UVA-SDAD ARI PORTFOLIO IN-PROGRESS REVIEW

Social and Decision Analytics Division (UVA-SDAD)
Army Research Institute for the Behavioral and Social Sciences (ARI)

Quarterly Project Meeting (Q4) and IPR

26 October 2021

UNIVERSITY of VIRGINIA

BIOCOMPLEXITY INSTITUTE

Agenda

ARI Updates – Kelly, Andy, and Chuck

Portfolio Overview – UVA Team

- Project 1: Leveraging Archival Data to predict Individual and Team Performance
- Project 2: Developing Predictive Models of U.S. Army Career Pathways

Open discussion – Everyone

ARI PORTFOLIO OVERVIEW

UNIVERSITY of VIRGINIA

BIOCOMPLEXITY INSTITUTE

Transdisciplinary UVA Team

Statisticians and Mathematicians

Josh Goldstein
Statistician

Aritra Halder
Statistician

Sallie Keller
Statistician

Eric Oh
Statistician

Vicki Lancaster
Statistician

Social and Behavioral Scientists

Nathaniel Ratcliff
Social Psychologist

Joanna Schroeder
Government

Stephanie Shipp
Economist

Joel Thurston
Social Psychologist

Information Technology & Project Management

Aaron Schroeder
Information Integration

Jim Walke
Project Manager

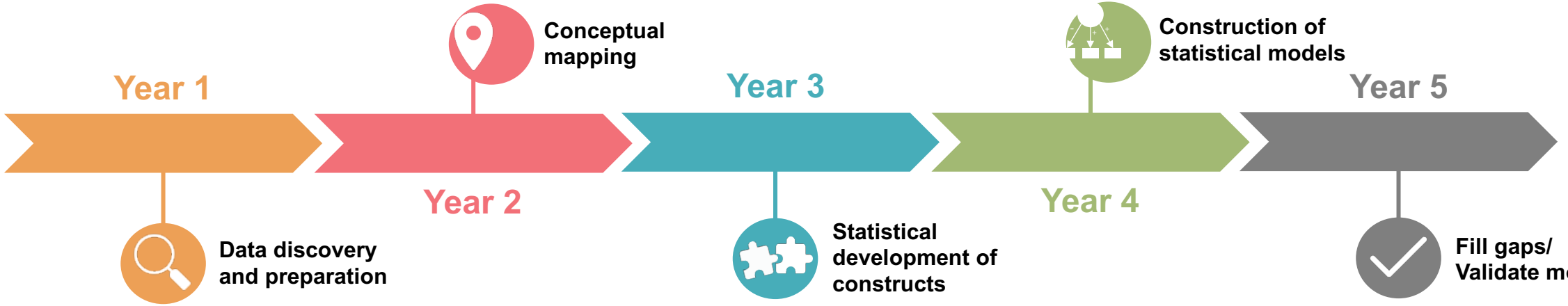
Leveraging Archival Data to Predict Performance

Problem: The Army possesses a trove of administrative data (e.g., personnel records, training scores), but has yet to fully leverage these data.

Purpose: Using modern data science techniques, we are developing models that integrate existing DOD data to make predictions about Soldier behavior and performance.

Payoff: Knowledge about how best to utilize data from disparate sources to form a holistic picture of Soldier and unit performance that can be used to:

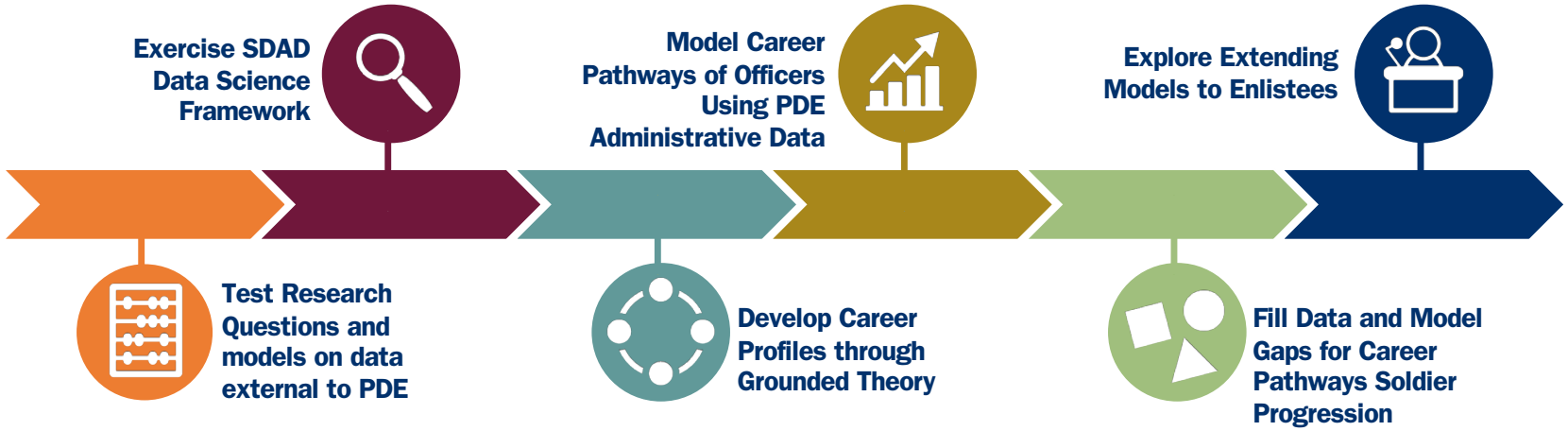
- Improve training
- Identify informative performance metrics
- Optimize talent management decisions across Soldier lifecycle



Developing Predictive Models of U.S. Army Career Pathways

Overview: ARI, in collaboration with the Biocomplexity Institute's Social & Decision Analytics Division of the University of Virginia, is developing statistical models of Soldier career progression based on existing Army administrative and survey data and other non-DOD data sources by:

- Leveraging expertise developed within the Person Data Environment (PDE);
- Developing longitudinal characterization of a Soldiers' knowledge, skills, and abilities acquisition through incorporation of individual and contextual information.

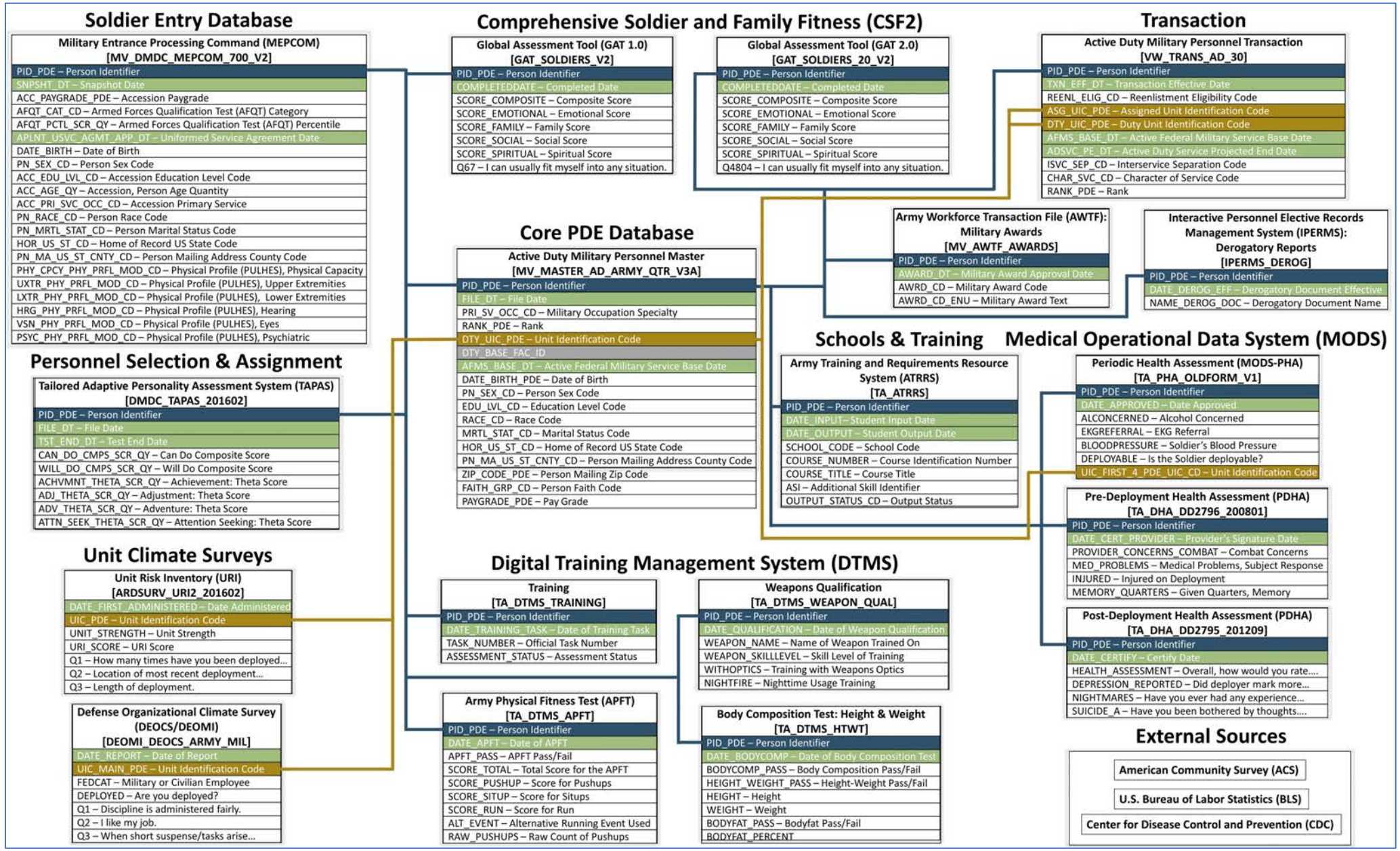


Person-Event Data Environment (PDE)

- Data enclave allows researchers to remotely access data
  - Maintained by Army Analytics Group Research Facilitation Lab (AAG-RFL)
  - Applied for and obtained Common Access Cards (CACs)
  - Registered to work in the PDE environment
  - Requested access to multiple data sources in the PDE
  - Regular meetings with RFL about software updates, PDE updates, and data inquiries

- Data Sources
  - Army administrative data sources (e.g., demographics, training history, accessions, and attrition data)
  - Adding non-DOD data sources (e.g., American Community Survey, Quarterly Census of Employment and Wages)
  - Archiving processes to ensure reproducibility

Data Source Map and Linkages



Gold Standard of Soldier Characteristics

Basic demographic variables between data sources needed to be reconciled and accessed for optimal data quality and usage.

Column Name	Description	Original Table
PID_PDE	Enlistee's Unique ID	Master
PN_SEX_CD	Gender	Master
RACE_CD	Race Code	Master
INIT_ENT_TRN_END_DT	Initial Entry Training End Date	Master
DATE_BIRTH_PDE	Person Birth Date	Master
PN_BIRTH_PLC_CTRY_CD	Person Birth Place Country Code	Master
HOR_ZIP_CODE_PDE	Home of Record Zip Code	Analyst
ACT_SCORE	ACT Score	Analyst
SAT_SCORE	SAT Score	Analyst
AP	ASVAB: Auditory Perception Score	Analyst
CO	ASVAB: Combat Score	Analyst
.	.	.
.	.	.
.	.	.

Demographic Schema

Data Source (Raw Variable)	PDE Variable Name	Description	Type	Levels	Model
Master (PID_PDE)	PID_PDE	Person identifier within PDE.	Alphanumeric	533,763: Numerous (e.g., PDERE706BESV)	NA
Master (PN_SEX_CD)	PN_SEX_CD	Soldier sex.	Categorical	2: Male; Female	Predictor/Moderator
Master (RACE_CD)	RACE_CD_RE	Soldier race.	Categorical	6: White; Black; Asian; AI/AN; NH/PI; Mixed Race/Other	Predictor/Moderator
MEPCOM 2 (PN_ETH_AFF_POST_2003_CD)	ETHNICITY	Soldier ethnicity.	Categorical	2: Hispanic of Latino; Non-Hispanic or Latino	Predictor/Moderator
Master (DATE_BIRTH_PDE), MEPCOM 1 (DATE_BIRTH), MEPCOM 2 (DATE_BIRTH)	DATE_BIRTH.CB	Soldier date of birth.	Date	Numerous	TBD
Master (MRTL_STAT_CD)	MRTL_STAT_FIRST.CB	Marital status at first record.	Categorical	8: Married; Never Married; Divorced; Legally separated; Widow(er); Annulled; Interlocutory decree; Unknown	Predictor/Moderator
Master (EDU_LVL_CD)	EDU_LVL_RD_FIRST	Highest education level of Soldier at first record.	Categorical	6: Less than high school; High school diploma, GED, or some college; Associate degree; Bachelor degree; Advanced degree	Predictor/Moderator
Family (CHLDRN_QY)	CHLDRN_QY_FIRST	Number of dependent children of Soldier at first record.	Numeric	(min) 0–ℝ (max)	Predictor/Moderator
Master (HOR_US_ST_CD), MEPCOM 1 (STATE), MEPCOM 2 (PN_MA_US_ST_CD)	HOR_STATE.CB	Home state of Soldier at accession.	Categorical	65: AE; AK; AL; AP; AR; AS; AZ; CA; CO; CT; DC; DE; FL; FM; GA; GU; HI; IA; ID; IL; IN; JA; KS; KY; LA; LE; MA; MD; ME; MH; MI; MN; MO ; MP; MS; MT; NC; ND; NE; NH; NJ; NM; NV; NY; OH; OK OR; OT; PA; PR; PW; RI; SC; SD; TN; TX; UT; VA; VI; WA; WI; WV; WW; WY	Predictor/Moderator

Data Codebooks (raw variables and derived variables for linked datasets)

Leveraging Our ARI Projects

Economies of scale using shared PDE resources:

- RFL allows us to treat ARI work as one big project
- PDE Data acquisition
- Contextual understanding of literature and data
- Conceptual variable profiling
- Data profiling
- Data quality checks
- Identification and creation of data codebooks
- Data linkage
- Data exploration
- Deriving key variables of interest (e.g., first-term attrition)

LEVERAGING ARCHIVAL DATA TO PREDICT INDIVIDUAL AND TEAM PERFORMANCE

UNIVERSITY of VIRGINIA

BIOCOMPLEXITY INSTITUTE

Outline

- Project Activities to Date
  - Literature Review and Conceptual Performance Model
  - Conceptual Profiling of Data Sources
  - Longitudinal Stability of the Global Assessment Tool (GAT)
  - 20-Year Data Benchmarking
  - Phase 1 Modeling: Simple Relationships
- Current /Next Steps
  - Phase 2 Modeling: Complex Relationships

Performance in the Army

- Studying Army performance is challenging
- Task-focused performance metrics do not always capture the social component of performance
- Want to expand performance criteria beyond task accomplishment
- **Premise**: Administrative Data Repositories may offer new opportunities to capture Soldiers' social and performance characteristics

