performance areas to be rated and are directed to the category names to see the associated performance areas. To develop increased familiarity with each performance area, the training guides users through an exercise in which they are asked to match sets of example behaviors to the correct performance area⁹.

⁹ Training slides were developed for each service. The example provided was developed to train Air Force raters.

The training also addresses four common rating errors: stereotyping, first impression, general impressions, and leniency/severity. The user is required to view a brief description of each error type and a statement of the optimal rating approach.

To increase its effectiveness, future training modules should include audio capability to allow for narration. In common military testing circumstances, training with audio capability will require the availability of headphones to avoid distracting other personnel.

4.8 Tailoring for Service-Specific Administration

Tailoring the generic rating forms and training resulted in only a few variations among Services, which should not affect construct measurement in a manner that would prohibit validation studies from drawing conclusions from data obtained across Services. In addition to editing Service-specific forms by changing the names of Services and service members, the PRS forms and associated training were tailored for each Service as follows:

- The last behavior on the Showing Support for Peers scale was tailored to include Service-specific lists of core values.
- The Marine Corps and Navy versions of the Adapting to Stressful Situations scale required the addition of a bullet on recognizing and addressing signs of stress in peers. This verbiage was added to reflect those Services' use of the Operational Stress Control and Readiness (OSCAR) model for stress management.
- Example behaviors on all Services' versions of the Physical Fitness scale were reworded to indicate physical fitness standards are Service-specific and not DoD-wide.
- An example behavior in the Navy's version of the Non-productive Behaviors scale included the term "muster" instead of "formations."
- For all applicable scales, the term "technical training" was changed to "A-School" in the Coast Guard form.
- An example behavior in the Non-productive Behaviors scale was modified to include the term "muster" instead of "formations" in the Navy form.

5.0 DEVELOPMENT OF SELF-REPORT ASSESSMENTS

This section describes the effort to extend coverage of the criterion domain by developing three cross-service self-report measures:

- EOT assessment – to measure dimensions of attitudinal constructs, performance constructs and organizational outcome constructs (physical and technical performance, reprimands, and awards);
- IU assessment – to measure dimensions of attitudinal, performance, and outcome constructs as with the EOT assessment;
- Exit survey – to provide a more fine-tuned measure of attrition, focusing on detailed reasons for attrition, withdrawal cognitions, psychosocial well-being, and work satisfaction.

All three assessment types were designed for use across Services, and each include the minimal number of items required to measure the targeted constructs. In application, researchers will be able to supplement the core set of items with Service-specific items in support of their unique research needs.

5.1 Assessment Development Process

The process employed to develop assessments included two steps: (a) selecting items for each construct from existing instruments and (b) adapting or tailoring the terminology and situations in items to be relevant for each Service. Existing criterion measures and their linkage to constructs of interest were identified early in the project and this information was used to select relevant sources and items for each construct. We documented item stems and response options by construct. This provided a systematic means of ensuring assessment content would cover the range of requirements across Services. We then developed a single set of items for each construct and genericized those items to be applicable across Services.

Tailoring was performed to ensure all terms, language, and technical content were appropriate for each Service, while maintaining a high level of consistency in construct measurement. For example, Services have different titles for enlisted officer ranks and for events such as physical fitness tests. Services also have different technical performance qualifications (i.e., rifle marksmanship is not required for all enlisted personnel in the Air Force). Most of the by-service tailoring of the assessments was performed by one or two SMEs per Service identified by the CMAP. The Services provided SMEs as follows: Army (12), Navy (2), Air Force (1), Marine Corps (1), and Coast Guard (2). These SMEs were tasked with reviewing and providing written feedback on the assessments. The only true focus group was conducted with Army SMEs, who reviewed the Army's version of exit survey.

The remainder of this section is organized by assessment type and describes the development and content of each assessment. Because our objectives were to provide coverage of the targeted criterion constructs, while standardizing that coverage across Services, discussion of each assessment's development focuses on the extent to which we were able to accomplish those objectives.

5.2 EOT Assessment

An EOT assessment was developed to be administered at the end of basic training or technical training.¹⁰ The assessment measures a variety of attitudinal, performance, organizational outcome constructs. Attitudinal constructs to be measured include work satisfaction, morale, organizational commitment, withdrawal intentions, and person-environment (PE) fit. Performance constructs include psychosocial well-being dimensions of CWB and adapting to stress. Finally, organizational outcome constructs included those oriented on physical and technical performance, reprimands (disciplinary actions), and awards.

An existing survey, the Army's Initial Military Training Army Life Questionnaire (IMT ALQ; Knapp & Kirkendall, 2020), which has been used with considerable success in previous validation studies, addressed a large proportion of the constructs to be measured. As such, we used the structure of that questionnaire as the basis of our EOT assessment and added items from other available instruments to fill gaps. When available, item statistics were reviewed to identify items to drop, thereby reducing assessment length.

An initial draft of the EOT assessment included 70 items across seven constructs. For each construct, Table 9 identifies construct dimensions, the number of items per dimension, and the sources from which items were obtained and adapted for cross-service use.

Table 9. Content of the Initial EOT Assessment

Performance Category	Dimension	Item Count	Source(s) ^a
Withdrawal Cognitions	– Continuance Intentions	2	1
	– Attrition Cognitions	2	3, 5
	– Reasons for Joining	1	6
	– Post-Active Duty Plans	1	6
	– Reasons for Staying/Leaving	14	Exit Survey
Morale	– Morale	1	2
Organizational Commitment	– Affective Commitment	3	2
Person-Environment Fit	– Person-Organization Fit	3	1
	– Person-Job Fit	3	1
	– Adjustment	3	1
Work Satisfaction	– Work/Life Balance	2	1
	– Job Satisfaction	3	4
Organizational Outcomes	– Awards	1	1
	– Reprimands	7	1
	– Technical Performance	3	1
	– Physical Performance	3	1
Psychosocial Well-being	– Counterproductive Work Behaviors	9	1, 7
	– Adapting to Stress	9	1, 2
Total Items		70	

^a 1 = End-of-Training Army Life Questionnaire (ALQ); 2 = Navy Recruit Training Center (RTC) Grad Survey; 3 = Army Life Survey; 4 = In-Unit ALQ; 5 = Air Force Life Questionnaire; 6 = DoD 2000 Military Exit Survey; 7 = Counterproductive Work Behavior Literature.

¹⁰ The terms *basic training* and *technical training* are used by the Army, Navy, and Air Force. The Marine Corps refers to all initial training as *formal training*, which consists of Boot Camp, School of Infantry, and Military Occupational School (the latter two representing technical training). The Coast Guard uses the terms basic training and “A” School.

Aside from global changes, such as using Service names and Service-specific terminology for members (e.g., Soldier, Airman, Marine), the tailoring process identified numerous changes needed to ensure each Service's version of the assessment was appropriate. The resulting variations, however, had only a minor impact on the number of items in some assessments, and none negatively impacted the consistency of construct measurement across Services. Examples of Service-specific tailoring requirements included:

- Services use different terms for punishments/reprimands (i.e., Article 15s and Non-Judicial Punishments), rifle marksmanship requirements (i.e., Basic Rifle Marksmanship and Marksmanship), and graduation awards.
- Rifle marksmanship is not required for certain AFSC, which resulted in the addition of the response option, "not applicable to my AFSC," in the Air Force version of the assessment.
- Services use different names for physical fitness tests: PFT and Physical Fitness Assessment (PFA). Additionally, the Marine Corps assessment includes two items on physical fitness, one for a PFT score and one for a Combat Fitness Test (CFT) score.

After tailoring, the number of items in each Service's version of the assessment were:

- Army = 69
- Navy = 70
- Air Force = 71
- Marine Corps = 71
- Coast Guard = 70

Differences in item totals resulted from variations in the items required to measure organizational outcome dimensions, such as physical fitness.

5.3 IU Assessment

We developed the IU assessment to be administered to service members between the end of their first year in-unit and the end of their first term. Though the constructs measured were the same as those measured by the EOT assessment, item content differed due to the shift in tenure. For example, the IU assessment included items on medals received and promotions and did not include items on adjustment and technical performance. Additionally, numerous item stems and response options were edited for the different conditions.

The primary item source for the IU assessment was the Army's IU ALQ (Knapp & Kirkendall, 2020). Gaps in construct coverage were addressed by adding items from various other instruments. Item statistics were reviewed to identify items that could be dropped.

As shown in Table 10, the final untailored version of the IU assessment includes 69 items across seven constructs and identifies the constructs measured, the dimensions of those constructs, the number of items per dimension, and the sources from which items were obtained and adapted for cross-service use.

Table 10. Content of the IUAssessment

Construct	Dimension	Item Count	Source(s)^a
Withdrawal Cognitions	- Continuance Intentions	2	1
	- Attrition Cognitions	2	3, 5
	- Reasons for Joining	1	6, 4
	- Post-Active Duty Plans	1	6
	- Reasons for Staying/Leaving	14	Exit Survey
Morale	- Morale	1	2
Organizational Commitment	- Affective Commitment	3	2
Person-Environment Fit	- Person-Organization Fit	3	1
	- Person-Job Fit	4	1
Work Satisfaction	- Work/Life Balance	2	1
	- Job Satisfaction	5	1
Organizational Outcomes	- Awards/Promotions	4	1
	- Reprimands	7	1
	- Physical Performance	2	1
Psychosocial Well-being	- Counterproductive Work Behaviors	9	1, 7
	- Adapting to Stress	9	1, 2
Total Items		69	

^a 1 = In-Unit Army Life Questionnaire (ALQ); 2 = Navy Recruit Training Center (RTC) Grad Survey; 3 = Army Life Survey; 4 = Army Transition Survey; 5 = Air Force Life Questionnaire; 6 = DoD 2000 Military Exit Survey; 7 = Counterproductive Work Behavior Literature.

Assessment tailoring was performed by the same SMEs who reviewed the EOT assessment. A sample of the minor variations across Services included:

- Services have different types of promotions (i.e., accelerated promotions versus meritorious promotions and battlefield promotions versus combat promotions). Additionally, not all Services have battlefield or combat promotions. These variations resulted in minor differences in the numbers of items across assessment versions.
- Services score their physical fitness tests differently. The Coast Guard's scoring is pass/fail and the Navy PFA item uses a Likert scale, which affected item construction across Services.