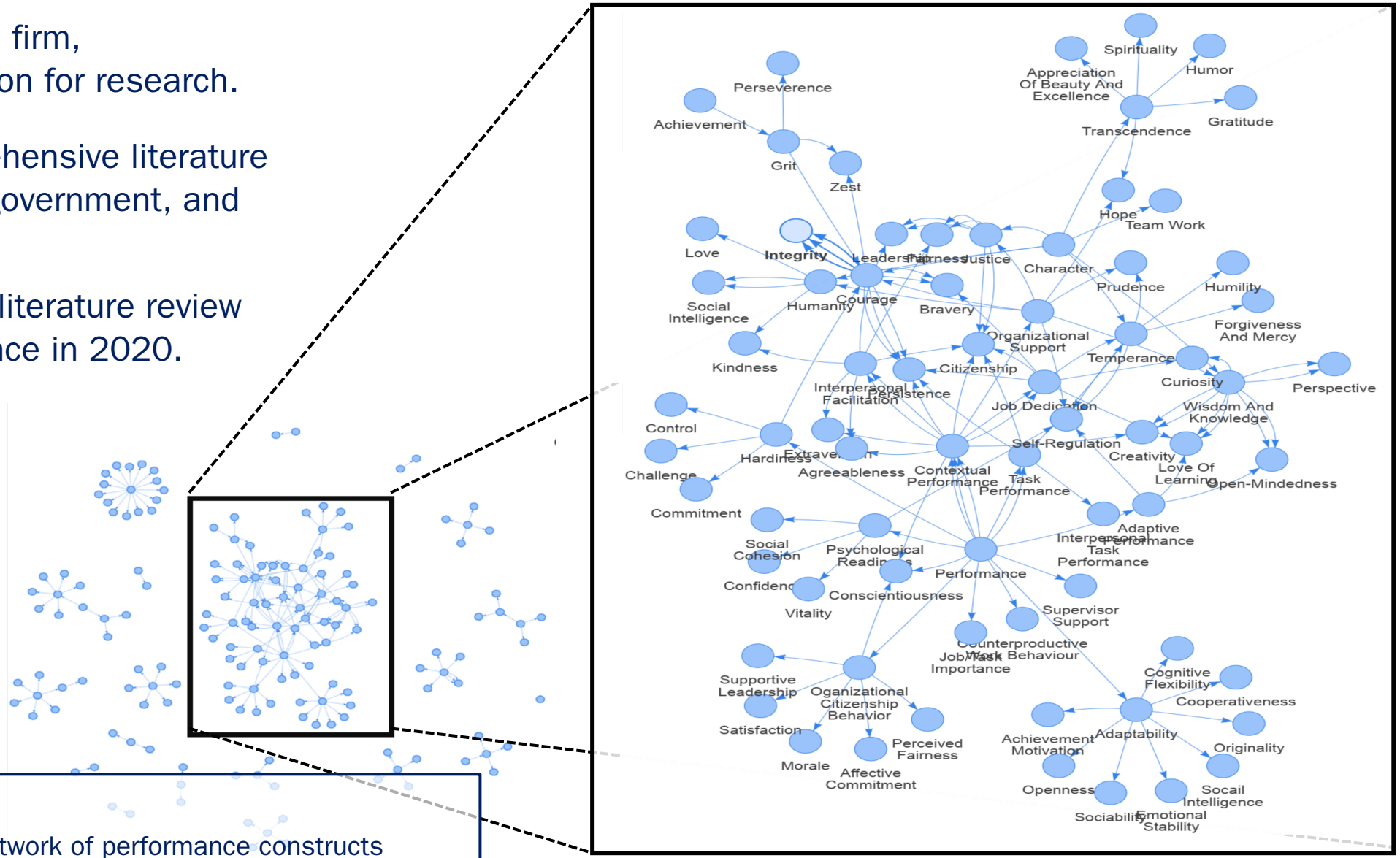
Official U.S. Army Flickr: https://commons.wikimedia.org/wiki/File:Flickr_-_The_U.S._Army_-_Expert_Field_Medical_Badge_testing.jpg

Literature Review

Problem: Need to obtain a firm, empirically-based foundation for research.

How: Conducted a comprehensive literature review across academic, government, and DOD sources.

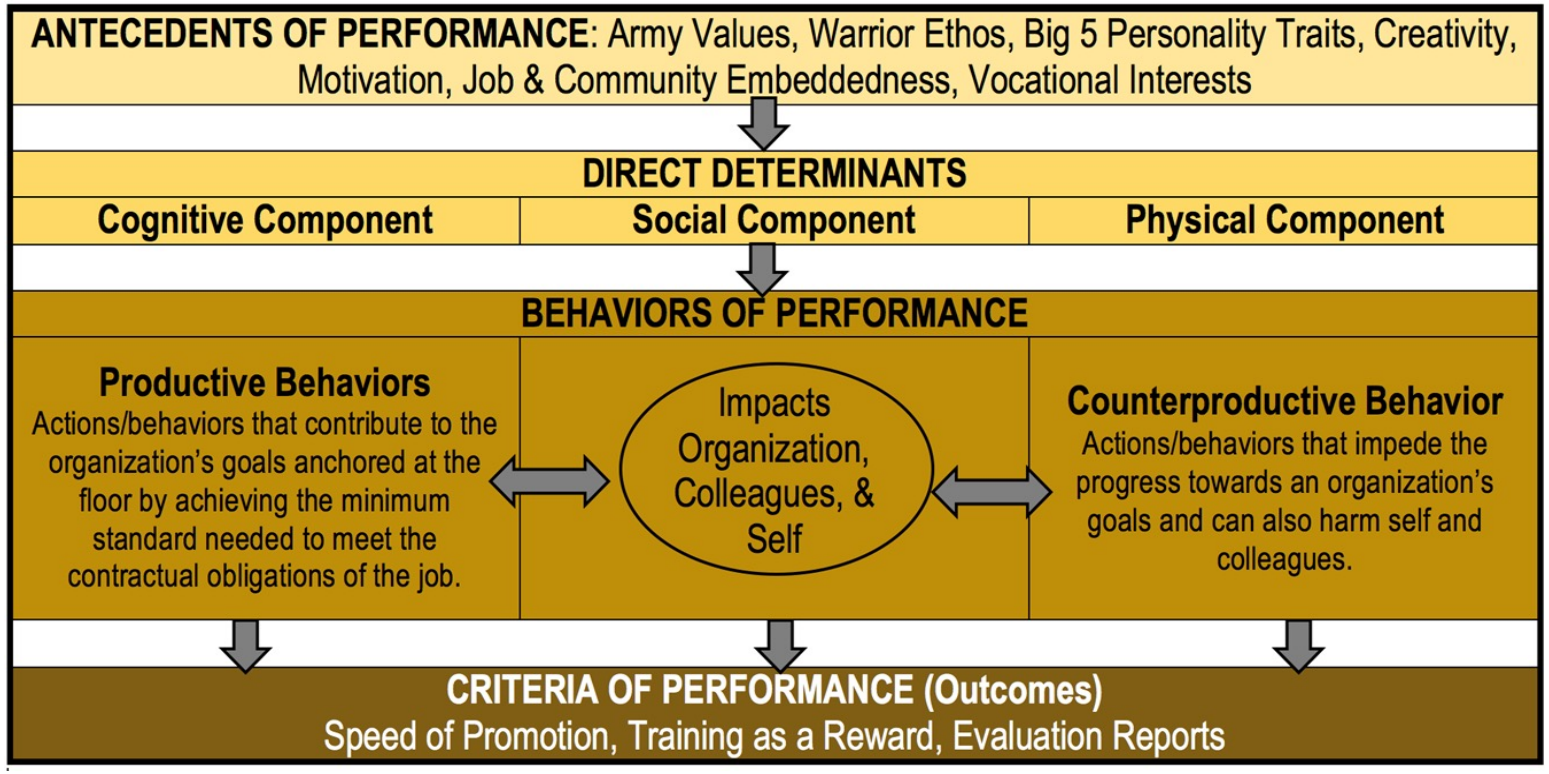
- Updated previous literature review on work performance in 2020.



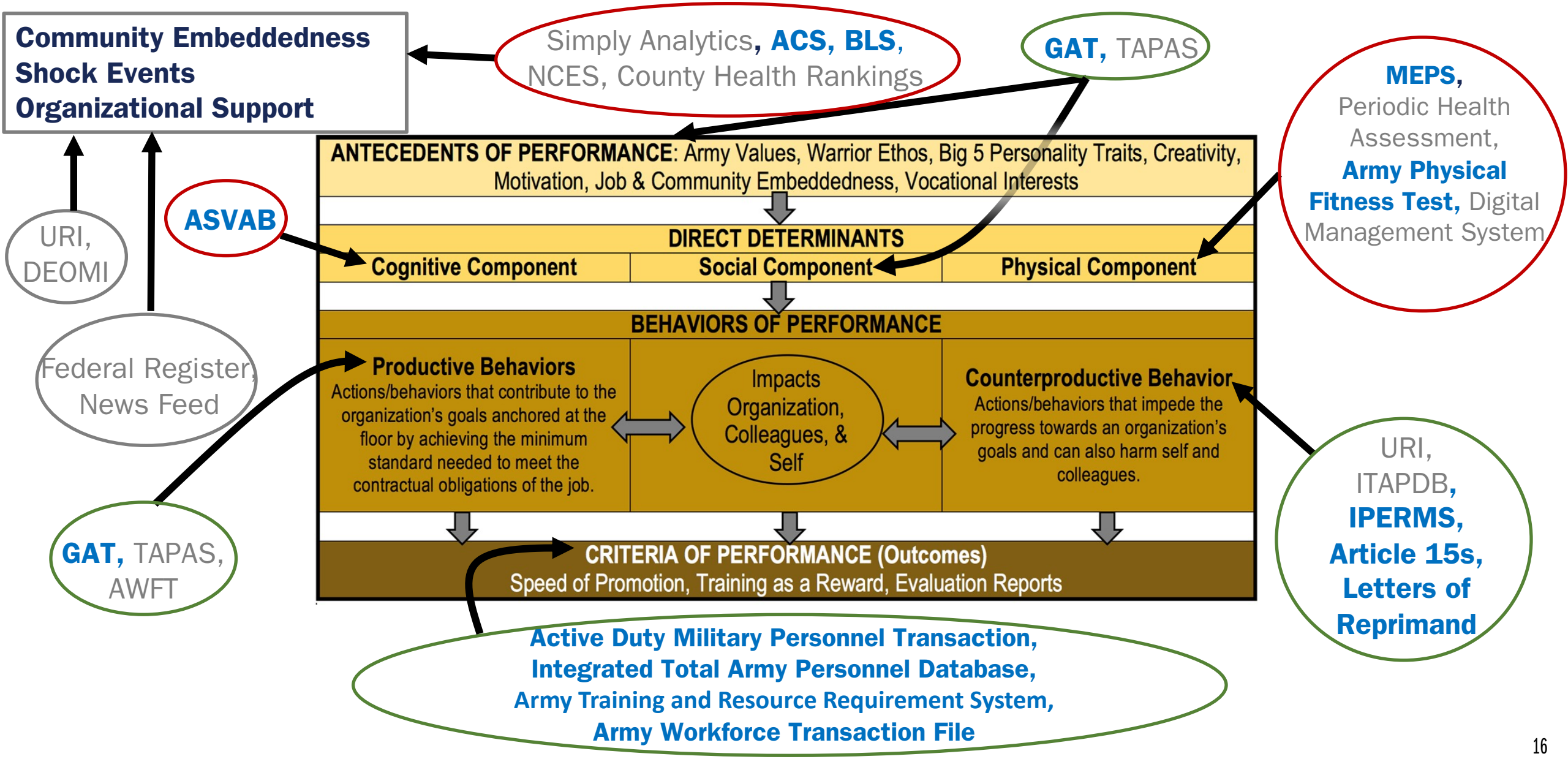
Products & Transitions:

- Interactive nomological network of performance constructs
- History of performance research in the U.S. Army (in preparation)

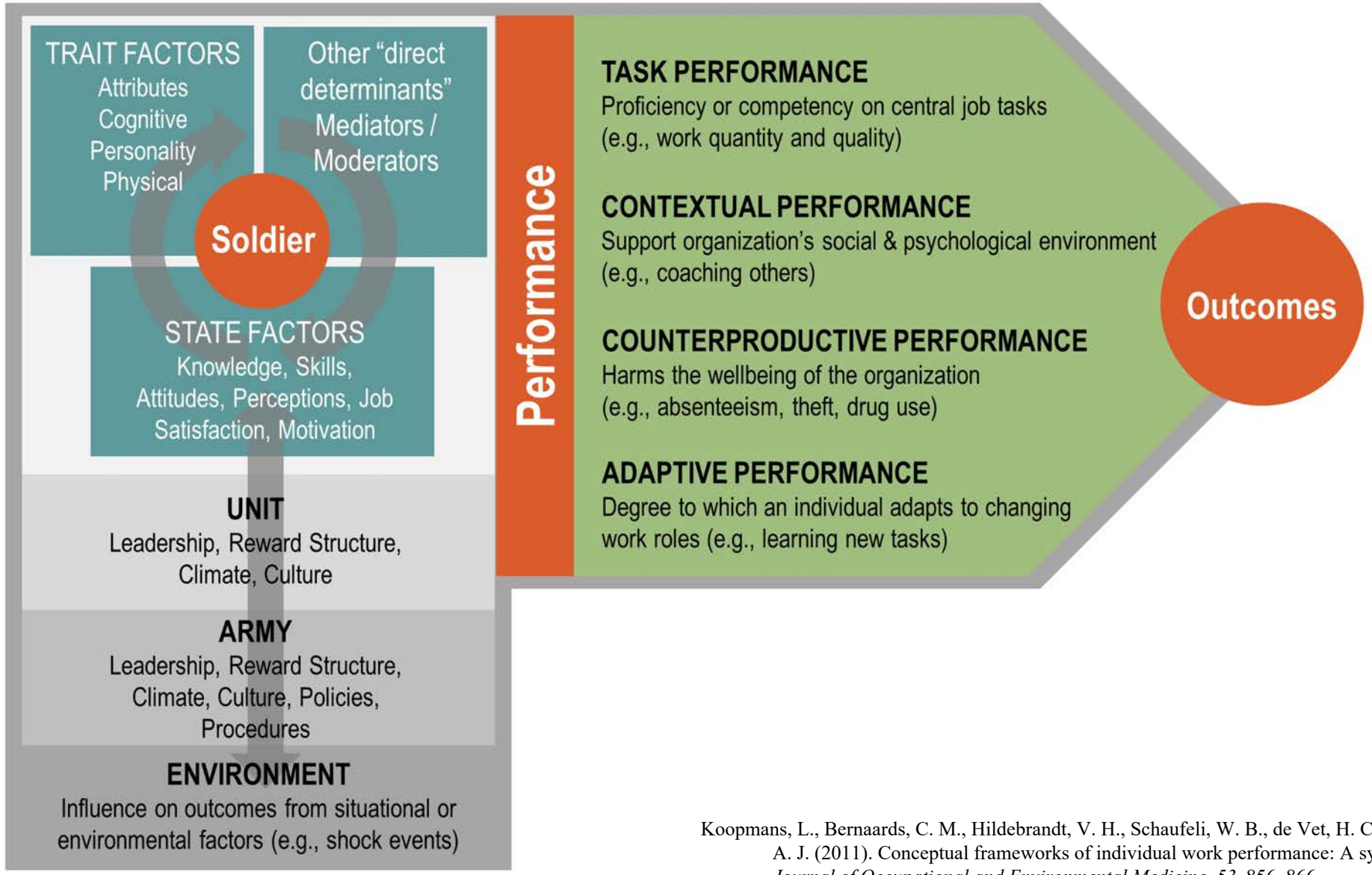
Conceptual Performance Model (Original)