After tailoring, the number of items in each Services' version of the assessment were:

- Army = 67
- Navy = 67
- Air Force = 66
- Marine Corps = 69
- Coast Guard = 66

Similar to our findings for the EOT assessment, differences in item totals resulted from variations in the items required to measure organizational outcome dimensions, such as physical fitness.

5.4 Exit Survey

Exit surveys are administered to those members who are leaving the Service. The purpose of these surveys is to support efforts to reduce attrition by identifying the reasons why service members leave the military. To meet this objective, the surveys address a broad range of topics to cover the numerous reasons for attrition. Though common themes in content exist across surveys, significant variance exists in the specific questions asked, as well as in question wording and available response options. These differences pose challenges for those who would use the data for cross-service research or operational purposes.

To address these challenges, the present effort developed a condensed exit survey appropriate for cross-service use. The new survey contains a core set of items that addresses common attrition measurement themes. The new survey fine-tunes and standardizes measurement across Services and generates information that aligns with and compliments the standardized attrition variables described earlier in this section.

Survey development began with a systematic examination of each available military exit survey to identify potential content for the new “core” survey, to include:

- 2000 Military Exit Survey (Hoover et al., 2001)
- Air Force Military Exit Survey (Air Force Survey Office, 2017)
- Marine Corps Data Collection Strategy Study: Final Report (Binggeli et al., 2016)
- Marine Corps Enlisted Retention Survey (U.S. Marine Corps, 2018)
- Coast Guard Organizational Assessment Survey (U.S. Coast Guard, 2016)
- Sample Survey of Military Personnel (Personal communication, 2018)
- Status of Forces Survey (Hunter, 2006)
- Marine Corps Milestone Exit Survey (U.S. Marine Corps, 2018)
- Army Transition Survey (Giacolone, 2000)

Existing survey content was reviewed that either directly or indirectly addressed reasons for attrition. We then analyzed the content to identify a potential list of topics and sorted relevant items into those topics. The broad topics, which would later serve as sections of the survey, included (a) background information (e.g., number of dependents, financial situation, and reasons for joining the Service), (b) military life (e.g., most recent duty station, number of deployments, promotion expectations), (c) separation conditions (e.g., eligibility for reenlistment, date of separation), and (d) explicit separation factors, which included the most direct measurement of an individual’s self-reported reasons for leaving (e.g., dissatisfaction with chain of command, desire to continue education, psychological impact of serving). For each topic, we identified various types of questions. For example, three question types that addressed reasons for leaving were “to what extent did each of the following factors influence your decision to leave the Service,” “how have the following issues impacted your desire to leave the military,” and “which was the most important factor in your decision to leave the Service.” We then formulated questions that would capture the most information in the most efficient manner possible. For each question, we developed consolidated lists of relevant response options from across existing surveys.

The overall content selection and item design processes resulted in a three-level organization of content. The first level represented sections of the survey. The second level represented the topics covered in each section. Most of these topics were addressed by single items, which represent the third level. The final untailored survey included 56 items, and its organization is shown in Table 11. In the Sources column of Table 11, we identified the first existing instrument in which we found a source item for the topic. Most items were represented in multiple instruments.

The first two sections of the survey (i.e., Background and Military Life) provide insight into an individual's decision to separate from the Service when considered with his/her responses to the more direct Separation Factor items, found at the end of the survey. The third section measures an individual's withdrawal cognitions, or thoughts and actions indicative of a desire to leave Service. It also documents specific causes of separation, which should align with administrative attrition data. The last section of the survey, Separation Factors, enables the most direct assessment of a person's own thoughts about their reasons for leaving.

Table 11. Content of the Exit Survey

Section	Topic	Item	Item Count	Source(s)^a
Background	Family	Dependents	1	1
		Marital Status	1	1
	Career Intentions	Career Intentions	1	1
		Reasons for Joining	1	1
Military Life	Years of Service	Years of Service	1	2
	Rank	Rank	1	2
	Occupation	Occupation	2	1
	Duty Station	Last Duty Station	1	1
	Housing	Housing	1	1
	Deployments/Assignments	Number of Deployments	1	1
		Permanent Changes of Station	1	1
		Times away from Station	1	1
		Time away from Station	1	1
	Work Demands	Hours per Week	1	1
		Hours per Week in Occupation	1	N/A
		Held Second Job	1	1
		Times Tested but not Selected	1	1
	Promotions	Times Passed Over	1	1
		Expectations	1	1
		Stress Level	1	3
	Adapting to Stress	Strategies	4	4, 5
		Recommend the Military	1	8
	Financial Situation	Financial Condition	1	1
		Basic Expenses Covered	1	1
Your Separation	Withdrawal Cognitions	Actively Recruited	1	1
		Thoughts About Leaving	3	7
		Preparation for Leaving	1	1
		Post-Military Plans	1	6
	Reclassification	Attempted	1	1
		Prohibited	1	1
		Desire to Remain	1	1
	Type of Separation	Honorable or Other	1	1
		Voluntary/Involuntary	1	1
	Eligibility for Reenlistment	Eligibility	1	1
		Likelihood of Staying	1	1

(continued)

Table 11. Content of the Exit Survey (continued)

Section	Topic	Item	Item Count	Sources ^a
Separation Factors	Influencers	Influencers	1	1
	Separation Factors	5 Most Important Factors Ranked	1	1, 2, 3, 6, 7, 8, 9
	Occupation-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Assignment-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Deployment-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Leadership-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Unit Readiness-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Morale-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Pay/Allowances-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Benefit-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Promotion-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Morale-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Stress-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Peer-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Personal/Family-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
	Physical/Psychological-related Factors	Most Important Factor	1	1, 2, 3, 6, 7, 8, 9
Total Items			56	

^a 1 = 2000 Military Exit Survey, 2 = Army Transition Survey, 3 = Navy Exit from Training Survey, 4 = In-Unit Army Life Questionnaire, 5 = Army Life Survey, 6 = Marine Corps Data Collection Strategy Study, 7 = Sample Survey of Military Personnel, 8 = Air Force Military Exit Survey, 9 = Coast Guard Organizational Assessment Survey.

Assessment tailoring was performed by the same SMEs who reviewed the EOT and IU assessments. Modifications resulted in variations among assessments, such as:

- The Marine Corps version of the survey includes three items for occupational specialty (i.e., primary Military Occupational Specialty [MOS] Code, secondary MOS, and most recently billeted MOS), which increased the item total in this section of the Marine Corps survey.
- An item asking where the service member lived in their last permanent duty station included different response options across versions of the survey (e.g., the Navy version's response options included "aboard ship.")
- The Coast Guard version of the survey does not include the item asking if the service member had been passed over for promotion, as this is not relevant to enlisted personnel in the Coast Guard.
- The Marine Corps version of the survey did not include joining the National Guard or Reserves as response options for the item asking what the service member planned to do following separation.

After tailoring, the number of items in each Services' version of the assessment were:

- Army = 58
- Navy = 57
- Air Force = 58
- Marine Corps = 59
- Coast Guard = 57

6.0 DEVELOPMENT OF THE SJT

In this section, we describe the first phase of an ongoing effort to develop a cross-service criterion SJT. This initial phase involved constructing an item pool by identifying existing SJT items and modifying them for cross-service use. As discussed in the DoD-Wide Criterion Measurement Research Plan (Appendix C), additional research will be needed to finalize and validate the SJT.

The SJT assesses test-takers' judgments about various problems they might encounter as E-3s and E-4s. Each SJT item consists of two parts: (a) a description of a hypothetical scenario and (b) several possible responses to that scenario. Examinees are asked to evaluate the effectiveness of different responses by comparing them to one another (e.g., indicating the best and worst responses) or by separately rating the effectiveness of each response. See Figure 2 for an example item.

- One of your teammates is performing an assigned task exactly the way it is supposed to be done but is taking too much time to get it done. Your team has several more tasks to do today, and this person was told to work faster so the entire team would not have to work late. An hour later, you notice that this person hasn't made much progress. Because of this, the whole team will have to work late tonight. How effective is each of the following actions you could take?
- Talk to the person and find out if anything is bothering them.
 - Offer to assist the person until the job is completed. Talk with them later to find out why they were so slow.
 - Offer to switch tasks with the person so I can get their work done.
 - Loudly tell the group that it is this person's fault we will have to work late. This will create peer pressure and get the person to work faster.
 - Tell the team leader that is not right that the entire team will have to work late if everyone else gets their work done.
 - Ask the team leader if I can leave when my own work is done.
 - Motivate the person and lift morale by cracking jokes about how slow they are.
 - Motivate the person by encouraging them to think about the team and not just about themselves.
 - Offer to train the person to complete the task faster.

Figure 2. Example SJT Item.

6.1 Test Specifications

We began the SJT development process by determining the desired features of the operational SJT, including test length, item format, reading level, and number of forms. The desire to minimize testing time (i.e., 15-20 minutes or less) was a key factor in multiple respects. First, it constrained us to no more than 20 items per form (assuming one minute per item). In turn, this constraint suggested a response format that would allow for adequate reliability with fewer items. As such, we chose to have test-takers rate the effectiveness of each response option, which yields higher internal consistency reliabilities than other formats (Ployhart & Ehrhart, 2003; as cited in Lievens et al., 2008). Relatedly, the effectiveness rating format maximizes information

by eliciting specific evaluations of each response option, rather than incomplete ordinal comparisons between options. Put differently, knowing which option a test-taker considers “best” is less informative than knowing how they perceive each individual option.

Since the SJT was intended for low-stakes research testing, we determined that one form of the measure would suffice. Based on the projected culling of items across various stages of item review, we set a minimum target of 42 scenarios going into the military SME review (discussed below). We also targeted at least four response options per scenario for the operational SJT and planned to generate extra options to allow for later reductions. To reduce cognitive load, we aimed to write the items at an 8th grade reading level or lower.

6.2 Identifying Potential Items

To build our item pool, we began by compiling items from existing Service-specific SJTs. Our goal was to identify items that could be modified to apply at the E-3 and E-4 levels across all Services. The initial pool included a total of 119 items from four Army SJTs (see Table 12 for all source SJTs). However, the items varied considerably in their applicability to E-3s and E-4s, as well as the apparent adaptability of their content to other Services. In addition, they were not designed to assess the same constructs as our SJT, making it unclear how effectively they would cover the target performance dimensions.