Conceptual Performance Model (Original)



Conceptual Performance Model



Koopmans, L., Bernaards, C. M., Hildebrandt, V. H., Schaufeli, W. B., de Vet, H. C. W., & van der Beek, A. J. (2011). Conceptual frameworks of individual work performance: A systematic review. *Journal of Occupational and Environmental Medicine*, *53*, 856–866. https://doi.org/10.1097/JOM.0b013e318226a763

Conceptual Profiling

Problem: There are a vast number of variables available to us in the PDE, but their meaning and utility is unclear

How: Using conceptual and methodological profiling, conceptually identify and categorize variables to generate informative meta-data for data sources

Methods: Conceptual profiling and methodological profiling processes

Payoff: This effort provides deeper understanding into the available PDE data sources for use in identifying useful variables to include in modeling

Products & Transitions:

- Conceptual and Methodological Profiling spreadsheet with detailed meta-data on the variables and tables available to us in the PDE
- Methodology paper on conceptual profiling for peer review (in preparation)

Conceptual Profiling

1	[PDE] ENT_NAME	[PDE]	Name	Data Type	Construct	Instru	Measure	Measure	Mapping to	Performanc	Item Stem	Item Text	Operational	Response	Response Values	Reve	Citation
6	GAT_SOLDIERS_V2	FLAG_CO	GAT 1.0	administra	consent	na	individual	attribute	trait	na	na	na		dichotomo	0 = no; 1 = yes	na	
7	GAT_SOLDIERS_V2	GENDER	GAT 1.0	administra	soldier sex	na	individual	attribute	trait	na	na	na	Reported gender	dichotomo	1 = male; 2 = female	na	
8	GAT_SOLDIERS_V2	PID_PDE	GAT 1.0	administra	personal	na	individual	attribute	trait	na	na	na	Personal identifier	free	text	na	
9	GAT_SOLDIERS_V2	Q10	GAT 1.0	designed	family	Q10	individual	perceptual	state	na	During the	How	Assesses overall	likert	1 = not at all	no	Peterson &
10	GAT_SOLDIERS_V2	Q100	GAT 1.0	designed	work engagement	Q100	individual	perceptual	state;	contextual	How well	My work is	Assesses feeling	likert	23 = 1 = Not like me	no	Wrzesniewsk
11	GAT_SOLDIERS_V2	Q103	GAT 1.0	designed	work engagement	Q103	individual	perceptual	state;	contextual	How well	I would	Assesses feeling	likert	23 = 1 = Not like me	no	Wrzesniewsk
12	GAT_SOLDIERS_V2	Q104	GAT 1.0	designed	work engagement	Q104	individual	perceptual	state;	contextual	How well	I am	Assesses feeling	likert	23 = 1 = Not like me	no	Wrzesniewsk
13	GAT_SOLDIERS_V2	Q106	GAT 1.0	designed	work engagement	Q106	individual	perceptual	state;	contextual	How well	How I do in	Assesses feeling	likert	23 = 1 = Not like me	no	Wrzesniewsk
14	GAT_SOLDIERS_V2	Q113	GAT 1.0	designed	organizational	Q113	individual	perceptual	situational	na	Please	I trust my	Assesses three	likert	35 = 1 = Strongly	no	Mayer,
15	GAT_SOLDIERS_V2	Q115	GAT 1.0	designed	organizational	Q115	individual	perceptual	situational	na	Please	I think we	Assesses three	likert	35 = 1 = Strongly	no	Mayer,
16	GAT_SOLDIERS_V2	Q117	GAT 1.0	designed	organizational	Q117	individual	perceptual	situational	na	Please	My leaders	Assesses three	likert	35 = 1 = Strongly	no	Mayer,
17	GAT_SOLDIERS_V2	Q119	GAT 1.0	designed	organizational	Q119	individual	perceptual	situational	na	Please	My	Assesses three	likert	35 = 1 = Strongly	no	Mayer,
18	GAT_SOLDIERS_V2	Q124	GAT 1.0	designed	organizational	Q124	individual	perceptual	situational	na	Please	Overall, I	Assesses three	likert	35 = 1 = Strongly	no	Mayer,
19	GAT_SOLDIERS_V2	Q125	GAT 1.0	designed	friendship	Q125	individual	perceptual	state	na	na	How many	Assesses strength	likert	1 = none; 5 = 4 or	no	
20	GAT_SOLDIERS_V2	Q128	GAT 1.0	designed	friendship	Q128	individual	perceptual	state	na	na	I have a best	Assesses strength	dichotomo	0 = no; 1 = yes	no	
21	GAT_SOLDIERS_V2	Q131	GAT 1.0	designed	family closeness	Q131	individual	perceptual	situational	na	na	I am very	Assess close ties	dichotomo	7 = 1 = No -> -0.5;	no	
22	GAT_SOLDIERS_V2	Q132	GAT 1.0	designed	friendship	Q132	individual	perceptual	state	na	na	I have	Assesses strength	dichotomo	0 = no; 1 = yes	no	

- Over 3,500 individual variables from the PDE underwent conceptual and methodological profiling
- After training, team applied their judgments to determine indices of agreement on a random subset of variables

Conceptual Performance Model (CPM)

Master: Demographics (age, sex, race)
MEPCOM: AFQT Score, Home Location, Education
TAPAS: Personality Metrics
GAT: Psychosocial Metrics

Master: Duty Location, Marital Status, Education
GAT: Psychosocial Metrics
PHA: PULHES Score
ATMS: Training, Weapon Qual.
ATRRS: Training, Coursework
FAM SPON: Dependents
CTS: Deployment History

DEOCS: Climate Survey
URI: Unit Risk Inventory

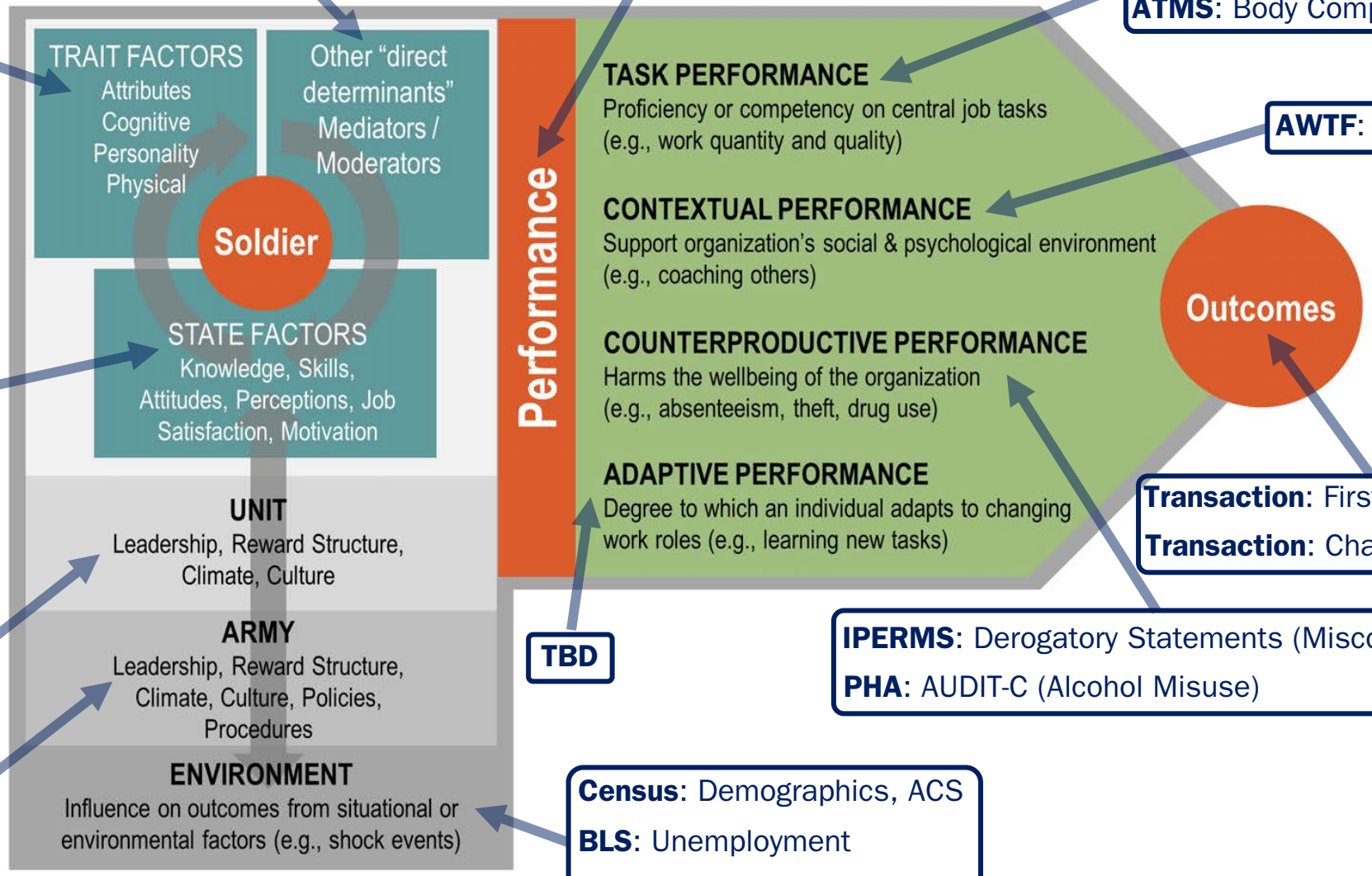
DEOCS: Climate Survey
Doctrine: FMs, ADRPs

Master: Rank Group, MOS

Master/Transaction: Speed of Promotion

ATMS: APFT Scores
ATMS: Body Composition

AWTF: Awards (Merit)



TBD

Transaction: First-Term Attrition
Transaction: Character of Service

IPERMS: Derogatory Statements (Misconduct)
PHA: AUDIT-C (Alcohol Misuse)

Census: Demographics, ACS
BLS: Unemployment
CDC: BMI

Conceptual Performance Model (CPM)

Master: Demographics (age, sex, race)
MEPCOM: AFQT Score, Home Location, Education
TAPAS: Personality Metrics
GAT: Psychosocial Metrics

Master: Duty Location, Marital Status, Education
GAT: Psychosocial Metrics
PHA: PULHES Score
ATMS: Training, Weapon Qual.
ATRRS: Training, Coursework
FAM SPON: Dependents
CTS: Deployment History

DEOCS: Climate Survey
URI: Unit Risk Inventory

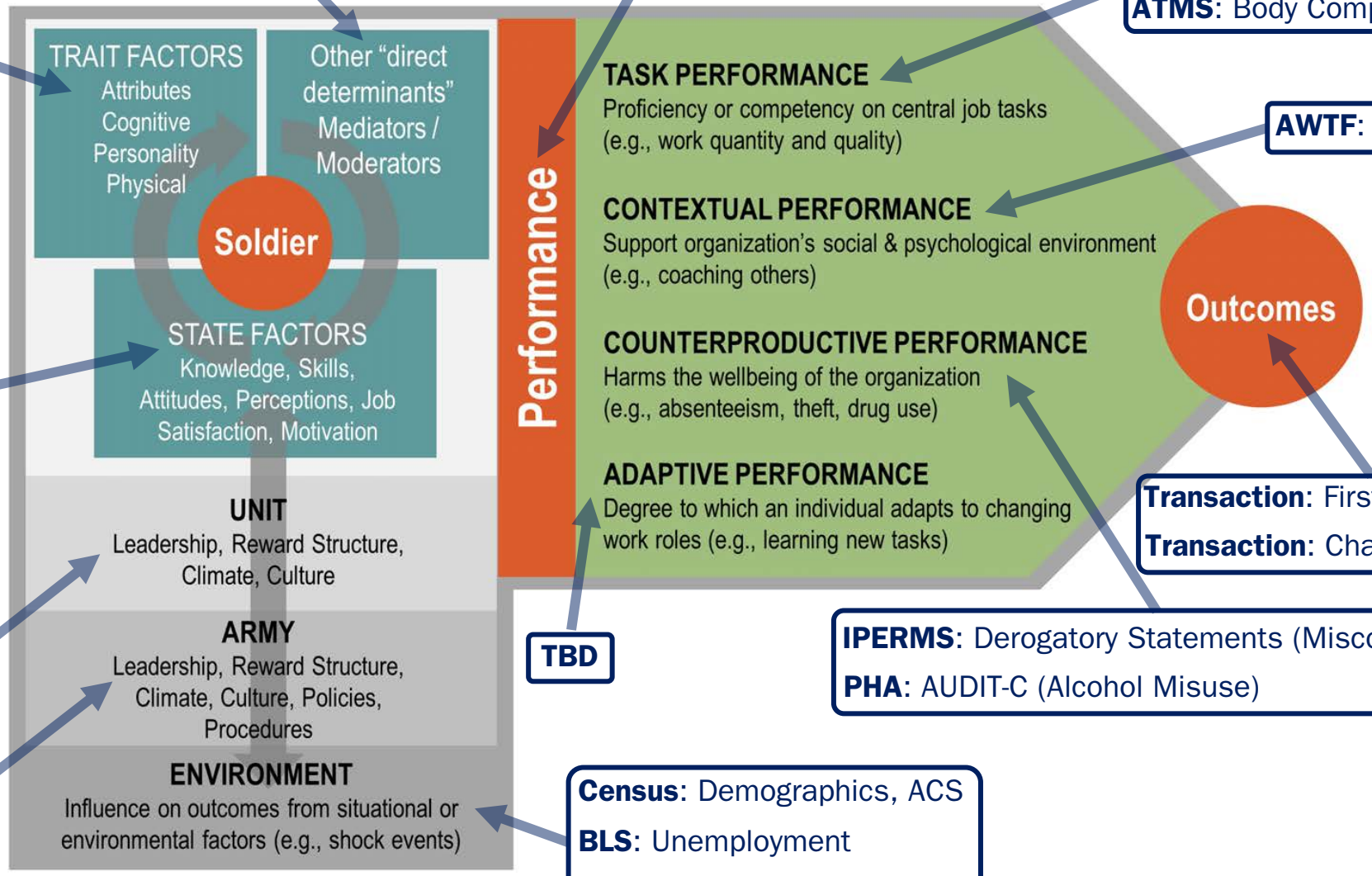
DEOCS: Climate Survey
Doctrine: FMs, ADRPs

Master: Rank Group, MOS

Master/Transaction: Speed of Promotion

ATMS: APFT Scores
ATMS: Body Composition

AWTF: Awards (Merit)



TBD

Transaction: First-Term Attrition
Transaction: Character of Service

IPERMS: Derogatory Statements (Misconduct)
PHA: AUDIT-C (Alcohol Misuse)

Census: Demographics, ACS
BLS: Unemployment
CDC: BMI

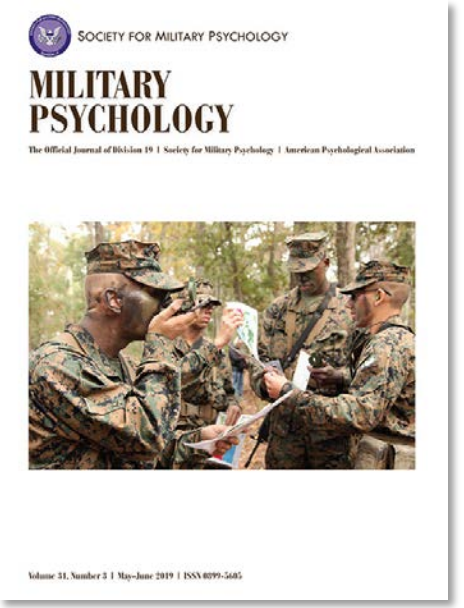
Longitudinal Stability of the GAT

Problem: The Global Assessment Tool (GAT) is measured annually, but the longitudinal psychometrics of the tool are undocumented

How: Examine whether GAT measures change over time at the population-level (e.g., mean-level changes, rank-order changes) and individual-level (e.g., reliable change).