Table 12. Existing SJT Sources for Item Development

Measure	Reference	Service	Items in Original	Items in Final Pool
Criterion SJT	Knapp & Tremble (2007)	Army	27	27
NCO SJT-X	Knapp, McCloy, & Heffner (2004)	Army	3	0
SJT	Knapp, McCloy, & Heffner (2004)	Army	40	22
Longitudinal Validation Second Tour (LVII) SJT	Hanson & Borman (1989; 1995)	Army	49	0 ^a
Tacit Knowledge Test	Sternberg et al. (1999)	Army	13	1
Weighted Airman Promotion System (WAPS) SJT	Sullivan et al. (2018)	Air Force	25	5
Coast Guard SJT	Cullen & Koch (2012)	Coast Guard	29	4

^aExcludes five near-duplicate items that were dropped after the preliminary review exercise, since we used the newer versions of these items from Knapp et al., (2004) Situational Judgment Test as a starting point.

Therefore, we conducted a preliminary review exercise to gauge the adequacy of our overall pool and to determine which items were the best candidates for modification. Several members of the project team¹¹ independently assessed the items in terms of (a) the degree of editing needed to generalize across Services, (b) the degree of editing needed to apply to E-3s and E-4s, and (c) alignment with the target performance dimensions (see Table 13). After the review exercise, we removed seven items from the pool due to the presence of near-duplicate items between two

¹¹ A total of six project staff participated. Three provided input on E-3/E-4 applicability, four provided input on cross-service generalizability, and five assessed linkages to the performance taxonomy.

of our source SJTs. Next, we removed an additional 61 items that were deemed unlikely to work at the E-3 and E-4 levels or to generalize across Services, even with extensive editing.

Table 13. Performance Dimensions Targeted by SJT

Performance Category	Dimension	Sub-Dimension
Technical Performance	– Decision Making, Problem Solving, and Innovation	– Decision Making, Problem Solving, and Innovation
Organizational Citizenship & Peer Leadership	– Planning and Structuring Work	– Providing Structure
		– Teamwork
		– Self-Management
	– Conscientious Initiative	– Self-Development
		– Persistence
	– Support for Peers	– Initiative
		– Helping Peers
		– Cooperating
		– Courtesy & Respect
	– Organizational Support	– Accepting Differences
		– Selfless Service
		– Support for the Organization
		– Integrity/Moral Courage

We then assessed the remaining 51 items for coverage of the target performance dimensions. Reviewers could map up to five of the 14 target subdimensions to each item. If at least two reviewers agreed on a subdimension, an item was considered to measure that subdimension. Similarly, we counted an item as measuring a major dimension if two or more reviewers mapped it to at least one of the corresponding subdimensions. As a secondary check, we also designated the dimension with the most SME endorsements for each item as that item’s primary dimension (with multiple primary dimensions in the event of a tie).

Ultimately, we determined that the five major dimensions were of greater interest than the subdimensions since (a) it would be difficult to adequately represent 14 subdimensions in a 20-item test while simultaneously accounting for other item and test characteristics, and (b) the subdimensions’ specificity made them less likely to endure after several rounds of item editing. As shown in Table 14, we achieved good coverage of all five performance dimensions with the initial 51 items.

Table 14. Initial Coverage of Target Performance Dimensions by Potential SJT Items

Dimension	All Mappings	Primary Dimension Mappings
Decision Making, Problem Solving, and Innovation	20	8
Planning and Structuring Work	37	14
Conscientious Initiative	14	9
Support for Peers	25	20
Organizational Support	16	5

After the review exercise, we obtained three additional SJTs designed for the Coast Guard, Air Force, and Army. Since the initial 51 items were already sufficient for our purposes, we did not conduct a full group review of the new items. Instead, one of the primary staff members responsible for item editing reviewed these items and selected 10 that showed promise for adaptation, bringing the total to 61 items.

6.3 Preliminary Item Editing

Three members of the project team revised the pool of 61 items in an iterative process, modifying scenarios and responses to be relevant across Services at the E-3 and E-4 levels. For example, we updated several responses that assumed a position of authority—such as giving orders—to reflect peer leadership instead (e.g., giving advice or encouragement). We also replaced Service-specific terminology (e.g., “squad,” “platoon sergeant”) with general terms (“team,” “supervisor”). Since the in-house project team lacked the expertise to identify and resolve every issue, our goal was simply to prepare the items as much as possible before conducting reviews with military SMEs.¹²

During this stage, we also drafted additional response options for items to allow for pruning at later stages. For two of our seven source SJTs, we were able to obtain extra response options from the original SJT development process. We generally included these responses, except when there was reason to believe a response was inherently flawed (e.g., weak psychometric statistics, questionable content).

6.4 Military SME Review

To ensure the items were appropriate for DoD-wide use, we conducted four-hour review workshops with SMEs from each Service. These workshops were conducted via teleconference and facilitated by two members of the project team who also contributed to the preliminary item search and editing. In some cases, our CMAP partners or their colleagues participated as additional facilitators.

We largely relied on the CMAP to identify appropriate SMEs in each Service; our main specification was that SMEs should be familiar with the jobs of E-3s and E-4s within their respective Services. As a result, the composition of SME groups varied somewhat across Services, although most SMEs were Non-Commissioned Officers (NCO) who supervised and/or trained E-3s and E-4s. See Table 15 for a summary of SME characteristics by Service.

¹² One of the three editors was a former Coast Guard E-6 with some cross-service expertise and provided initial SME input at this stage.

Table 15. Characteristics of the SJT Review Workshop Participants

Service	Workshop Order	N	Pay Grades	Job Specialties/Departments
Air Force	1-2 ^a	4	E-7	Integrated Instruments Flight Control Systems, Surgical Technician, Aircraft Maintenance
Army	3	2	E-7	Signal School Career Manager
Navy	4	10	E-7, E-8, O-5	Hospital Corpsman, Aircrew, Parachute Rigger, Fire Controlman, Gas Turbine System Technician, Submarine Officer
Coast Guard	5	6	E-6, E-7	Instructor ^b , Operations Specialist, Electronics Technician, Information Systems Technician, Storekeeper, Boatswains Mate
Marine Corps	6	8	E-4, E-5, E-6	Manpower and Reserve Affairs

^a Due to scheduling conflicts, we facilitated two consecutive Air Force workshops with two SMEs each. ^b All Coast Guard SMEs were instructors in various specialties at the same training center.

Each SME group reviewed all 61 items for relevance to E-3s and E-4s in their Service. To promote a thorough review of all items, we divided the pool into three blocks of 20 or 21 items and rotated the order of blocks for each new workshop. During the workshops, we took notes on any potential issues and solicited feedback to resolve them, which we implemented during or soon after each workshop. We updated the items in an iterative fashion, with each set of SMEs reviewing versions that incorporated the previous group's feedback. When suggesting edits, we encouraged SMEs to consider how they would affect an item's relevance to other Services. We also took notes on the reasons for various edits, which helped us avoid introducing new issues for Services that had already provided feedback earlier in the process.

6.5 Post-Workshop Review

After the workshops were completed, we conducted additional reviews and revisions to prepare the item pool for pilot testing. First, we revisited our notes from earlier workshops to confirm that known issues were not re-introduced during subsequent revisions. We also reviewed the items for spelling, grammar, clarity, and other stylistic issues.

Due to the sequential nature of the SME reviews, as well as substantial between-service differences in policies and day-to-day practices, we were ultimately less confident that certain items generalized to all Services. Therefore, we decided to conduct a final rating exercise to preserve this information for future steps that fall beyond the scope of the present effort. After consolidating any workshop notes that indicated potential threats to generalizability, both workshop facilitators independently rated each item's generalizability on a six-point scale (see Table 16).¹³

¹³ During this rating process, we made a final attempt to address issues that affected our generalizability ratings, when feasible. Reported ratings reflect these updates.

Table 16. Summary of Item Generalizability Ratings

Item Count	Rating Value	Rating Description
2	1	SMEs confirmed that item does not apply to one or more Services
0	2	Have reason to suspect that item does not apply to one or more Services
4	3	Item technically applies to all Services, but applicability is limited in one or more Services (e.g., scenario is substantially rarer in some Services; scenario may be contrived so that it could "technically" occur in a particular Service)
4	4	Item applies to all Services but has substantial limitations to within-service generalizability (e.g., requires hypothetical assumptions about the test-taker's identity--such as gender or parental status--that could limit applicability to many test-takers).
6	5	Item applies to all Services but has minor limitations to within-service generalizability (e.g., imposes hypothetical assumptions to test a universal problem, such as assuming the test-taker lacks specific knowledge they will need for a new assignment).
45	6	Item has minimal limitations to generalizability

Our primary concern was issues that could make a scenario implausible in one or more Services, but we also considered factors that could reduce the relevance to certain groups within a given Service. We did not account for potential between-service differences in scoring, since these can (and should) be accounted for in follow-up research by collecting Service-specific effectiveness ratings. In addition, our ratings should be interpreted with caution for Services that reviewed the items earlier in the process, since some items changed substantially during later workshops. To address this limitation, we recommend soliciting feedback on potential generalizability issues during future rounds of SME review and/or pilot testing.

After reviewing the items independently, the facilitators discussed the reasons for any discrepancies and decided on a final rating. As shown in Table 16, we scored all but two items in the 3-6 range, suggesting the item pool largely generalized across Services. The two items that received a score of 1 had fundamental generalizability issues that we were unable to resolve. We dropped these items from the pool, bringing the total number of items to 59. We assigned four additional items a score of 3, indicating they applied to all Services but might be *less* applicable to at least one Service. For example, one scenario involved the respondent being asked to report to duty on their day off, which SMEs indicated was unlikely in the Marine Corps due to their policy of assigning backups. The SMEs told us the scenario was plausible with added language indicating the assigned backup was unavailable. However, the original critique still implies that the scenario would be uncommon for Marines. (In addition, the assigned backup reference may be less relevant to other Services). We did not remove items of this variety from the pool, but we recommend additional scrutiny of their relevance in follow-up research.

We assigned a rating of 4 or 5 to nine items, indicating that these items applied across Services but imposed hypothetical assumptions that could affect their personal relevance to certain test-takers (e.g., asking the test-taker to imagine they are a parent when many test-takers are not). These assumptions were not an explicit focus of the revision process and were usually present in the original SJT items. Since they were relatively common in established military SJTs, we did not view such hypotheticals as inherently problematic. However, we chose to differentiate them in our ratings as one potential factor to consider during future steps of the development process,

such as form assembly. All else equal, it may be preferable to use an item with more universal relevance over a psychometrically similar item that requires a specific hypothetical perspective.