Methods: t-tests, RM-ANOVA, RM-MLM, RM-SEM, reliable change index

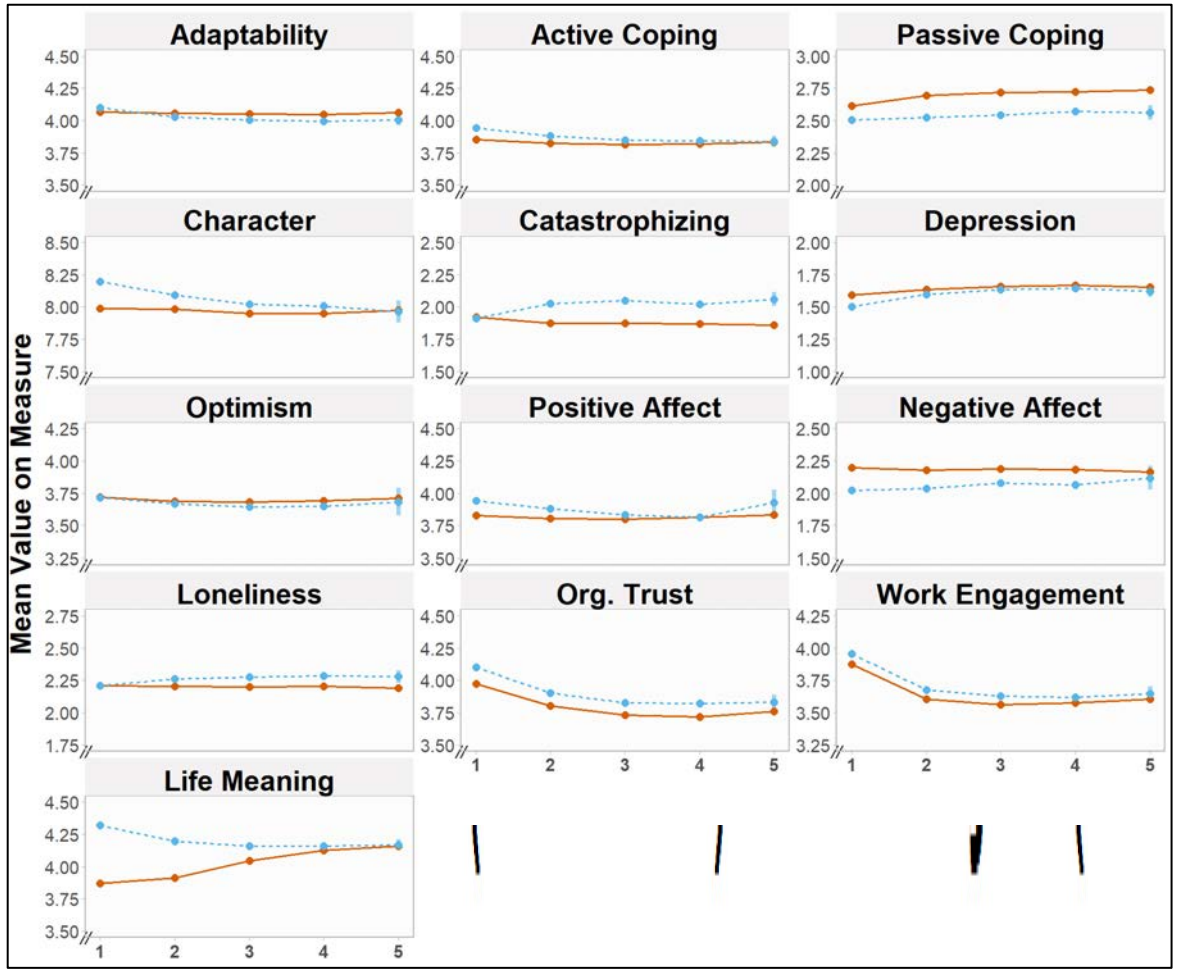
Payoff: This research helps inform the psychometric characteristic of the GAT measures and how to integrate them into performance models.



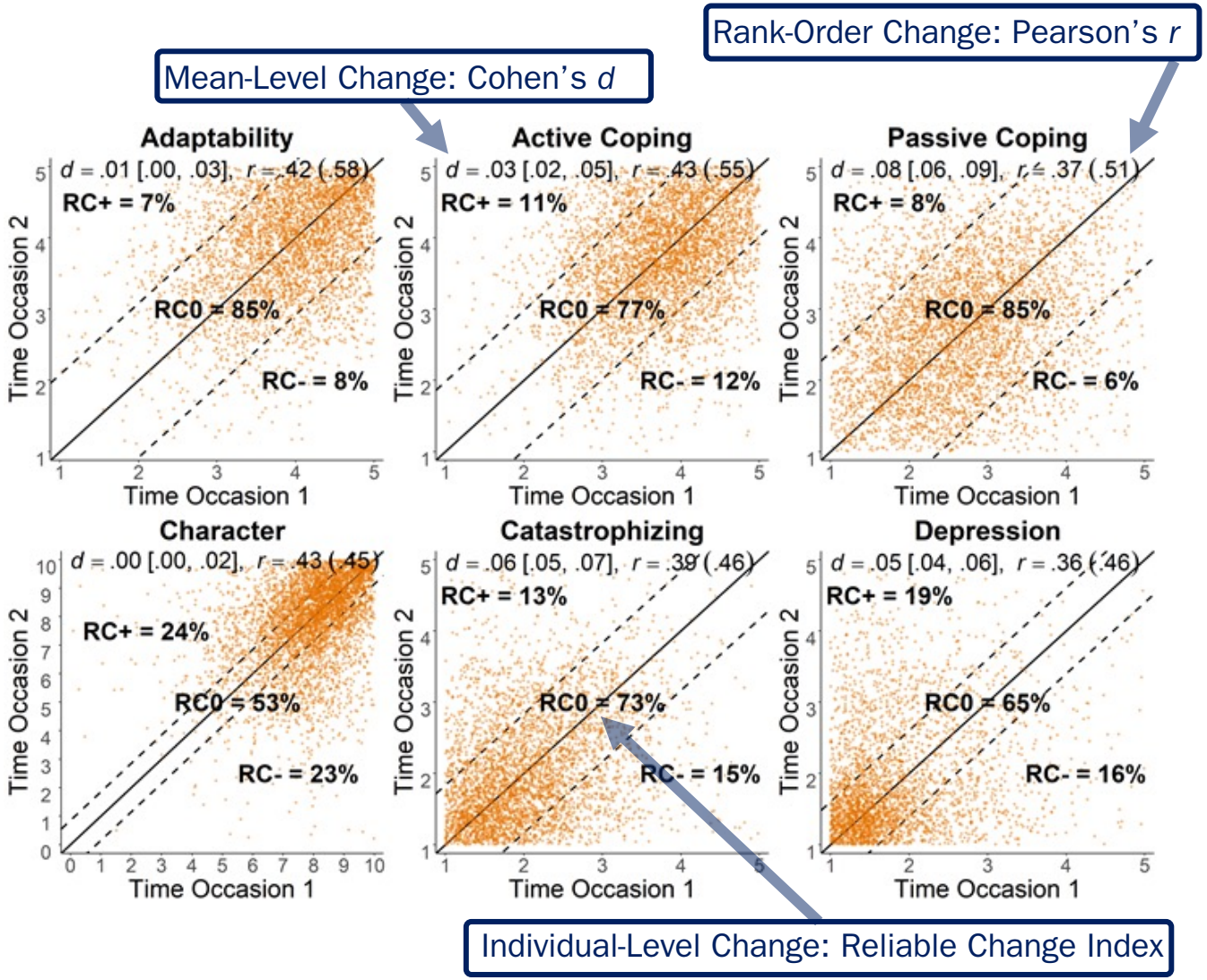
Products & Transitions:

- Comprehensive GAT ARI Tech Report (in DTIC: AD1147839, ARI TR 1397)
- Extracted analysis of the GAT (in press at *Military Psychology*)
- Draft of *Mil Psych* paper provided to the Army Resilience Directorate (ARD)
- Participation in Azimuth Check Steering Committee for redesign

GAT Stability – Selected Results

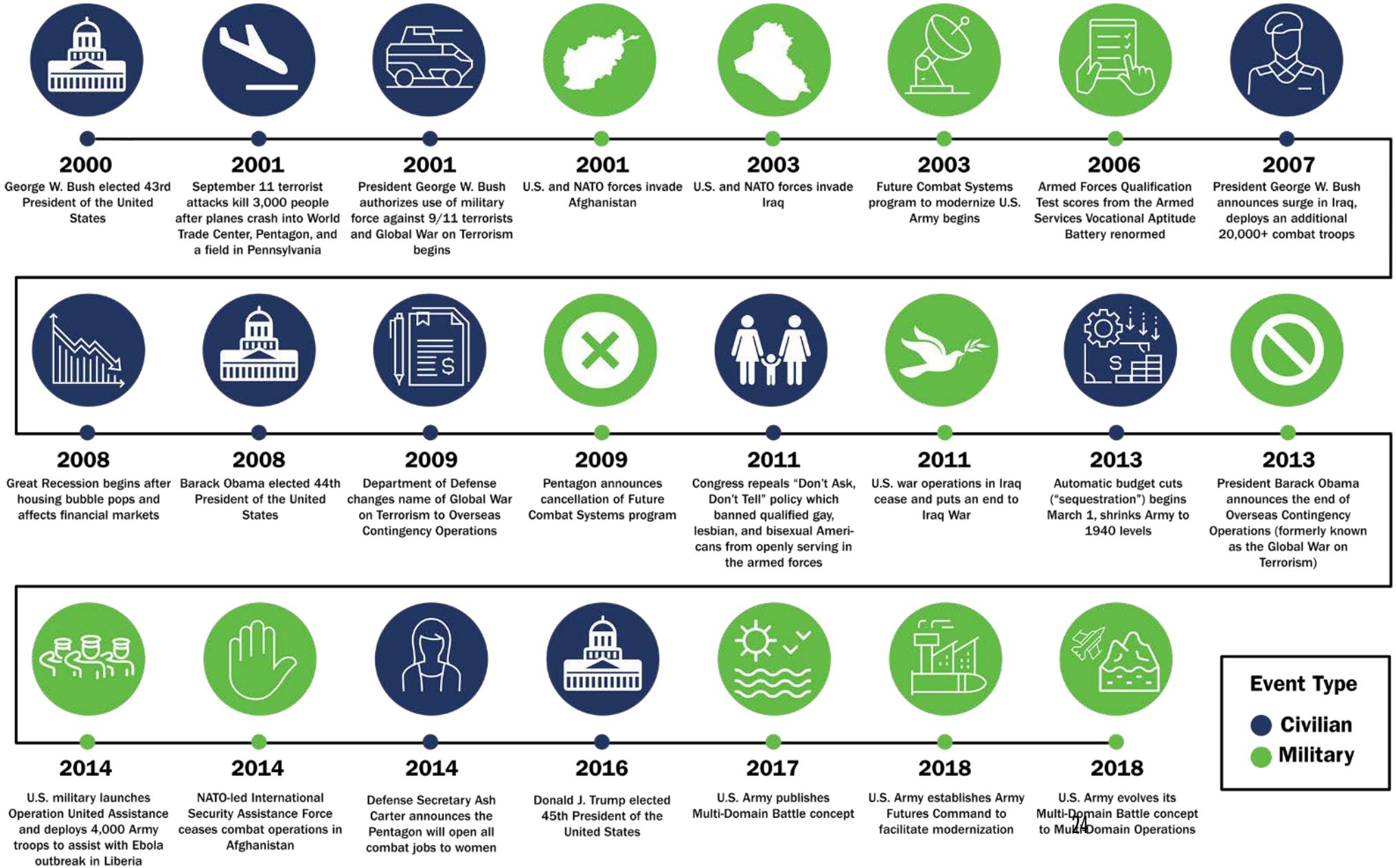


Measures mostly stable over time except for organizational trust, work engagement, and life meaning



Measures predominately stable across different indicators of longitudinal change

Timeline of Major Events Relevant to the U.S. Army from 2000-2019



20-Year Benchmark of the U.S. Army

Problem: Data are available for a 20-year time span, but it is not fully understood what trends occur over time and their potential broad impacts on changes in performance.

How: Benchmarking accession trends for U.S. Army Soldiers from 2000 to 2019.

Focus Areas:

Soldier Demographics and Home Location

Recruit Quality (Education & Armed Forces Qualification Test (AFQT) Scores)

Socioeconomic status

Soldier Health & Physical Fitness

Payoff: This research provides a quality check of our data with data reported publicly by the DOD and other DOD-affiliated research groups

Conceptual Performance Model (CPM)

Master: **Demographics (age, sex, race)**
MEPCOM: **AFQT Score**, Home Location, **Education**
TAPAS: Personality Metrics
GAT: Psychosocial Metrics

Master: Duty Location, Marital Status, **Education**
GAT: Psychosocial Metrics
PHA: PULHES Score
ATMS: Training, Weapon Qual.
ATRRS: Training, Coursework
FAM SPON: Dependents
CTS: Deployment History

DEOCS: Climate Survey
URI: Unit Risk Inventory

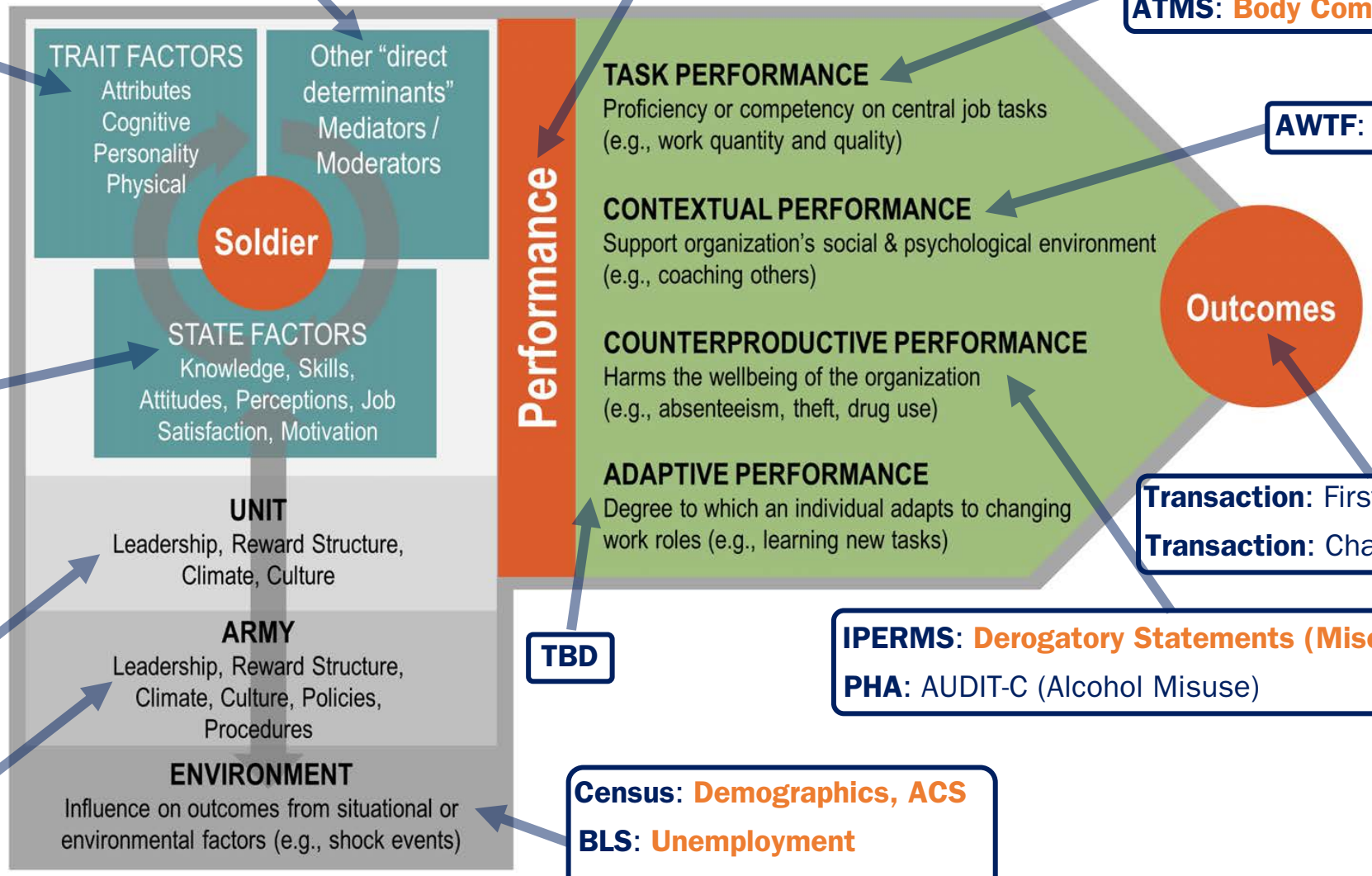
DEOCS: Climate Survey
Doctrine: FMs, ADRPs

Master: **Rank Group**, **MOS**

Master/Transaction: Speed of Promotion

ATMS: **APFT Scores**
ATMS: **Body Composition**

AWTF: Awards (Merit)



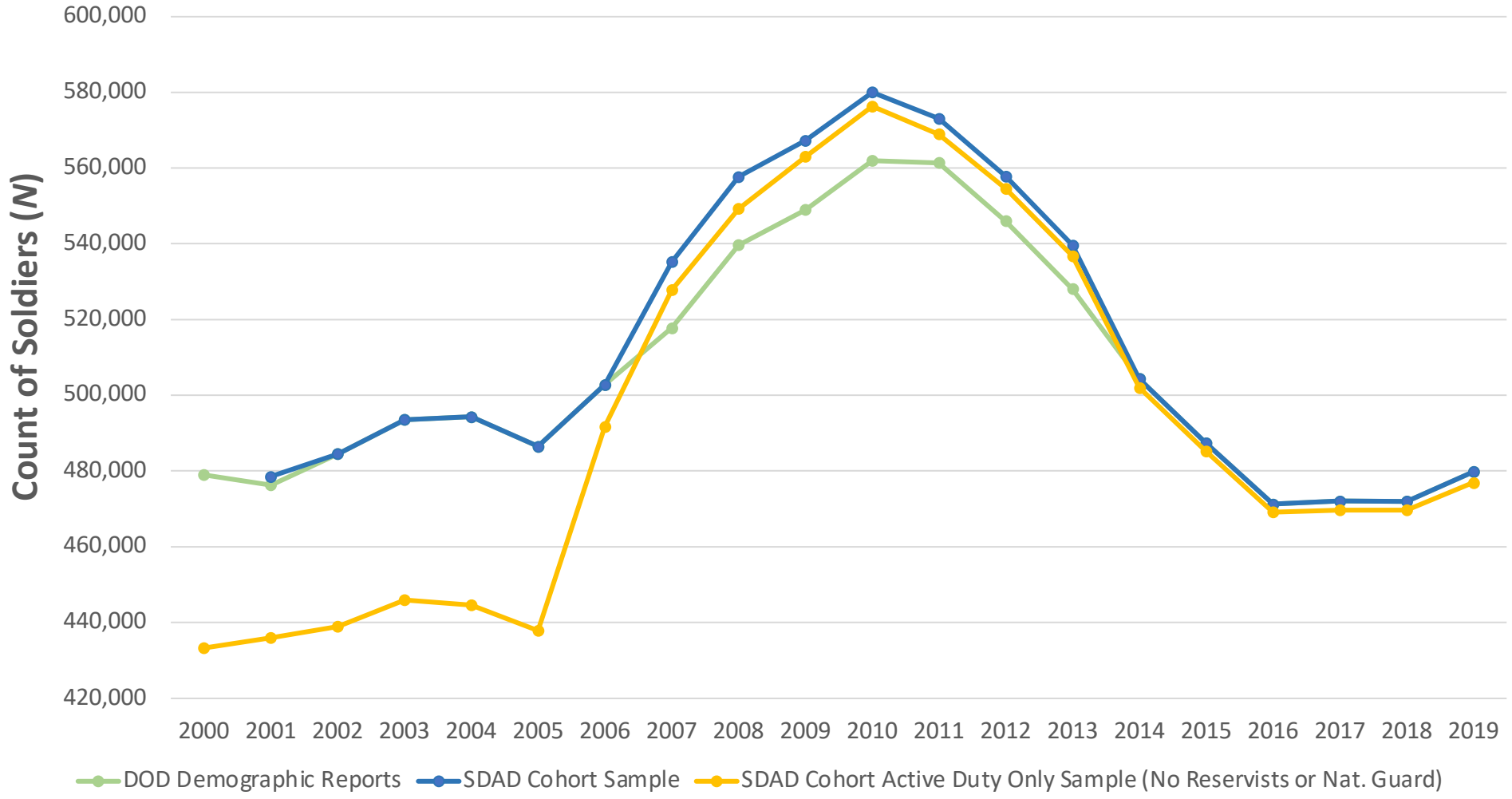
TBD

Transaction: First-Term Attrition
Transaction: Character of Service

IPERMS: **Derogatory Statements (Misconduct)**
PHA: AUDIT-C (Alcohol Misuse)

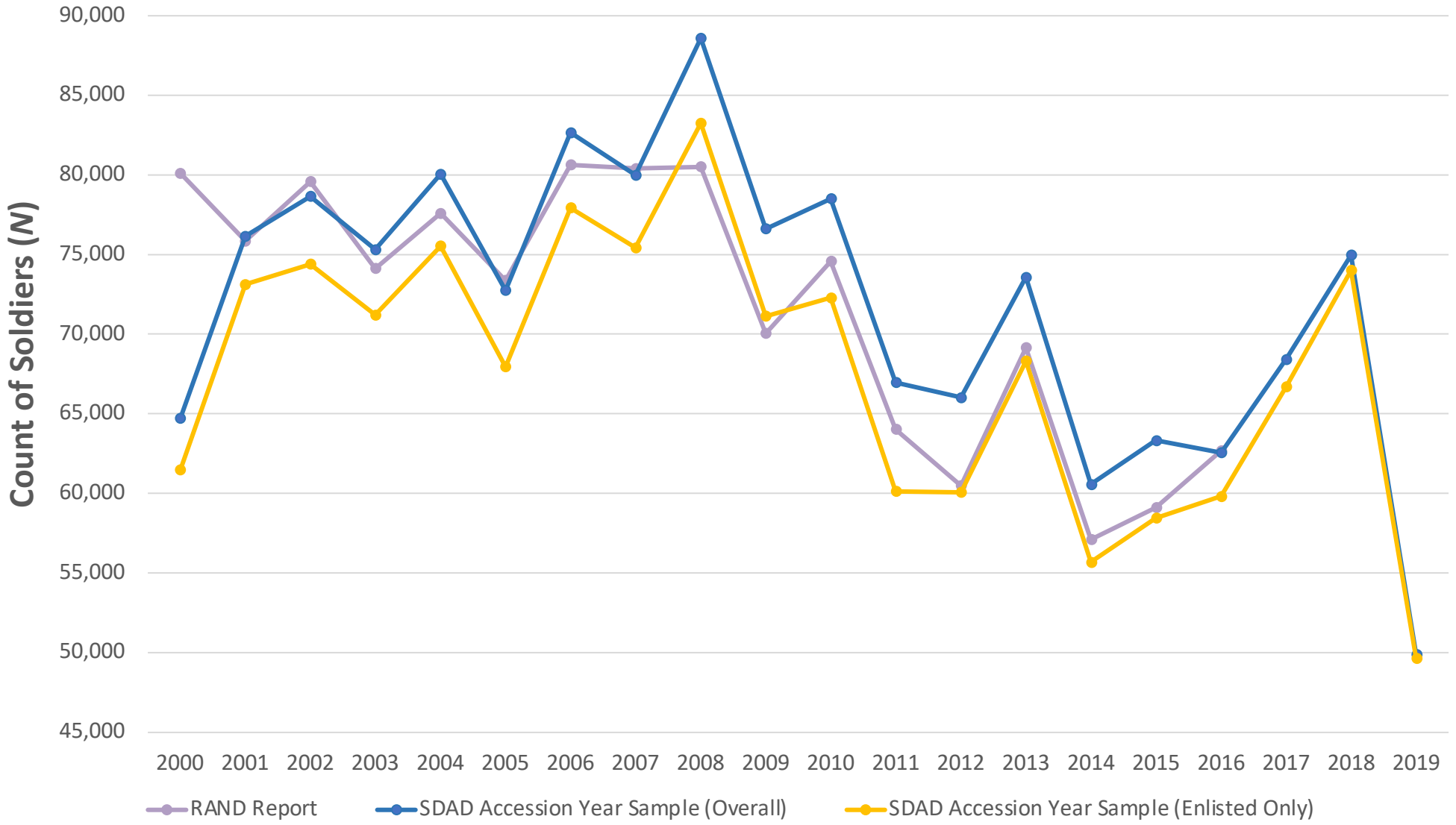
Census: **Demographics, ACS**
BLS: **Unemployment**
CDC: **BMI**

Crosswalk of DOD Demo Reports & PDE Data Top Level Cohort Numbers



Note. **DOD Demographic Reports include reservists and guard on 'active status'**

Crosswalk of RAND Report & PDE Data Top Level Accession Numbers



Overview of Modeling Phases

Problem: Examine feasibility of using administrative Army data sources to predict individual performance.

How: Each phase builds on previous phases to build statistical models of increased complexity.

Methods:

Modeling Phase 1 Analysis ***(complete)***

- Analysis of simple relationships using one predictor and one outcome variable
- Informed by prior research

Modeling Phase 2 Analysis ***(in-progress)***

- An analysis of complex relationships between the variables of interest.
- Examine feasibility of creating composite indices of performance potential among optimized predictors
- Will not account for changes over time (i.e., no time-varying variables or repeated measurements)

Modeling Phase 3 Analysis ***(planning)***

- Account for time as a component by modeling relationships using time-varying variables (i.e., variables that are measured at different time points across a Soldier's career)
- Modeling techniques to account for time as a factor:
  - Hidden Markov Models (HMM),
  - repeated SEM models (RSEM)
  - Multi-Level Models (MLM)

Phase 1 Performance Modeling

Problem: The Army possesses a trove of administrative data, but it is unclear if the data can be leveraged to construct models of Soldier performance

How: Conduct descriptive analyses, simple relationship analyses (using a single predictor and single outcome variable), and dominance analyses.

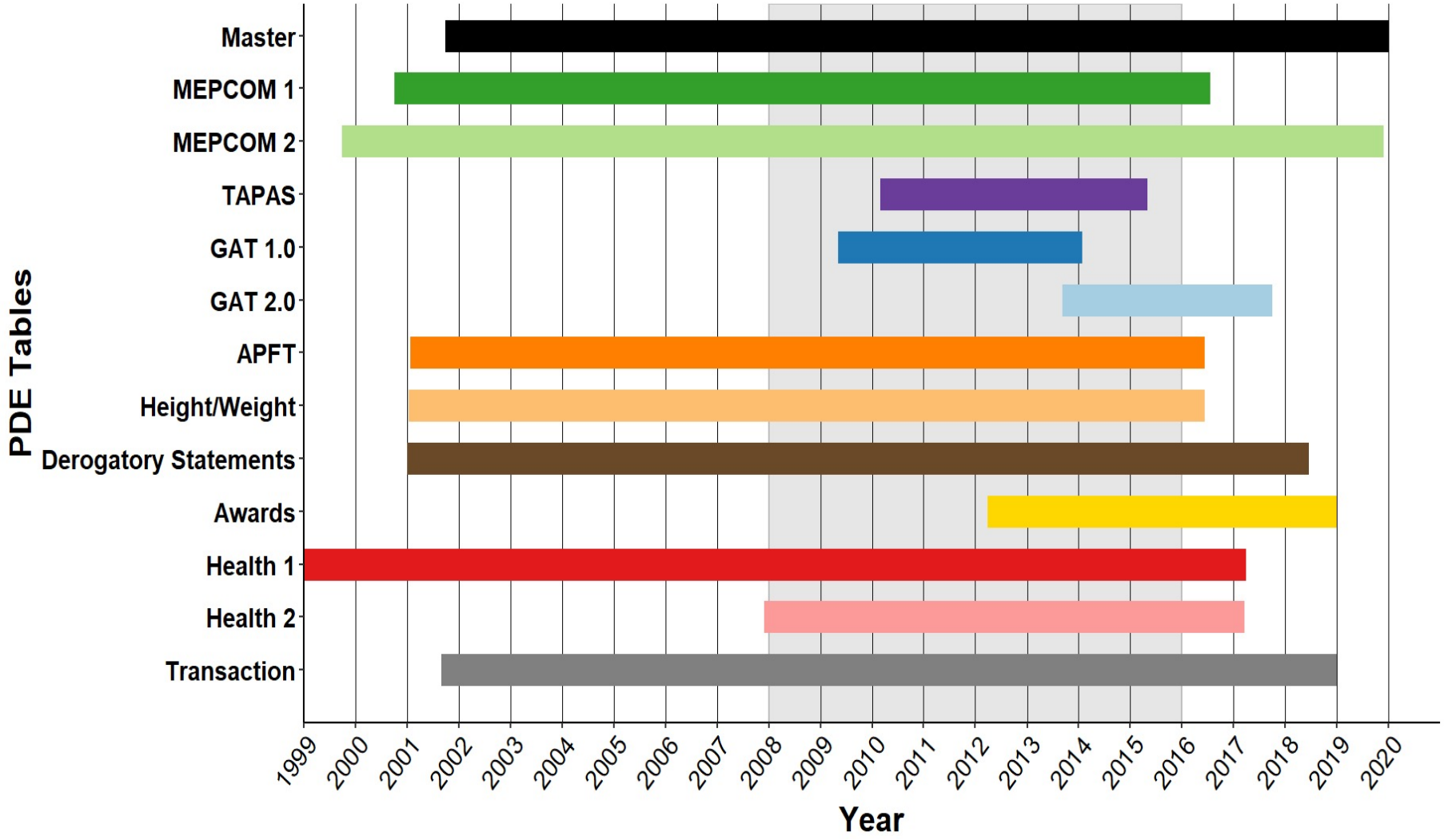
Methods: descriptive statistics, correlation, simple linear regression, generalized linear model, ANOVA

Payoff: This research could provide a baseline understanding of performance relationships in the Army and areas for further study.