The final overlength SJT form contains a total of 59 items and 502 response options, with the number of response options per item ranging from six to 12 (mean = 8.5; median = 8). With approximately triple the number of items and double the response options needed for the operational SJT, there should be significant flexibility for winnowing in subsequent development stages.

In addition to the items for pilot testing, the overlength SJT form includes test-taker instructions and a seven-point effectiveness rating scale (see Figure 3). We adapted the instructions from Knapp et al., (2004) and used the rating scale from Knapp and Tremble (2007). In keeping with the DoD-Wide Criterion Measurement Research Plan (Appendix C), the instructions assume that the SJT will be administered in a paper-and-pencil format. The Flesch-Kincaid Grade Level for the overlength form is 6.7, substantially below the target eighth grade reading level.

Effectiveness of the Action						
Ineffective action.		Moderately effective action.			Very effective action.	
The action is likely to lead to a bad outcome.		The action is likely to lead to a passable or mixed outcome.			The action is likely to lead to a good outcome.	
———— Low ————		———— Moderate ————			———— High ————	
1	2	3	4	5	6	7

Figure 3. SJT Effectiveness Rating Scale.

7.0 SUMMARY AND RECOMMENDATIONS

The cross-service criterion instruments developed for this project—measures of attrition, attitudes and performance—are important DoD-wide.

7.1 Summary

To fulfill the recommendations that resulted from the first phase of the project (Allen et al., 2020), researchers developed a set of cross-service criterion measures. The criterion measures were designed to measure the job performance, organizational outcomes, and attitude constructs for first-term enlisted personnel. Table 17 provides a crosswalk of the constructs targeted by each criterion instrument that was developed or proposed (in the case of administrative alignment variables). Each construct is addressed by at least one new criterion measure, and in most cases, by two or more.

Table 17. Summary of Performance, Organizational Outcome, and Attitude Constructs Targeted by Cross-Service Criterion Measures

Constructs	Cross-Service Criterion Measures					
	Align Admin. Variables	EOT Assessment	IU Assessment	Exit Survey	PRS	SJT
Performance Category/Dimension						
Technical Performance						
Task Performance					X	
Decision Making, Problem Solving, and Innovation					X	X
Communication					X	
Safety and Security					X	
Consciousness						
Organizational Citizenship & Peer Leadership						
Planning and Structuring Work					X	X
Conscientious Initiative					X	X
Support for Peers					X	X
Organizational Support					X	X
Psychosocial Well-Being						
Adapting to Stressful Situations		X	X	X	X	
Counterproductive Work Behavior	X	X	X		X	
Physical Performance						
Physical Fitness	X	X	X		X	

(continued)

Table 17. Summary of Performance, Organizational Outcome, and Attitude Constructs Targeted by Cross-Service Criterion Measures (continued)

Constructs	Cross-Service Criterion Measures					
	Align Admin. Variables	EOT Assessment	IU Assessment	Exit Survey	PRS	SJT
Organizational Outcomes						
Attrition	X			X		
Reprimands	X	X	X			
Performance Records		X	X			
Merit-based Awards		X	X			
Attitudes						
Work Satisfaction		X	X	X		
Morale		X	X			
Organizational Commitment		X	X			
Withdrawal		X	X	X		
Cognitions/Intentions						
Person-Environment Fit (PE Fit)		X	X			

Note. EOT = End-of-training, IU = In-Unit, PRS = Performance Rating Scales, SJT = Situational Judgment Test.

7.2 Recommendations for Evaluating New Measures: Criterion Measurement Research Plan

The final step of this project was to develop a research plan for evaluating the psychometric characteristics of the new set of cross-service measures. The research plan (provided in Appendix C) describes a strategy for conducting the data collections that will be required for thorough evaluation of each criterion measure. The plan describes the challenges and provides possible solutions for:

- Obtaining resources for data collection,
- Using technology for data collection,
- Enhancing data quality,
- Obtaining IRB approval,
- Evaluating the accuracy of attrition data,
- Maximizing usefulness of data collections,
- Streamlining training and in-unit data collection,
- Planning EOT and IU data collection.

The remainder of the research plan describes three proposed projects, listed in Table 18. While predictor validation is likely to be added to some of these projects, the research plan focuses on the evaluation of the criterion measures.

Table 18. Proposed Projects and Purposes for Evaluating Criterion Measures

Project	Purpose
Attrition	To evaluate procedures for constructing attrition variables, the integrity of attrition data.
End-of-Training (EOT)	To evaluate performance appraisal ratings and a self-assessment survey administered at the end of technical training.
In-Unit (IU)	To evaluate performance appraisal ratings, a situational judgment test, and a self-assessment survey administered to E-3 and E-4 Service members having 18 to 36 months of experience.

7.3 Conclusion

The new unified set of criterion procedures and measures can be used in joint-Service projects to evaluate the criterion-related validity of the ASVAB and other predictor measures. The research plan condenses ideas of many military personnel researchers and provides a useful starting point for planning future joint-Service validation efforts.

8.0 REFERENCES

- Air Force Survey Office (2017). *2017 Military Exit Survey*.
- Allen, M. T., Russell, T. L., Ford, L. A., Kirkendall, C. D., & Carretta, T. R. (2020). *Identification of criterion constructs and measures for joint-service enlisted personnel*. AFRLRH-WP-TR-2020-0019, Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Warfighter Interface Division, Collaborative Interfaces and Teaming Branch.
- Binggeli, J., Garrick, R., Lancaster, W., Bolger, D., Hinkson, D., Hardy, C., & Comer, K. (2016). *Marine Corps Data Collection Strategy Study: Final Report*. Booz Allen Hamilton, Inc.
- 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.
- 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.
- Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology*, 43, 313–334.
- Cullen, M. J., & Koch, A. (2012). *Development of a situational judgment test to measure integrity in Coast Guard cadets*. United States Coast Guard Academy.
- Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology*, 90(6), 1241–1255.
- Giacalone, R. A. (2000). *Analysis of the Revised Army Career Transitions Survey (ACTS) and comparison with the Fall 1996 Sample Survey of Military Personnel (SSMP): Results and recommendations* (SR 2000-03). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Gibson, J. L., Hackenbracht, J., & Tremble, T. R. (2014). An event history analysis of first-term soldier attrition. *Military Psychology*, 26, 55–66.
- Gonzalez-Mulé, E., Mount, M. K., & Oh, I.-S. (2014). A meta-analysis of the relationship between general mental ability and nontask performance. *Journal of Applied Psychology*, 99, 1222–1243.
- Hanson, M. A., & Borman, W. C. (1989, April). Development and construct validation of a Situational Judgment Test as a job performance measure for first line supervisors. In W. C. Borman (Chair), *Evaluating "Practical 10": Measurement issues and research applications in personnel selection and performance assessment* [Symposium]. Society for Industrial and Organizational Psychology 4th Annual Conference, Atlanta, GA, United States.
- Hanson, M. A., & Borman, W. C. (1995). *Development and construct validation of the Situational Judgment Test* (ARI Research Note 95-34). Personnel Decisions Research Institute.

- 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.
- Hoover, E. C., Randolph, J. S., Elig, T. W., & Klein, P. M. (2001). *Overview of the 2000 Military Exit Survey*. Defense Manpower Data Center.
- Hunter, T. M. (2006). *An analysis of the impact of job search behaviors on Air Force company grade officer turnover*. Air Force Institute of Technology.
- Knapik, J. J., Jones, B. H., Hauret, K., Darakjy, S., & Piskator, E. (2004). *A review of the literature on attrition from the military services: Risk factors for attrition and strategies to reduce attrition*. U.S. Army Center for Health Promotion and Preventative Medicine.
- 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). Brooks AFB, TX: Human Resources Directorate Manpower and Personnel Research Division.
- Knapp, D. J., & Kirkendall, C. D. (2018). *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., & Kirkendall, C. D. (2020). *Tier One Performance Screen initial operational test and evaluation: Capstone report* (Technical Report 1380). 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 21st-century Army NCO performance* (Technical Report 1145). 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* (TR 1205). 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.
- 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., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, 37, 426-441.
- O'Boyle, E., Jr., & Aguinias, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology*, 65, 79-119.
- Ployhart, R.E. & Ehrhart, M.G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selections and Assessment*, 11(1), 1-16. <https://doi.org/10.1111.1468-2389.00222>

- 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.
- Ramsberger, P. F., & Babin, N. E. (2005). *Recent ARI research on enlisted attrition from the United States Army* (FR-05-68). Human Resources Research Organization.
- Sipes, D.E., Strickland, W.J., Laurence, J.H., DiFazio A.S. & Wetzel, E.S. (2000). *Training Base Attrition: Analysis and Findings* (FR-00-38). Human Resources Research Organization.
- Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Tremble, T. R., Snook, S., Williams, W. M., Wagner, R. K., & Grigorenko, E. L. (1999). *Tacit knowledge in the workplace* (Technical Report 1093). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Strickland, W. J. (Ed.). (2005). *A longitudinal examination of first term attrition and reenlistment among FY1999 enlisted accessions* (Technical Report 1172). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Sullivan, T. S., Burgyone, T. C., McCloy, R. A., & Whetzel, D. L. (2018). *Review of Situational Judgment Test (SJT) prototype development process and materials* (Technical Report 061). Human Resources Research Organization.
- U.S. Coast Guard (2016). *Coast Guard Organizational Assessment Survey*.
<https://www.dcms.uscg.mil/Our-Organization/Assistant-Commandant-for-Human-Resources-CG-1/Strategy-and-HR-Capability-CG-1B/CG-OAS/>
- U.S. Marine Corps (2018). *Enlisted Retention Survey*.
- U.S. Marine Corps (2018). *Milestone Exit Survey*. <https://go.max.gov/EMLS>