Products & Transitions:

- Phase 1 Performance Modeling ARI Tech Report (under review with ARI)

Phase 1 Modeling Data Time



Grey area = sample used for Phase 1 analysis based on accession year (2008–2015)

CPM – Phase 1 Variables

Master: **Demographics (age, sex, race)**
MEPCOM: **AFQT Score**, Home Location, Education
TAPAS: **Personality Metrics**
GAT: **Psychosocial Metrics**

Master: Duty Location, Marital Status, Education
GAT: Psychosocial Metrics
PHA: **PULHES Score**
ATMS: Training, Weapon Qual.
ATRRS: Training, Coursework
FAM SPON: Dependents
CTS: Deployment History

DEOCS: Climate Survey
URI: Unit Risk Inventory

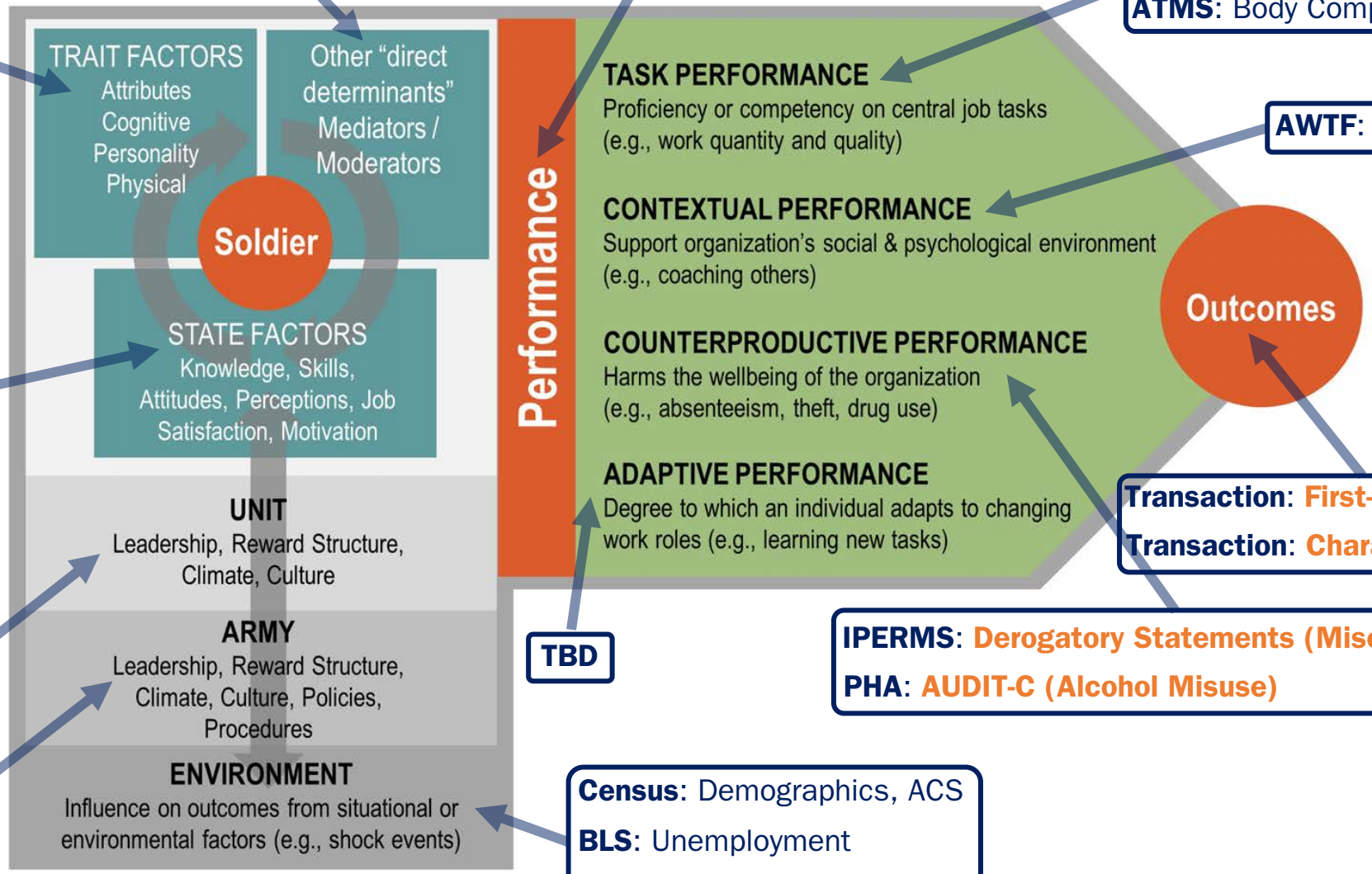
DEOCS: Climate Survey
Doctrine: FMs, ADRPs

Master: **Rank Group, MOS**

Master/Transaction: **Speed of Promotion**

ATMS: **APFT Scores**
ATMS: Body Composition

AWTF: **Awards (Merit)**



TBD

Transaction: **First-Term Attrition**
Transaction: **Character of Service**

IPERMS: **Derogatory Statements (Misconduct)**
PHA: **AUDIT-C (Alcohol Misuse)**

Census: Demographics, ACS
BLS: Unemployment
CDC: BMI

Phase 1 Modeling - Variable Selection

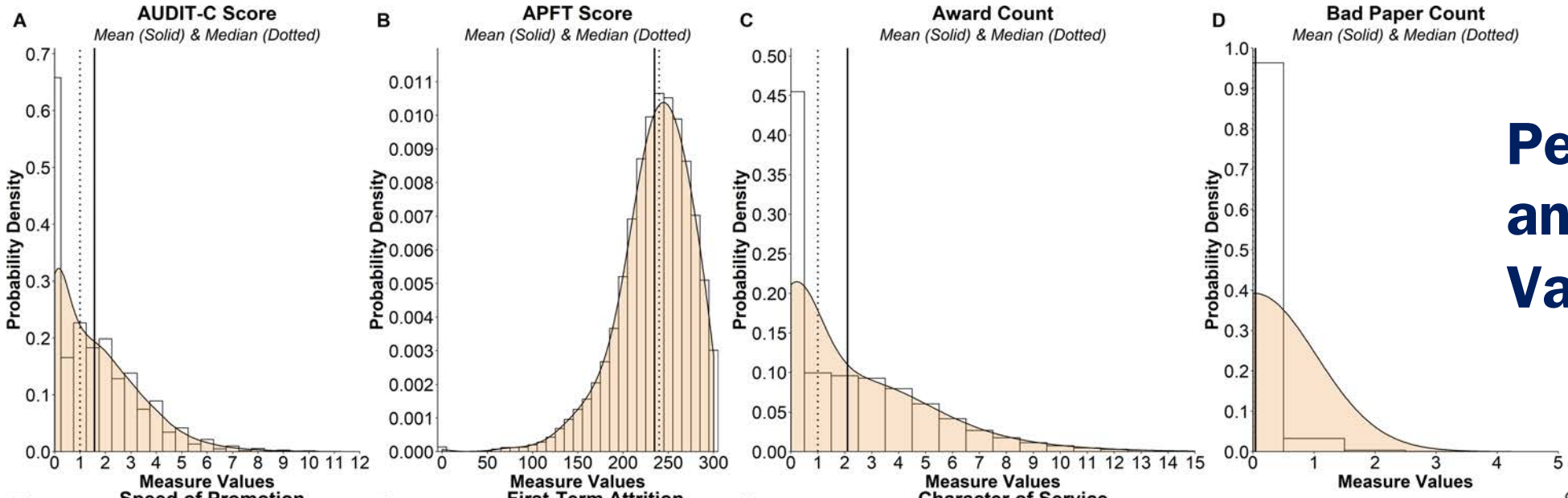
Variables	Variable Name	Type	Performance Component	Used/Derived	Expected Range	# Items
Predictor Variables						
Accession						
Rank Group	RANK_PDE_GRP	Army Function				
MOS Type	MOS_TYPE	Army Function				
Soldier Sex	PN_SEX_CD	Demographic				
Soldier Race	RACE_CD_RE	Demographic				
Age at Accession	AGE_ACC	Demographic				
AFQT Score	AFQT_PCTL.CB	Cognitive Ability				
PULHES Score	PULHES.MEAN	Health				
GAT						
Adaptability	adapt.scale.CB	Psychosocial	—	Derived	1–5	3
Active Coping	acope.scale.CB	Psychosocial	—	Derived	1–5	5
Passive Coping	pcope.scale.CB	Psychosocial				
Character	chr.scale.CB	Psychosocial				
Catastrophizing	catastro.scale.CB	Psychosocial				
Depression	depress.scale.CB	Psychosocial				
Optimism	optimism.scale.CB	Psychosocial				
Positive Affect	posaffect.scale.CB	Psychosocial				
Negative Affect	negaffect.scale.CB	Psychosocial				
Loneliness	lone.scale.CB	Psychosocial				
Organizational Trust	orgtrust.scale.CB	Psychosocial				
Work Engagement	wkengage.scale.CB	Psychosocial				
Life Meaning	lifemean.scale.CB	Psychosocial				
TAPAS						
Achievement	ACHVMNT_THETA_SCR_QY	Personality				
Adjustment	ADJ_THETA_SCR_QY	Personality				
Adaptation	ADPT_CMPS_SCR_QY	Personality				
Adventure	ADV_THETA_SCR_QY	Personality				
Attention Seeking	ATTN_SEEK_THETA_SCR_QY	Personality				
Commitment to Serve	CMTS_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Cooperation	COOPR_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Courage	COUR_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Dominance	DOMNC_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Even Tempered	EVTMP_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Intellectual Efficiency	INTLL_EFC_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Non-Delinquency	NON_DLNQY_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Optimism	OPTMSM_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Order	ORD_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Physical Condition	PHY_COND_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Responsibility	RSBY_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Sociability	SCBLTY_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Self-Control	SELF_CTRL_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Situational Awareness	SITNL_AWRNS_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Selflessness	SLFNS_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Team Orientation	TEAM_ORNTN_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Tolerance	TOL_THETA_SCR_QY	Personality	—	Used	$\mathbb{R}^- – \mathbb{R}^+$	varies
Outcome Variables						
AUDIT-C Score	AUDITC_TOTALSCORE_MEAN	Alcohol Misuse or Abuse	Counterproductive	Derived	0–12	3
APFT Score	APFT_TOTALSCORE_MEAN	Physical Fitness	Task	Derived	0–300	3
Award Count	award_count	Merit	Contextual	Derived	$0– \mathbb{R}^+$	—
Bad Paper Count	badpaper.overall	Misconduct	Counterproductive	Derived	$0– \mathbb{R}^+$	3
Speed of Promotion	SOP.RANK_HIGH.STDZ2	Rank Achievement Time	General	Derived	$\mathbb{R}^- – \mathbb{R}^+$	—
First-Term Attrition	ATTRIT_FIRST_TERM	Separation from Army	Outcome	Derived	2	—
Character of Service	CHAR_SVC_CD2	Terms of Separation	Outcome	Derived	2	—

Note: $\mathbb{R}^+$ = any real positive number, $\mathbb{R}^-$ = any real negative number. Number of items varies by GAT version (first number = GAT 1.0, second number = GAT 2.0).

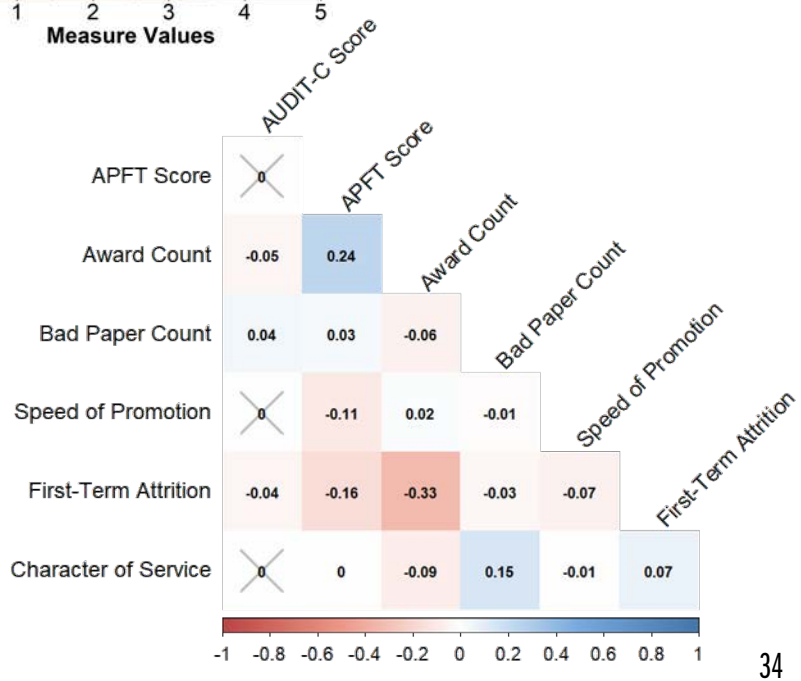
Informed by conceptual profiling process

A non-exhaustive total of **42 predictor variables** and **7 outcome variables** were selected for initial Phase 1 modeling of simple relationships

Phase 1 Modeling – Selected Results

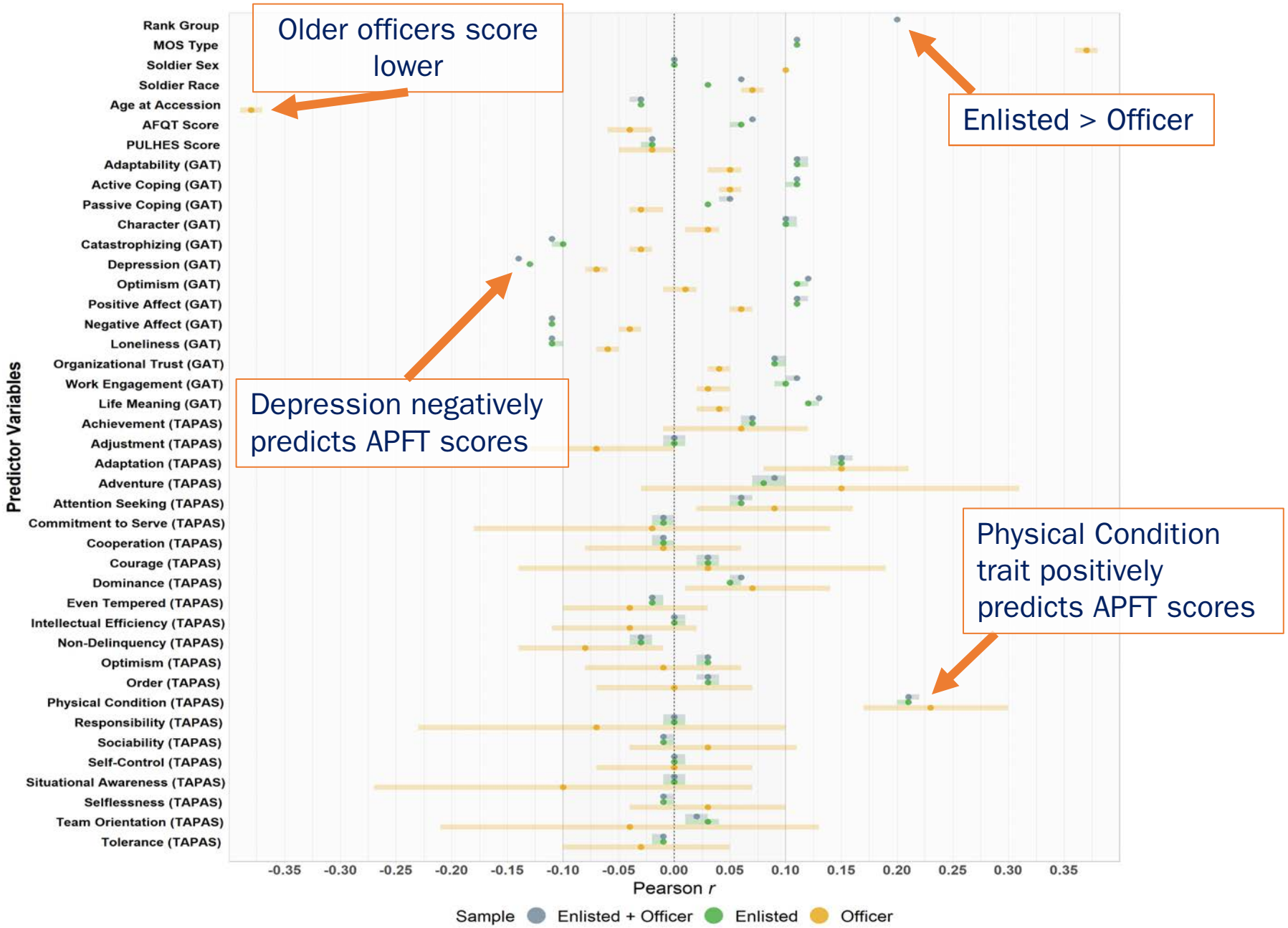


Performance and Outcome Variables



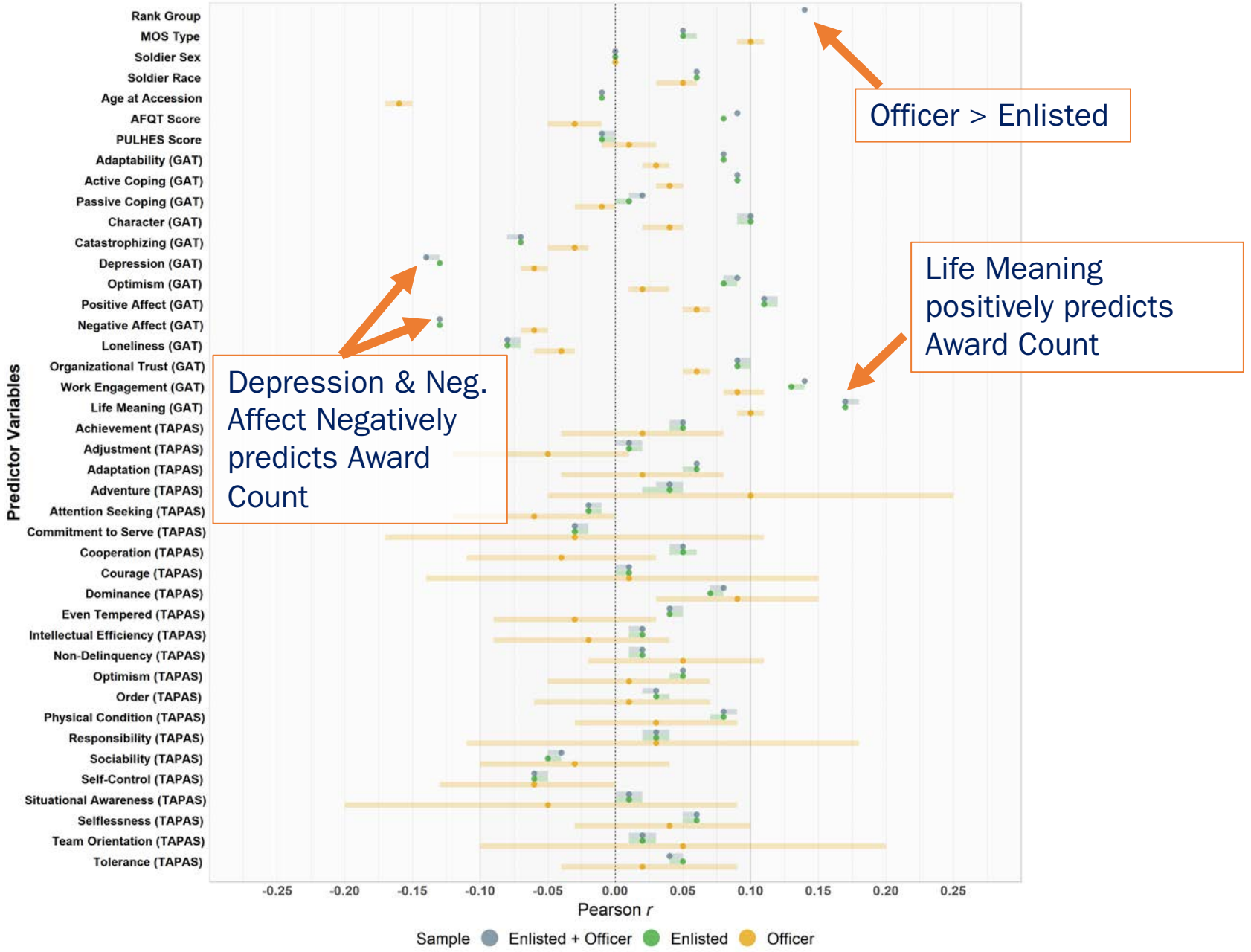
Phase 1 Modeling – Selected Results

Simple Relationship Analysis: Predictors with average **APFT Score** over a Soldier's career

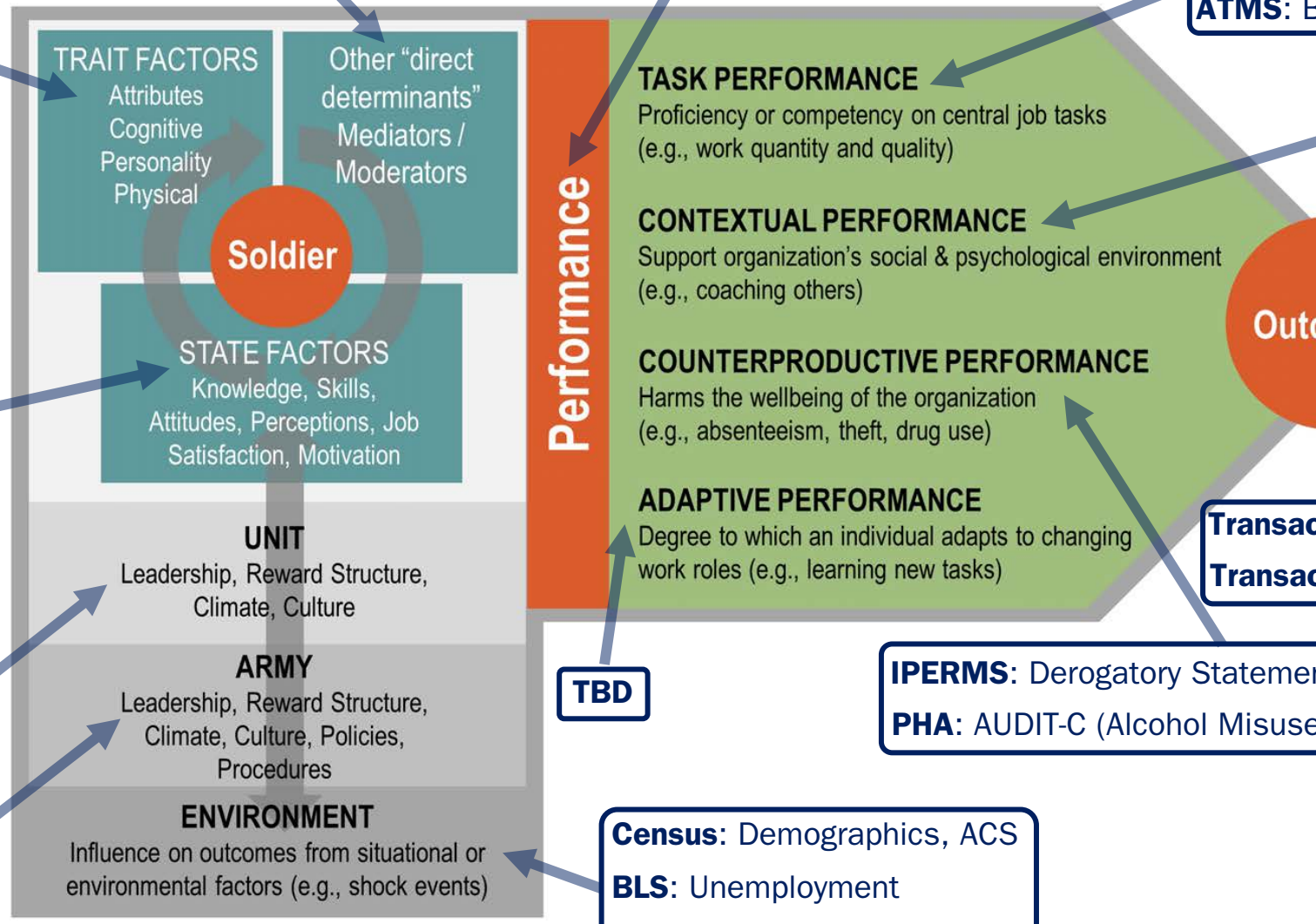


Phase 1 Modeling – Selected Results

Simple Relationship Analysis: Predictors with **Award Count** over a Soldier's career



Conceptual Performance Model



Master: Demographics (age, sex, race)
MEPCOM: AFQT Score, Home Location, Education
TAPAS: Personality Metrics
GAT: Psychosocial Metrics

Master: Rank Group, MOS

Master/Transaction: Speed of Promotion

ATMS: APFT Scores
ATMS: Body Composition

AWTF: Awards (Merit)

Master: Duty Location, Marital Status, Education
GAT: Psychosocial Metrics
PHA: PULHES Score
ATMS: Training, Weapon Qual.
ATRRS: Training, Coursework
FAM SPON: Dependents
CTS: Deployment History

Transaction: First-Term Attrition
Transaction: Character of Service

DEOCS: Climate Survey
URI: Unit Risk Inventory

TBD

IPERMS: Derogatory Statements (Misconduct)
PHA: AUDIT-C (Alcohol Misuse)

DEOCS: Climate Survey
Doctrine: FMs, ADRPs

Census: Demographics, ACS
BLS: Unemployment
CDC: BMI

CPM – Phase 2 Variables

Master: Waivers, 1st Duty location

MEPCOM: Education at accession (ROTC), Home location (Zip Code), Quality of recruit

Master: Marital Status, Dependents

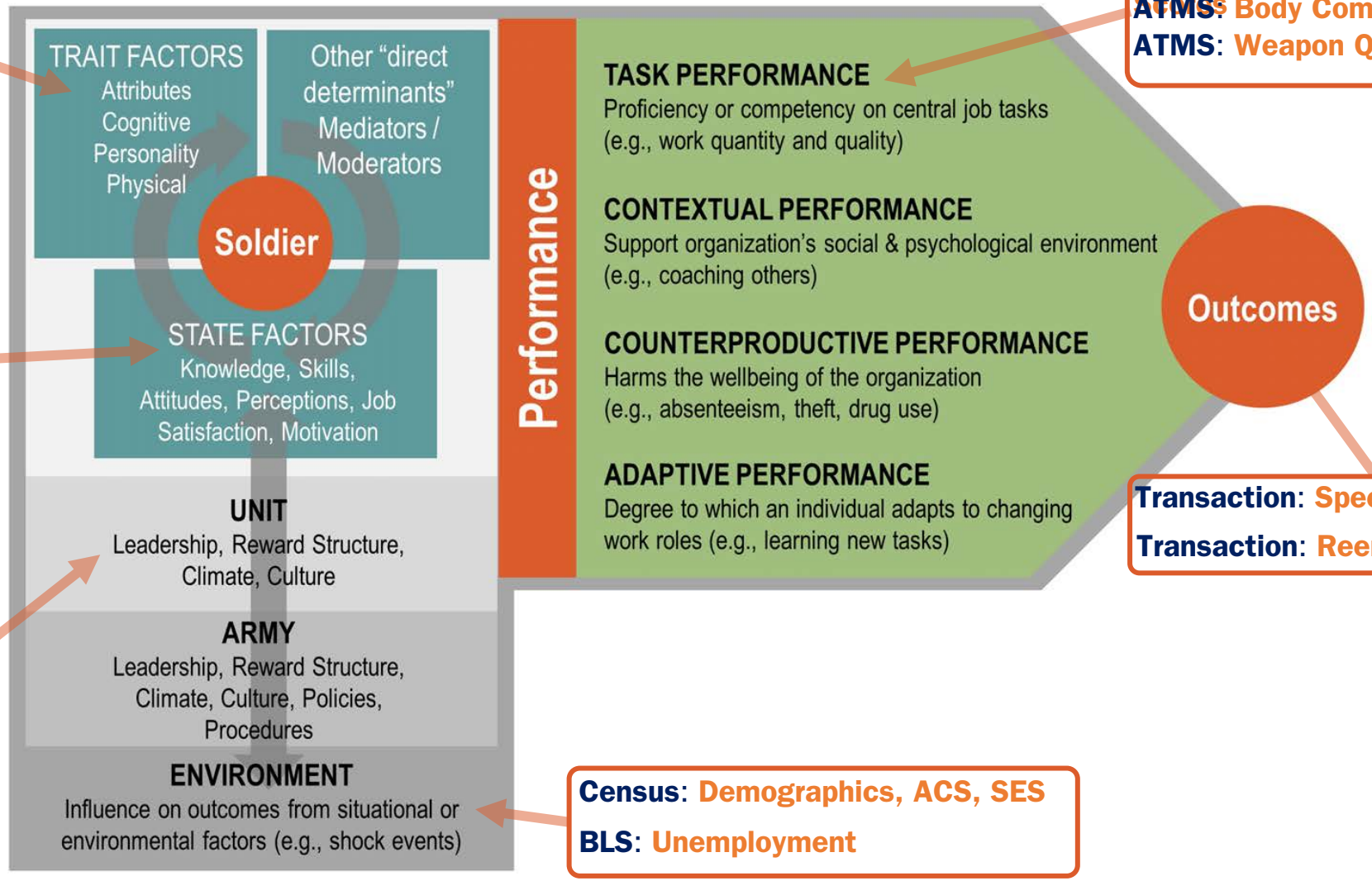
GAT: Family

FAM SPON: Dependents

CTS: Deployment History (frequency, time in between)