9.0 LIST OF ACRONYMS, ABBREVIATIONS AND SYMBOLS

ACFT	Army Combat Fitness Test
ADSO	Active Duty Service Obligations
AFRL	Air Force Research Laboratory
AFSC	Air Force Specialty Code
AF-WIN	Air Force Work Interest Navigator
APFT	Army Physical Fitness Test
ASVAB	Armed Services Vocational Aptitude Battery
ATRRS	Army Training Requirements and Resource System
AWOL	Absent without Leave
BCA	Body Composition Assessment
BTMS	Basic Training Management System
CeTARS	Corporate Enterprise Training Activity Resource Systems
CFT	Combat Fitness Test
CMAP	Criterion Measures Advisory Panel
CWB	Counterproductive Work Behavior
DIBRS	Defense Incident Based Reporting System
DMDC	Defense Manpower Data Center
DoD	Department of Defense
DoDM	DoD Manual
EOT	End-of-Training
FA	Fitness Assessment
IMT ALQ	Initial Military Training Army Life Questionnaire
IRB	Institutional Review Board
ISC	Inter-Service Separation Code
IU	In-Unit
JON	Job Opportunities in the Navy
LVII	Longitudinal Validation Second Tour
MCTFS	Marine Corps Total Force System
MILSPERSMAN	Military Personnel Manual
MOS	Military Occupational Specialty
NSIPS	Navy Standard Integrated Personnel System

OCSAR	Operational Stress Control and Readiness
PE	Person-Environment
PEVT	Person Event
PFA	Physical Fitness Assessment
PFT	Physical Fitness Test
PRS	Performance Rating Scales
PRT	Physical Readiness Training
RTC	Recruit Training Center
SJT	Situational Judgment Test
SME	Subject Matter Expert
SPD	Separation Program Designator
TAPAS	Tailored Adaptive Personality Assessment
TAPDB	Total Army Personnel Database
TTMS	Technical Training Management System
WAPS	Weight Airman Promotion System

APPENDIX A. Final Procedures for Aligning Administrative Variables

ATTRITION

Raw Variables Required

- Variables obtained in databases from Table 3.1 (pg. 8)
 - Accession date (i.e., date of entry)
 - Separation date (i.e., date of exit)
 - Separation reason

Constructed Attrition Type

- Ideally, obtain attrition as Inter-service Separation Codes (ISC) from DMDC
- If ISC are not obtained, then Service-specific separation codes should be recoded to ISC (crosswalk available in the For Official Use Only version of this report)
- ISC codes considered non-attrit should be coded as missing
 - 1 (Expiration of term of service)
 - 3–8 (Early release)
 - 40 (Officer commissioning program)
 - 52 (Retirement)
 - 100 (Immediate re-enlistment)
- ISC codes should be coded into broader attrition types (Table 3.2; pg. 10)
 - Create separate 0 or 1 indicator variables for whether a record has an attrition code in that broader attrition category

Constructed Attrition Time

- Overall time-to-attrition (i.e., time in service)
 - Compute difference between separation and accession dates
 - If no separation date, code as missing
 - If value is negative, manually check for issues including:
 - Possible data entry errors (e.g., accession date in the future)
 - Re-enlistment
 - Recode as missing any negative values that cannot be rectified
- Cumulative time-to-attrition
 - Create separate variables for cumulative attrition in monthly intervals up to 48 months (e.g., 1-month attrition, 2-month attrition, 3-month attrition..., 46-month attrition, 47-month attrition, and 48-month attrition)
 - For each cumulative time-to-attrition variable, code as
 - 0 if overall time-to-attrition is greater

- 1 if overall time-to-attribution is less
- If the separation code is considered non-attrib, recode any value of 1 to 0
- Censored data beyond the overall time-to-attribution value should be treated as missing
- Example
 - Overall time-to-attribution value is 12 months
 - 3, 6, and 9-month attrition = 0
 - 12-month attrition = 1
 - 15 through 45-month attrition = NA
- Overall attrition
 - Code as most recent non-missing cumulative time-to-attribution value
 - If no non-missing cumulative time-to-attribution value, code as missing

TRAINING

Raw Variables Required

- Variables obtained in databases from Table 3.3 (pg. 14)
 - Training grades
 - Training status codes (e.g., graduation, recycle/washback, and fail/washout)

Constructed Training Status

- Dichotomous indicator variables
 - Graduated from training
 - 0 = Did not graduate from training
 - 1 = Graduated from training
 - Recycled/washed back in training
 - 0 = Did not recycle/wash back in training
 - 1 = Recycled/washed back at least once in training
 - Failed/washed out in training
 - 0 = Did not fail/wash out in training
 - 1 = Failed/washed out at least once in training
- If academic versus non-academic reasons for recycling or failing are available (Table 3.4; pg. 16), the above recycled and failed variables above can be broken down into academic and non-academic versions
 - Recycled in training for academic reasons
 - 0 = Did not recycle in training or recycled for non-academic reasons
 - 1 = Recycled in training for academic reasons
 - Recycled in training for non-academic reasons
 - 0 = Did not recycle in training or recycled for academic reasons

- 1 = Recycled in training for non-academic reasons
 - Failed in training for academic reasons
 - 0 = Did not failed in training or failed for non-academic reasons
 - 1 = Failed in training for academic reasons
 - Failed in training for non-academic reasons
 - 0 = Did not fail in training or failed for academic reasons
 - 1 = Failed in training for non-academic reasons
- Status count variables
 - Number of times recycled prior to graduation
 - If the graduation indicator variable = 0, code as missing
 - If the graduation indicator variable = 1, count the number of recycles indicated in the service member's training records
 - Number of times recycled prior to failing out
 - If the fail indicator variable = 0, code as missing
 - If the fail indicator variable = 1, count the number of recycles indicated in the service member's training records
- If training statuses are not available, but training grades are, recode grades into pass/fail status based on the course criteria
- If both training statuses and grades are available, perform the above variable construction procedures and compare them as a quality control measure. Examples of possible issues that may be identified if the training statuses and recoded grades do not agree:
 - If separate training records or databases are merged, training records should be verified to ensure that the data was merged accurately
 - Verify whether the training statuses and grades are describing the same training event or course, or instead pertain to different ones

PHYSICAL FITNESS

Raw Variables Required

- Variables obtained in databases presented in Table 3.5 (pg. 18)
 - Physical fitness test scores
 - Physical fitness test classification

Standardized Physical Fitness Test Scores

- Standardize via z-scores:

$$z_i = \frac{x_i - \bar{x}}{s}$$

- where z_i is the z-score for an individual, x_i is the raw score for an individual, $\bar{x}$ is the mean score for the sample, and s is the score standard deviation for the sample

- If scores are extracted as part of a standalone datafile, they should be standardized across individuals in the datafile
- If the scores are part of a continually updated longitudinal datafile, scores should be standardized across individuals each year, so as not to recompute scores for everyone with every update of the datafile

Physical Fitness Test Classification

- Although classification variables are available in the raw databases, they need to be aligned across Services and recoded as:
 - P = Passed physical fitness requirements
 - F = Did not pass physical fitness requirements
 - W = Obtained a waiver for physical fitness requirements

DISCIPLINARY INCIDENTS

Raw Variables Required

- Variables should ideally be obtained from the Defense Incident-Based Reporting System (DIBRS)
 - Date of Offense
 - Date of Adjudication
 - Offense Identifier (see Tables 7 and 8 of DoDM 7730.47-M-V2 for codes)
 - Offense Result (attempted or completed)
 - Involvement (principal, accessory, conspiracy, or solicit)
 - Action Taken (no action, administrative, non-judicial, or judicial)

Count Disciplinary Incident Descriptor Variables

- Offense Identifier variables should first be recoded into broader categories of offenses against person, property, or society (see Tables 7 and 8 of [DoDM 7730.47-M-V2](#)).
- Aggregate disciplinary incident records for each service member by constructing separate numeric variables to count the number of times a service member has been indicated for each of the following raw DIBRS variables:
 - Overall offenses, which should be subcategorized into:
 - Offense against person
 - Offense against property
 - Offense against society
 - Offense attempted
 - Offense completed
 - Involved as principal
 - Involved as accessory, conspiracy, or solicit

- Action taken: None
- Action taken: Administrative or non-judicial
- Action taken: Judicial

APPENDIX B. Training for Performance Rating Scales

FRAME OF REFERENCE TRAINING FOR PERFORMANCE RATING SCALES

Air Force Version

1

Learning Objectives

This training will help you make more accurate ratings. After completing this training, you will be familiar with:

- The areas of performance you will be rating
- The scales you will use to make your ratings
- Ways to avoid common rating errors

2

What Performance Will I Be Rating?

You will be rating Airmen on twelve performance areas. The areas fall into four broad categories of performance. **Click on the categories** to see the performance areas.

Technical Performance

Task Performance
Making Decisions and Solving Problems
Communicating with Other
Maintaining Safety and Security

Organizational Citizenship and Peer Leadership

Planning and Structuring Work
Showing Persistence and Taking Initiative
Showing Support for Peers
Showing Support for the Organization

Physical Performance

Physical Fitness

Maintaining Own Well-being

Adapting to Stressful Situations
Avoiding Unproductive Work Behaviors
Avoiding Harmful Work Behaviors

3

Understand the Performance Dimensions

To make accurate ratings, you need to understand the performance areas you will be rating. As each set behaviors appears on the left, **click on the performance area on the right that best matches the behaviors.**

Behaviors

- Takes responsibility for own learning
- Persists with extra effort.
- Takes initiative to accomplish objectives.
- Volunteers for additional assignments.

Performance Areas

Task Performance

Making Decisions and Solving Problems

Communicating with Others

Maintaining Safety and Security

Planning and Structuring Work

Showing Persistence and Taking Initiative

Showing Support for Peers

Showing Support for the Organization

Physical Performance

Adapting to Stressful Situations

Avoiding Unproductive Work Behaviors

Avoiding Harmful Work Behaviors

Presentation Notes:

- The user will view and match behaviors for all performance areas.

4

Understanding the Rating Scales

In addition to understanding the performance you will be rating, you should also understand how to use the rating scales.

You will be using two types of scales: a relative scale and a frequency scale.

Relative Rating Scale

	Below Average		Average	Above Average		Well Above Average	Not Observed/ Cannot Rate
Ratee	Lowest 25%	Lowest 40%	Average 40-60%	Top 40%	Top 25%	Top 10%	
Airman 1							
Airman 2							

Frequency Rating Scale

Ratee	Never	1 Time	2-3 Times	4-5 Times	More Than 5 Times	Not Observed/ Cannot Rate
Airman 1						
Airman 2						

Presentation Notes:

- When user clicks the button, he/she will see callout bubbles come in to describe the rating levels and the frequencies.

5

How to Use the Relative Rating Scale

	Below Average		Average	Above Average		Well Above Average	Not Observed/ Cannot Rate
Ratee	Lowest 25%	Lowest 40%	Average 40-60%	Top 40%	Top 25%	Top 10%	
Airman 1							
Airman 2							

You will use the relative rating scale to rate Airmen on the first ten performance areas:

1. Task Performance
2. Making Decisions and Solving Problems
3. Communicating with Others
4. Maintaining Safety and Security
5. Planning the Structuring Work
6. Showing Persistence and Taking Initiative
7. Showing Support for Peers
8. Showing Support for the Organization
9. Physical Performance
10. Adapting to Stressful Situations

6

How to Use the Relative Rating Scale

	Below Average		Average	Above Average		Well Above Average	Not Observed/ Cannot Rate
Ratee	Lowest 25%	Lowest 40%	Average 40-60%	Top 40%	Top 25%	Top 10%	
Airman 1							
Airman 2							

When using the relative rating scale:

- Indicate how well each Airman has performed *in comparison to other Airmen that you have observed*.
- Because most Airmen perform their jobs to standard (i.e., minimum acceptable proficiency) or better, rating a Airman as 'below average' does NOT necessarily mean that the Airman has not performed to standard.

Now test your understanding of relative rating scale in a practice rating exercise.

Start the Practice Rating Exercise

7

Practice Rating Exercise

Think of five Airmen you know and rate them on task performance.

Task Performance

Example Behaviors:

- Performs job-specific and Air Force-wide tasks proficiently.

How well does each Airman perform tasks compared to other Airmen you have observed?

	Below Average		Average	Above Average		Well Above Average	Not Observed/ Cannot Rate
Ratee	Lowest 25%	Lowest 40%	Average 40-60%	Top 40%	Top 25%	Top 10%	
Airman 1							
Airman 2							
Airman 3							
Airman 4							
Airman 5							

See How You Did

Presentation Notes:

- Everyone will see each of the next three slides, regardless of how they performed on the practice rating exercise.
- For the peer ratings version, this will say: "Think of five of your peers and rate them on their task performance."

8

Practice Rating Exercise Feedback

Did you rate everyone or almost everyone as average?

This is NOT recommended.

It is unlikely that everyone performs the same in any area.

Ratee	Below Average		Average	Above Average		Well Above Average	Not Observed/ Cannot Rate
	Lowest 25%	Lowest 40%	Average 40-60%	Top 40%	Top 25%	Top 10%	
Airman 1			X				
Airman 2			X				
Airman 3				X			
Airman 4			X				
Airman 5			X				

9

Practice Rating Exercise Feedback

Did you use only the high end of the scale?

This is NOT recommended.

It is unlikely that everyone is a top performer.

Ratee	Below Average		Average	Above Average		Well Above Average	Not Observed/ Cannot Rate
	Lowest 25%	Lowest 40%	Average 40-60%	Top 40%	Top 25%	Top 10%	
Airman 1						X	
Airman 2						X	
Airman 3					X		
Airman 4						X	
Airman 5					X		

10

Practice Rating Exercise Feedback

If you used the full range or close to the full range of the scale, great job!

This is the recommended approach.

Relative performance is typically distributed across the scale. In any group, some are average, some are above or well above average, and some are below or well below average.

	Below Average		Average	Above Average		Well Above Average	Not Observed/ Cannot Rate
Ratee	Lowest 25%	Lowest 40%	Average 40-60%	Top 40%	Top 25%	Top 10%	
Airman 1	X						
Airman 2			X				
Airman 3		X					
Airman 4					X		
Airman 5				X			

11

Frequency Rating Scale

Ratee	Never	1 Time	2-3 Times	4-5 Times	More Than 5 Times	Not Observed/Cannot Rate
Airman 1						
Airman 2						
Airman 3						

You will use the frequency scale to rate Airmen on the last two performance areas:

- Unproductive Work Behaviors
- Harmful Work Behaviors

When using the frequency scale:

- Consider how often you have observed this behavior in the Airman during the specified timeframe.
- Not all Airmen will display all the behaviors.

12

Avoid Common Rating Errors

Now that you know how to use the scales, review the four common errors that you should avoid when making performance ratings. Click on each type of error to see more information.

1. Stereotyping

You should avoid letting irrelevant factors (e.g., race, sex, and age) influence your ratings.

Rate each person based on his/her performance only.

2. First Impression

You should avoid letting first impressions or your most recent interactions with the person influence your ratings.

Rate each person based on his/her typical performance.

3. General Impressions

It is unlikely that any person performs the same in all areas. You should avoid letting general impressions cause you to rate a person the same in all areas.

Rate each person based on his/her strengths and weaknesses.

4. Leniency/Severity

It is unlikely that everyone performs the same in any area. You should avoid rating everyone the same (high, low, or average) in any area.

Rate each person based on his/her performance.

13

Summary

This training has provided information that should help you make more accurate ratings by familiarizing you with:

- The areas of performance you will be rating
- The rating scales you will use
- Ways to avoid common rating errors

14

APPENDIX C. DoD-Wide Criterion Measurement Research Plan

I. Objective

The DoD recently completed the *Criterion Measures Project* to develop DoD-wide 1st term criterion measures—measures of attrition, attitudes and performance that are important DoD-wide. These measures are listed in Table C.1. Training measures are intended for administration at the end of technical training. In-unit measures are intended for administration to first-term enlisted personnel approximately 18 months into the first term of enlistment (i.e., E3 - E4 level). The Exit Survey is intended for personnel who have exited Service or plan to. While some of the new measures tap technical proficiency in a general way, the new measures are not intended to assess detailed occupation-specific task performance.

Table C.1. New DoD-Wide Criterion Measures

Criterion Measure	Estimated administration time	Target Enlisted Population	
		End-of-Training (EOT)	In-Unit (IU)
Procedures for aligning administrative data (e.g., physical fitness scores, training outcomes, attrition data) across Services	N/A	X	X
Self-Report Assessment	~15 minutes	X	X
Performance rating scales to be completed by peers and supervisors/trainers	~20 minutes	X	X
Situational Judgment Test	~30 minutes		X
Exit Survey ^a	~15 minutes		

^a Intended for personnel who have exited or plan to leave Service during or at the end of their first term.

The objective of the research described in this document is to evaluate the new criterion measures.

II. Challenges and Possible Solutions

Research projects that involve collecting data in the field require a lot of support, resources, and planning. Several challenges are listed in Table C.2 along with potential ways of addressing them.

Table C.2. Data Collection Challenges and Possible Solutions

Challenges	Possible Solutions
Obtaining resources	• Obtain support at high levels within DoD and each Service.
Using technology	• House automated criterion measures and the data on a DoD database. • Purchase military-owned data collection devices. • Use a mobile wireless networking router to access the Internet, where needed.
Enhancing data quality	• Use trained staff to collect data. • Piggyback on other data collection efforts, when possible. • Develop instructional training for unproctored data collections.
Obtaining IRB approval	• Submit one IRB protocol approval request for all Services.
Evaluating accuracy of attrition data	• Compare attrition variables from administrative records to Exit Survey responses for personnel who have exited or plan to exit.
Maximizing usefulness	• Use the criterion data to also validate predictors.
Streamlining training and in-unit projects	• Adopt a concurrent design (as opposed to a longitudinal one)
Planning EOT data collection	• Select high density occupations for inclusion. • Select schoolhouses where data collection is feasible.
Planning IU data collection	• Task intact units for data collection.

Obtaining Resources

Challenges

Data collection is resource intensive. If measures are to be administered by automated devices, the measures must be programmed and tested for administration on those devices. If measures are administered on computers, the Services may need to obtain access to computer facilities or portable computers. If the data collections are face-to-face, data collectors must be identified and trained. Any paper-and-pencil measures need to be printed on scannable forms and shipped to/from data collection sites. If the Services have insufficient staff, contractors must be hired. Facilities and equipment must be secured. Finally, Service personnel time is required for completing assessment tools, including time (a) for points of contact (POC) to coordinate site visits, (b) for supervisors and peers to complete performance rating scales, and (c) for individual service members to complete assessments.

Possible Solutions

The most important step toward securing resources is to obtain appropriate high-level support within DoD and each Service. The Services will need financial support as well as the endorsement of high-ranking officials.

Using Technology

Online data collection methods eliminate the need to print, process, store, scan, and destroy data collection forms. In short, automation reduces data transmittal and preparation time/errors. Technology also allows the use of alternative assessment formats such as video and audio presentation of measures.

Challenges

There are three main factors that limit the Services' ability to conduct data collections on-line: (a) a DoD-certified server, (b) electronic data collection devices, and (c) Internet access.

Certified Server. Servers and software used on them to collect data from military personnel must meet various security certification requirements. It is difficult for contractors to obtain appropriate certifications for their servers and software. For this reason, the Services' data collection efforts often take place on their own servers. It would be highly desirable for this project to house automated criterion measures and the data on a DoD server.

Electronic Devices. Another limitation has to do with the availability of electronic devices for data collection. Over the years, the Services have used several methods for computer administration. Off and on since the 1980s, data collection teams have carried computers to data collection sites. The Army often uses computers at Digital Testing Facilities (DTFs) or Education Centers at training installations. In at least one instance, the Air Force rented 100 computers for testing at a site. For some projects, the Services have purchased laptops and other peripheral devices (e.g., ear buds) specifically for data collection use. Some personnel have suggested that measures that can be administered directly over smart phones would be highly desirable. The DoD is in the process of evaluating a wide array of devices for administering the Armed Services Vocational Aptitude Battery (ASVAB). So far, they have learned that smart phones and tablets can be used to administer ASVAB subtests, provided that the subtests are programmed with a smart phone in mind (Personal communication with Gwen Van Trieste and Jeff Barnes, September 13, 2019). Results regarding the quality of ASVAB scores are pending.

While using service members' personal smartphones might be convenient, there could be a few problems. First, some service members may not have smart phones. Also, given the variability in smart phones, it could be difficult to program the measures such that they could be administered over any smart phone. Finally, there might be IRB objections to *requiring* use of service members' personal devices for data collection.

For criterion data collection, military-owned devices would offer important advantages and some disadvantages. The Services could purchase a particular type of device in quantity (e.g., tablet), eliminating the need to develop and test apps across a variety of devices. The DoD could also install Common Access Card (CAC) readers to those devices. CAC readers are required for many military data collections used to collect Personally Identifiable Information (PII) from military personnel. Owning the devices, however, would require storing, securing, and shipping them. Here, lessons could be taken from the Army Research Institute (ARI) which has used its own laptops in numerous data collections.

Regarding the type of device, tablets might be the best option. Tablets are relatively inexpensive (perhaps \$200 each) compared to laptops, notebooks, or smart phones, and they have been used successfully on the ASVAB Devices project. They weigh less than laptops or notebook computers and therefore they cost less to ship to installations. Finally, the larger screen (compared to a smart phone) would make it easier to administer measures that do not fit easily on a smart phone screen such as performance rating scales.

Internet Access. A third limitation is access to the Internet at various installations. The Services have used a variety of methods for conveying electronic data to a database. In the early years, data were stored on disks or thumb drives which were later read and uploaded to a central database. In recent years, on-line administration has eliminated the need for disks because data are saved to a central server during administration. This requires access to the Internet, which may not be available to data collection staff at military installations.

The ASVAB Devices project is using mobile Wi-Fi routers to access the Internet where Wi-Fi is unavailable. Such routers require cell coverage, but typically research staff have been able to find facilities within range of cell towers at military installations. Mobile Wi-Fi routers require contracting with Internet Service Providers.

Possible Solutions

In summary, possible solutions for addressing on-line data collection challenges include:

- House automated criterion measures and the data on a DoD server.
- Purchase military-owned data collection devices (e.g., tablets).
- Use mobile Wi-Fi routers for Internet access as needed on installations.

Enhancing Data Quality

Criterion measures do not require the same security as predictor measures (where compromising the items is a concern). Therefore, the Services have used several different levels of proctoring for data collection:

- *No proctoring.* Selected participants receive an email link to the instrument (survey, performance rating scales) from an authority figure. This method is advantageous when (a) the target population is spread across geographically diverse locations and (b) the data collection instrument is relatively simple. Response rates can be very low.
- *POC contact only.* To enhance response rates, the Services have also sent the link to a POC at a military installation along with the names of participants, and the POC has tasked participants to go to the link.¹
- *POC administration.* Following instructions from research staff, a military POC tasks personnel to assemble in one place and administers the measures to the group. This

¹ All research is voluntary. We only ask that the participant go to the link. The informed consent form at the link allows subjects to decline participation.

method has been used to collect criterion data from trainees at training facilities. POCs must be trained in data collection methods.

- *Trained research staff administer measures in group settings.* POCs are always necessary for providing access to research participants but once participants are made available, the most reliable way to collect data is to use trained research staff.

Challenges

While administering measures using trained research staff is expensive (i.e., staff labor hours, travel expenses), in most situations there is not a good alternative. Response rates are low when there is no proctoring and extensive follow-up can be required to achieve adequate participation.² Methods relying on military POCs or trainers can be problematic because those personnel are typically already tasked with a full workload. It cannot be assumed that, on a broad scale across Services, POCs or trainers could be made available.

The Services have conducted projects wherein a large population/sample received an email from an authority figure with a link to a self-assessment. This method can work reasonably well when the population is very large because, even with a low response rate, the final sample size may be large enough for analysis. This method is more challenging when administering peer and supervisor ratings. In our experience, it is not possible to identify peers and supervisors ahead of the data collection using administrative data. So, typically research participants identify their peers and supervisors in proctored sessions, and those individuals are then tasked with making performance ratings. Alternatively, to collect performance ratings from peers and supervisors in an unproctored, online fashion, respondents must identify their peers and supervisors (and provide contact information such as an e-mail address) when responding to their self-assessments. In turn, links to performance rating forms can be emailed to the peers and supervisors of those respondents. This method is cumbersome, but it worked adequately for collecting supervisor ratings for junior officers in an ARI study (Optimizing Officer Classification I (OCLASS I)). The report is in preparation. However, collecting both peer and supervisor ratings for enlisted personnel in this manner could be very challenging.

Possible Solutions

When performance ratings need to be collected, plan for proctored data collection, requiring research staff to travel to data collection sites. If the Services have other avenues in place, such as access to trained POCs, those methods should be used.

When the only measure to be administered is a self-report, consider unproctored, online administration. As noted below, we recommend this approach for collecting Exit Survey data.

² One way to boost participation is through incentives. Incentives are usually not allowed for military personnel, but perhaps that possibility could be investigated further. For example, could all personnel receive a \$5 gift card for simply accessing the link and completing the informed consent form? Even \$50,000 for 10,000 participants would be fairly economical compared to the cost of travel and personnel time involved in face-to-face data collections.

Obtaining IRB Approval

Challenges

Institutional Review Board (IRB) approval can be laborious and time-consuming. To minimize the burden on individual Service's research staff, instruments and data collection procedures should be standardized. That way, one IRB should be sufficient for reviewing the entire project for all of the Services. As needed, Memoranda of Understanding (MOU) or other documents should be put in place to acknowledge the Services' agreement to follow one IRB. It is important to note, however, that the "One IRB" approach will only work if the validation is similar across Services. If the Services want to include Service-specific predictors or conduct their projects independently, each Service would need its own IRB due to use of different measures and staff.

Possible Solutions

As needed, MOU or other documents could be put in place to acknowledge the Services' agreement to follow one IRB.

Evaluating Accuracy of Attrition Data

Challenges

Two challenges to evaluating the new attrition alignment methods are that (a) it can take years for attrition data to mature for a given cohort of accessions and (b) it is difficult to know what the truth is, if administrative records are incorrect.

Possible Solutions

The Exit Survey includes 'reasons for leaving' that could be used to evaluate the accuracy of administrative data. Comparing the two would be a way of cross-checking the accuracy of data. Where mismatches in reasons for leaving occur, a subsample of the population could be interviewed over the phone.

One approach would be to collect Exit Survey data from a large sample of enlisted personnel that have left active duty in the previous fiscal year. There would be no need to wait for attrition data to mature because all participants have left active duty. One problem with this approach is that it may not be possible to get email addresses/contact information for personnel who exited in the prior year (Personal communication with Thomas Carretta, September 24, 2019). Another problem with this approach is that it could require Office of Management and Budget (OMB) approval because civilians are being surveyed (Personal communication with Sofiya Velgach, November 6, 2019), and the OMB process is quite lengthy. Given those drawbacks, surveying personnel who have left the Service is probably unworkable.

Another approach would be to identify all enlisted personnel who have notified the military of their intent to leave the Service. Those personnel would be asked to complete the Exit Survey. Several months later, after the personnel exits have been recorded in military databases, Exit Survey responses could be compared to administrative records. While it might be easier to access emails for the sample, this approach would require waiting for the administrative data to mature over time.

Several Services currently conduct operational Exit Surveys. Those operational surveys could be used in lieu of the newly developed survey if (a) the data are identifiable such that records can be matched to administrative data for each service member and (b) the items on the operational survey match those on the new Exit Survey.³

The Exit Survey would be housed in a DoD server and administered (unproctored) through an email link. If all personnel leaving active duty for a specific time frame were surveyed, the DoD should obtain a large sample of survey respondents, even if response rates are low.

Maximizing Usefulness

Challenges

Criterion data collection is expensive and time-consuming. It may be difficult to convince decisionmakers to support a laborious criterion data collection, if the only result is the evaluation of the criterion measures.

Possible Solutions

The best way to maximize the usefulness of the criterion data is to make the criterion evaluation part of a predictor validation project. Using the hard-earned data for predictor validation is likely to be much more palatable than collecting criterion data by itself.

The predictive validity of the ASVAB and any other predictors that have been administered pre-enlistment could be examined as long as the criterion data are collected in a manner that allows matching subjects' records (e.g., name and date of birth, DoD ID, or SSN).

Validation 101

Criterion-related validity is the gold standard for pre-accession assessments like the ASVAB and TAPAS. It means that scores on the assessment are significantly correlated with, or “predict,” a criterion of interest, such as performance ratings or attrition.

Validation studies can be **concurrent**, meaning the predictor and the criterion data are gathered at about the same time, or **predictive**, when the criterion data are collected well after the predictor data. Predictive validation studies are longitudinal.

³ The new exit survey includes a small core set of items derived from the services' operational exit surveys. We fully expect a high degree of concordance between the operational and new surveys.

If a Service has an experimental predictor which has not been administered pre-enlistment, that Service could add predictor administration to its data collection efforts. This would allow the Service to evaluate criterion-related validity using a concurrent validation research design. As noted before, Service-specific additions would have IRB implications.

The Services and possibly the Manpower Accession Policy Working Group (MAPWG) would need to think about the predictors needing validation and their expected relations to the criteria. For example, since the ASVAB is not intended to predict psychosocial well-being or counterproductive work behaviors, it would not be expected to show substantial relationships with criterion variables that measure those constructs. Non-cognitive predictors like the Tailored Adaptive Personality Assessment System (TAPAS) would be expected to show a relationship with those criterion variables. The focus of this research plan is administration and analysis of the criterion measures.

Streamlining Training and In-unit Data Collection

Challenges

Assuming the project will serve both predictor and criterion validation purposes, one challenge is to design the project to minimize (a) the impact on the Services' operations in terms of personnel time and resources and (b) the delay between administering measures and receiving results.

The end-of-training (EOT) and in-unit (IU) validation projects could either be part of one large longitudinal validation project or could be conducted as separate concurrent validation projects. In the longitudinal design all Services would track one cohort from accession to the end of the first term of enlistment. Individuals who participated in the EOT validation would be sought for inclusion in the IU validation. Longitudinal tracking is logistically difficult because the initial cohort must be tracked and retested later. In longitudinal validation, there is some loss in the sample (due to attrition or failure to track personnel). Initial sample sizes must be very large to ensure sufficient sample sizes in later data collections. Finally, it can take years to obtain results from longitudinal studies.

In a concurrent design, the EOT and IU projects would use separate samples, and the two projects could be conducted at the same time. The main advantage of the longitudinal design is that individual's scores on the EOT assessments (rating scales, self-report survey) could be compared to their scores on the IU assessments. However, we do not view this as critical and given the difficulties in conducting longitudinal data collection, we recommend a concurrent validation strategy for this project.

Possible Solutions

Adopt a concurrent data collection strategy.

Planning EOT Data Collection

Challenges

While the focus of the current project is on DoD-wide criterion measures (i.e., not on occupation-specific performance), data collected at the end of technical training must, by necessity, be collected in or from schoolhouses for specific occupations. Low-density occupations often have small classes and a low flow-rate through school, making it difficult to obtain sufficiently large samples of trainees in a reasonable period of time.

Possible Solutions

Prior research efforts have handled these issues in a variety of ways. First, selecting high density occupations for inclusion helps build the sample size quickly. Second, training schoolhouse staff to administer data collection instruments reduces the cost of travel and burden on the Services' personnel research organizations. Often schoolhouse staff are already overburdened. So, in some instances, contractors have been hired to work with schoolhouses to improve collection of final school grades or other measures.

Planning IU Data Collection

Challenges

While schoolhouses provide a natural structure and place for collecting data from trainees, research staff must gain access to in-unit personnel through the command structure. The performance rating scales require participation of peers and their supervisors. In the past, this has been handled by tasking intact units to provide all 1st term enlisted personnel and their supervisors.

Possible Solutions

Work with the command structure to task personnel and their supervisors, possibly from intact units for the data collection. The Army Research Institute for the Behavioral and Social Science (ARI) has a good deal of experience conducting research in this way and could provide some lessons learned for other Services.

III. Research Plan

The remainder of this document describes three proposed projects, listed in Table C.3. While predictor validation is likely to be added to some of these projects, this research plan focuses on the evaluation of the criterion measures.

Table C.3. Three Projects and Their Purposes

Attrition	To evaluate procedures for constructing attrition variables, the integrity of attrition data.
End-of-training (EOT)	To evaluate performance appraisal ratings and a self-assessment survey administered at the end of technical training.
In-Unit (IU)	To evaluate performance appraisal ratings, a situational judgment test, and a self-assessment survey administered to E-3 and E-4 service members having 18 to 36 months of experience.

Attrition Project

Overview

The Services will email a link to the Exit Survey to all enlisted personnel who have notified the Service of their intent to leave. If the Service routinely collects Exit Survey data, the Service could forego collecting new Exit Survey data provided that it meet the following two criteria: (1) the Service's Exit Survey items must overlap with the new joint-service Exit Survey, and (2) the Service must be able to provide identifiable survey data for this project, (i.e., identifiable data that can be matched to administrative records). Unidentifiable survey data would not be useable.

Survey responses will be stored in a database for several months while the administrative data mature. After most respondents are likely to have left, each Service will create a database comprising administrative data for those personnel. Research staff will construct attrition variables using procedures from the Criterion Measures Project and will construct analytic databases.⁴ Research staff will examine correlations between attrition variables constructed from the databases and reasons for leaving as indicated in the Exit Survey.

Assumptions

- The Services can identify personnel who intend to leave active duty Service in the next fiscal year.
- The Services will program and host the Exit Survey if current active duty personnel are surveyed. If it is determined that the survey will be administered to personnel who have exited and are now civilians, contractors or others may be able to host the survey.

⁴ In this document, "research staff" refers to either DoD personnel or civilian contractors.

Project Tasks

This project involves the following seven tasks:

1. **Obtain IRB Approval.** Research staff will work with appropriate Service POCs to complete IRB forms, prepare the Informed Consent / Privacy Statement, and finalize procedures for handling Personally Identifiable Information (PII).
2. **Automate the Exit Survey.** The Exit Survey will be automated such that it can be administered either on a smart phone, tablet, or on a desktop/laptop computer. Beta test versions will be evaluated in-house and tested on a small scale for each Service.
3. **Administer the Survey.** Links to the survey will be sent to the target population. Links will be embedded in an email from a high-ranking official explaining the purpose and importance of the survey. If the Service has met the two criteria for providing operational Exit Survey data, the Service would instead provide Exit Survey data for the target population.
4. **Assemble Attrition Database(s).** Each Service will construct a database of administrative records for the target population. The database must include identifiers that allow matching of personnel to Exit Survey data. Research staff will clean and screen databases and construct attrition variables using the procedures defined in the Criterion Measures Project.
5. **Analyze Data.** The procedures for constructing attrition variables will be evaluated by identifying problems or errors in trying to construct the variables. The integrity and validity of the attrition variables will be evaluated by comparing scores on attrition variables constructed from the databases to scores from the Exit Survey.
6. **Phone Survey.** To further assess the accuracy of data, research staff will flag cases where there is a mismatch between the reasons for leaving reported in administrative records and the reasons for leaving reported on the Exit Survey. A random sample of 50 cases will be drawn. Staff will follow-up with interview questions to identify reasons for the mismatch.
7. **Document Results.** The procedures and results will be documented in a technical report and briefed to appropriate audiences.

End-of-Training (EOT) Validation

Overview

Criterion data will be collected from selected schoolhouses. Participants will complete peer ratings and a self-assessment survey. Trainers may also be asked to provide performance ratings. Administrative data (e.g., training grades) will be collected for trainees. Data will be analyzed to evaluate the psychometric properties of the criterion measures and the latent structure of the criterion domain. It is important to note that the EOT validation study does not include the Situational Judgment Test (SJT) – a criterion measure developed only for experienced 1st term personnel.

Assumptions

- The Services will provide POCs for the data collection
- Measures will be administered by paper-and-pencil methods in schoolhouses.
- The Services will provide data collection facilities.
- Each Service will provide data for at least 300 trainees.

Project Tasks

This project involves the following seven tasks:

1. **Obtain IRB Approval.** Research staff will work with appropriate Service POCs to complete IRB forms, prepare the Informed Consent / Privacy Statement, and finalize procedures for handling PII.
2. **Finalize EOT Data Collection Procedures.** Schoolhouses and POCs will be identified and brought on board. Data collection staff (schoolhouse POCs, the Services' research staff, or contractors) will be identified. Manuals and checklists will be developed for POCs and research staff. Data collection staff will participate in data collection training.
3. **Prepare Measures.** Performance Rating Scales (PRS) for Peers and/or Trainers and the Self-Report Assessment will be printed on scannable forms and prepared in accordance with the final data collection procedures.
4. **Collect Data.** Performance appraisal ratings and the self-report assessment will be administered to trainees in proctored sessions. At the close of the face-to-face data collection, the Services will provide administrative data (i.e., training outcomes, physical fitness results) for the trainees who were tested.
5. **Analyze Data:**
 - a. **Develop database.** Merge administrative data with self-report assessment responses and performance appraisal ratings. Conduct data quality checks.
 - b. **Evaluate criterion measure psychometric properties.** Examine the internal consistency of the self-report survey and interrater agreement of peers and supervisors for the PRS. Estimate subgroup differences.
 - c. **Evaluate the construct validity of criterion measures.** Examine the intercorrelations between different assessments to assess construct validity. Correlations between similar constructs measured by different methods should converge. Conduct factor analyses to examine the latent structure of the criterion domain.
 - d. **Finalize criterion scoring.** Develop and evaluate composite scores, as appropriate.
6. **Conduct Attrition Follow-up.** Approximately two years after the main study, the research team will collect administrative data and construct attrition variables. Attrition results will be compared to the self-report assessment responses to determine the extent to which the self-report assessment data predicts attrition.

7. **Document Results.** The procedures and results will be documented in a technical report and briefed to appropriate audiences.

In-Unit (IU) Validation

Overview

Criterion data will be collected from in-unit Service members. Participants will complete peer ratings, an SJT, and a self-assessment survey. Supervisors will also provide performance ratings on participants. Administrative data (e.g., training grades, physical fitness scores, disciplinary actions) will be collected for participants. Data will be analyzed to evaluate the psychometric properties of the criterion measures and the latent structure of the criterion domain.

Assumptions

- The Services will provide POCs for the data collection
- Measures will be administered by paper-and-pencil methods.
- The Services will provide data collection facilities.
- Each Service will provide data for at least 300 first term personnel with approximately 18 months of experience.

Project Tasks

This project involves the following seven tasks:

1. **Obtain IRB Approval.** Research staff will work with appropriate Service POCs to complete IRB forms, prepare the Informed Consent / Privacy Statement, and finalize procedures for handling PII.
2. **Finalize EOT Data Collection Procedures.** POCs for selected organizational units will be identified and brought on board. Data collection staff (the Services' research staff, or contractors) will be identified. Manuals and checklists will be developed for POCs and research staff. Data collection staff will participate in data collection training.
3. **Prepare Measures.** Performance Rating Scales (PRS) for Peers and Supervisors, the SJT, and the Self-Report Assessment will be printed on scannable forms.
 - a. **Pilot test the SJT.** The overlength version of the SJT delivered at the close of the Criterion Measures Project will need to be pilot tested before the main validation study. We recommend administering the SJT to approximately 50 – 100 1st term personnel with approximately 18 months of experience in each Service (i.e., a total sample of 250 - 500 across Services). Pilot test data will be used to shorten and finalize the SJT.
 - b. **Develop the SJT key(s).** Research suggests that the best SJT keys are based on the responses of people who are highly proficient in the qualities that the SJT assesses. For this project we recommend collecting responses from at 25 – 50 senior non-commissioned officers (E8 – E9 NCOs) from each Service to serve as

the key. Basing the key on senior NCO judgments also enhances the credibility of the key. The reliability and other psychometric properties of keys constructed within and across Services will be evaluated to determine whether the key should be within or across Services.

- c. **Print forms.** Performance Rating Scales (PRS) for Peers and Supervisors, the SJT, and the Self-Report Assessment will be printed on scannable forms.
4. **Collect Data.** Performance appraisal ratings, the SJT and the self-report assessment will be administered in proctored sessions to 1st term personnel with approximately 18 months of experience. At the close of the face-to-face data collection, the Services will provide administrative data (i.e., training outcomes, physical fitness results) for the personnel who were tested.
5. **Analyze Data:**
 - a. **Develop database.** Merge administrative data with self-report assessment scores, SJT scores, and performance appraisal ratings. Conduct data quality checks.
 - b. **Evaluate criterion measure psychometric properties.** Examine the internal consistency of the self-report survey and the SJT. Examine interrater agreement of peers and supervisors for the PRS, where possible. Estimate subgroup differences.
 - c. **Evaluate the construct validity of criterion measures.** Examine the intercorrelations between different assessments to assess construct validity. Correlations between similar constructs measured by different methods should converge. Conduct factor analyses to examine the latent structure of the criterion domain.
 - d. **Finalize criterion scoring.** Develop and evaluate composite scores, as appropriate.
6. **Conduct Attrition Follow-up.** Approximately two years after the main study, the research team will collect administrative data and construct attrition variables. Attrition results will be compared to the self-report assessment responses to determine the extent to which the self-report assessment data predicts attrition.

Document Results. The procedures and results will be documented in a technical report and briefed to appropriate audiences