DEOCS: Climate Survey

URI: Unit Risk Inventory



ATMS: APFT, ACFT

ATMS: Body Composition

ATMS: Weapon Qualification

Transaction: Speed of promotion

Transaction: Reenlistment

Census: Demographics, ACS, SES

BLS: Unemployment

Current/Next Steps

Develop research questions for Phase 2 modeling

- Examine complex relationships with more than one predictor variable
- Examine complex relationships with more than one outcome variable
  - Construct performance composite indicators from multiple performance variables (e.g., multivariate tests, composite indices)
- Examine complex relationships with new variables

  (e.g., quality of recruits, deployment history, weapon qualification, home geography, SES at entry)

Developing Predictive Models of U.S. Army Career Pathways

UNIVERSITY of VIRGINIA

BIOCOMPLEXITY INSTITUTE

Outline

- Conceptual Framework and Overview of Models (Joel)
- Sequence Analysis on BGT Data (Joanna)
- Probit Models (Aritra)
  - Synthetic Generation
- Current /Next Steps (Josh)

Qualitative Informs Quantitative

Document analysis + interviews

- Review of Army admin documents from 2000 - 2020 (i.e., doctrine, directives, memos, DTIC technical reports, press releases)
  - Identify official Officer promotion trajectory and accompanying characteristics
  - Uses thematic coding and hybrid inductive-deductive approach
  - Team coding with reliability checks
- Interview w/ LTC Lockhart and LTC Danderson (TMTF)

Iterative learning

- Document analysis & discussions –> Discovery and analysis of new documents
- Wardynski et al. background readings that informed TMTF

Outcomes

- Provide rich description that informs and aids in the formation, interpretation, and validation of research questions and statistical models

Qualitative Outcome: Research Questions

How do predictors and covariates interact to create career pathways?

- Are there reoccurring relationships between

  (1) source of commission,

  (2) postings, deployments, and promotions, and

  (3) MOS that can be grouped to create a taxonomy of careers?

- Do the key factors we identified (i.e., West Point, ground combat experience, seniority) predict more successful careers,

  e.g., do we see disproportionate promotion rates among combat arms compared to their baseline in the total Soldier population?

MOS is used as shorthand to indicate area of assignment and specialization for all Soldiers

Conceptual Model

Initial inputs Predictors & covariates present prior to or at Soldier accession	Career Pathways Predictors & covariates that occur during Army career; interactive and iterative	End states Soldier endstate at point of separation from the Army, i.e., end point of their career path
Primary Qs: Do we have this data? If not, where and how do we get it?	**Primary Qs:** Do we have this data? And how is it best represented in our models? Are new models needed?	**Primary Qs:** Do we have this data? And how is it best represented in the model(s)?
Demographics, e.g., race, ethnicity, sex, age, SES (education, home location income, etc.), family status (married, children)	Initial and subsequent MOS assignments, e.g., combat arms, combat support • What is the likelihood of switching MOS?	Total time in service and final rank achieved • How are we defining a complete and/or successful Army career (e.g., 20+ years, LTC/COL)?
MEPS data, initial physical testing, psychosocial state/trait assessments, e.g., TAPAS, individual differences in personality traits • Personality (absolute, relative to peers) • Intelligence (absolute, relative to peers) • Health/fitness (absolute, relative to peers)	Initial and subsequent postings • What are common "sequences" of postings by MOS? • Requires useful taxonomy of postings which might be challenging; at a minimum can we verify Wardynski et al. claim that HRC prioritizes alternating between CONUS and OCONUS postings? • How do we track or distinguish assignments in primary MOS vs. assignments in other areas (e.g., key developmental area)? Ask RFL what data is available on Soldier assignments	Turnover dynamics (voluntary & involuntary) • What is the likelihood (or time required) of passing VOLUNTARY courses?
Waivers (different types)	Initial and subsequent deployments • When and how frequently are people deployed? • What are common deployment "sequences" by MOS? • Do we have access to full list of deployments or only Overseas Contingency Operations? Another RFL Q?	Major life events, e.g., marriage, divorce, children, death of family
Selection source, source of commission (for officers), e.g., West Point vs. ROTC vs. OCS	Promotion dynamics • What is the likelihood of getting to a point where you can be considered for promotion at each rank by MOS? • What is the likelihood of non-promotion vs below zone, in zone, and above zone at each rank by MOS?	Performance dynamics and ongoing physical and psychosocial assessments (e.g., counterproductive behavior like alcohol and drug use, GAT scores) • What is the likelihood of "good" or "bad" OER ratings; need to define, e.g., use top block vs. not? • What is the likelihood of medical events over career (what classes of event?) • What is the likelihood of ART15, NJP over career? • What is the likelihood of receiving award or commendation? • How do we account for or incorporate personality, intelligence, health/fitness testing, possession/demonstration of Army leadership attributes and competencies into model?
Prior military experience	Education and training, e.g., Masters/PhD attended, credentials received • What is the likelihood (or time required) of passing key REQUIRED courses for MOS (e.g., AIT)? • What is the likelihood of being assigned to particular VOLUNTARY courses (e.g., Ranger school)?	External and contextual factors, e.g., DOD policy, congressional mandates • Seasonal effects (e.g., Quarterly) • Year-specific (period) effects • Force strength caps • Change in $ value education benefits
Temporal and socioeconomic factors • Economic conditions near home of record and general economic conditions • Other social conditions near home of record (e.g., crime, education levels)		
Other external and contextual factors, e.g., policy, congressional mandates • Force strength caps • $ value of education benefits • $ value of retirement benefits • Relaxed requirements, increase in waiver issuance (just for accessions? Or for other events?) • $ value of signing bonus		

Conceptual Model

Initial Inputs
Predictors & covariates present prior to or at Soldier accession

Primary Questions:
- Do we have this data?
- If not, where and how do we get it?

Career Pathways
Predictors & covariates that occur during Army career; interactive & iterative

Primary Questions:
- Do we have this data?
- How is it best represented in our models?
- Are new models needed?

End States
Soldier endstate at separation from the Army, i.e., end point of career

Primary Questions:
- Do we have this data?
- How is it best represented in the models?

Methods Aligned with Research Questions

Models	Inference	Relevant Questions
Sequence Analysis and Clustering • Exploratory approach for classifying a typology of career trajectories • Inputs are states (events in a career) and time • Clustering to classify similar sequences into types of career trajectories • Inputs for cluster analysis are a sequence dissimilarity metric and clustering method		Reoccurring relationships between (1) the source of commission; (2) initial and subsequent postings, deployments, and promotions; and… *Compare estimated effects for individuals from different source of commission. Difference or similarity? Confidence intervals help here for comparing difference* (3) MOS or MOS groupings that can be grouped to create a taxonomy of career pathways? *Subset by starting state for individuals and fit the model for each starting state and compare the effects to see if a group of MOS's share the same dependence on covariates*.
Bayesian Multinomial Regression • Links career trajectories and covariates within one modeling framework • Embarrassingly parallel algorithm • Suffers from sample size constraints (Higgs and Hoeting 2010) • Computational efficiency increased using a Gibbs sampling through *probit regression* • Extensions to quantify spatial and temporal variation	Full scale Bayesian inference on effect sizes for various covariates (under the link function). Quantifies dependence of career trajectories on covariates and space-time. Drawbacks: Works on discrete/categorical ordered career states	What is the likelihood of renewing contracting (and if so, for how long) or acceptance of new ADSO (by rank and MOS)? What is the likelihood of incurring additional obligations (e.g., ADSO) (via PCS, schoolhouse choice, etc.) (by rank and MOS)? What is the likelihood involuntary turnover before (or after) end of contract/ADSO (by rank and MOS)? *Estimate the probability for such occurrences using either a MNL or a survival framework*
Multivariate Survival (Time to Event) Models • Extends the Bayesian MNL model to continuous outcomes. Rich literature on applications (Crowder M.J. 2012) • Easily incorporate temporal and spatial extensions • Ability to deploy in the PDE	Ability to incorporate continuous outcomes through time to event definition. Full scale Bayesian inference available (for space-time extensions as well) under appropriate choice of hazard function.	What combination of predictors and covariates produce "opportunity cost" effects that lead people to stay vs. leave the Army? *A Multivariate survival regression framework is one of the useful options here*

SEQUENCE ANALYSIS

UNIVERSITY of VIRGINIA

BIOCOMPLEXITY INSTITUTE

Developing Predictive Models of U.S. Army Career Pathways

Problem:

Critical gap in the Army's capability to manage talent over a Soldier's career pathway. The Army needs an integrated means to measure talent management.

Our approach:

1. Link the veteran Burning Glass Technologies resume data with data from the Occupational Information Network
2. Construct a typology of career pathways using Sequence and Clustering Analyses

Proof of concept:

This project tests statistical approaches and explores Veteran career progression in the first ten years after leaving the military using non-DOD data sources.

Geographic Area	Unit of Analysis	Time Period
Washington, D.C. Maryland Virginia	Veteran Resumes	2016 to 2018

Data Sources

- Burning Glass Technologies (BGT)
  - Proprietary data collected from publicly available sites between 2016 and 2018
  - Includes anonymized information on education, certifications, skills, geographic location, job history
- Occupational Information Network (O*NET)
- Standard Occupation Classification (O*NET-SOC) codes classify jobs
  - O*NET-SOC codes also correspond with O*NET job zones, which are grouped based on the amount of preparation needed for a position (zone 1-least; zone 5- most preparation)

Geographic Level	Unit of Analysis	Time Period Used in Analysis
DC, MD, VA, (**DMV**) and Texas	Individual (Resume)	Resumes collected from 2016-2018

BGT resume data quick facts

Value of O*NET and BGT Data for Testing Modeling Approaches

Model formulation with O*NET and BGT data is a preliminary step to Army data

- Both datasets allow for creation of career pathways, e.g., O*NET Job Zones and Army Rank
- Both datasets contain spatial and temporal information
- Both datasets have covariates to build a predictive modeling framework

Approaches to address differences in the datasets

- Building in transition rules ("hard checks")
  - Certain types of transitions (e.g., transitions to lower job zones, a single job zones over 30 years) are common in the BGT data, but uncommon or nonexistent in Army data
- Pulling data from Texas to test model performance
  - Potential bias in building models that perform well for transitions in Washington D.C., Maryland, and Virginia

Using Sequence and Clustering to Analyze Career Progression

Career progression is captured by transitions between employment, unemployment, and education

- Transitions are ordered into **sequences of states** (job zones, education, and unemployment)
- Sequences are standardized for periods of 10, 20, and 30 years after exiting the military

Sequences with similar characteristics are grouped using cluster analysis

- Sequence clustering involves choosing a dissimilarity metric, a method of clustering, and the number of clusters.
- The optimal cluster solution is chosen based on criteria of cluster quality and our ability to interpret the cluster.

We use sequence clusters to develop a typology of career trajectories

Data Linkage-BGT Resume & O*NET Job Zones

Raw BGT Resume data

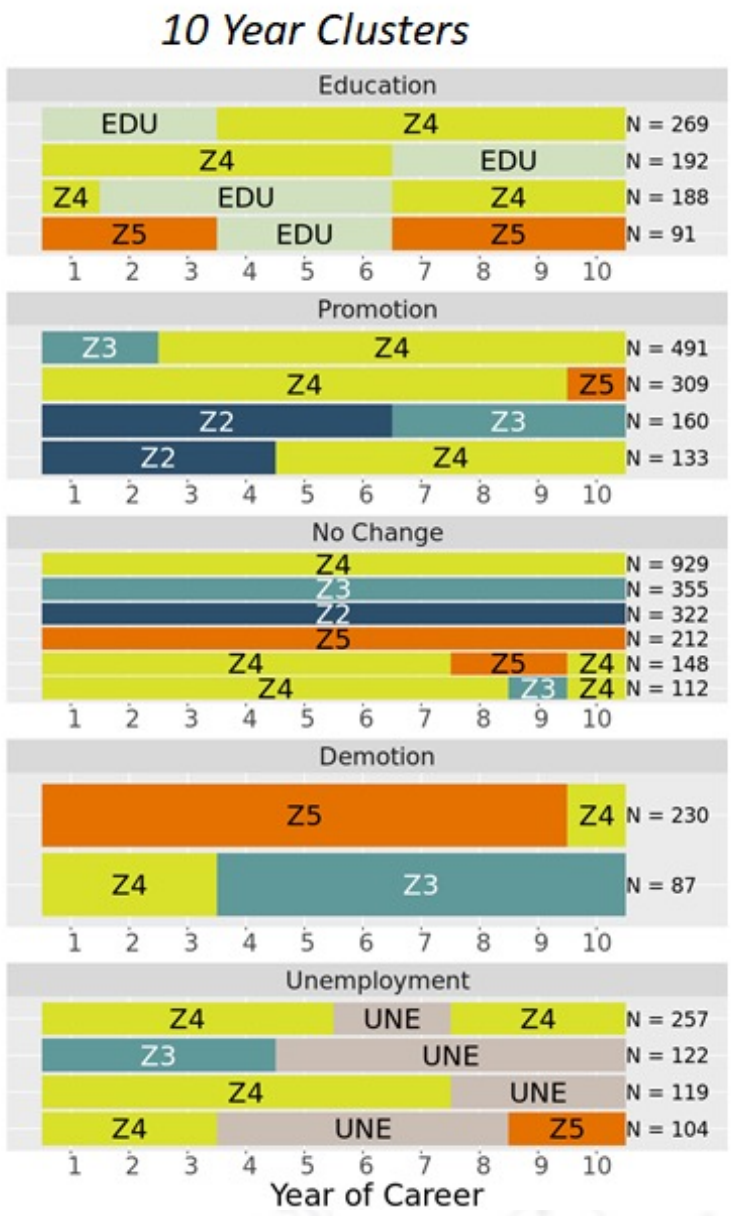
id	onet	startdate	enddate
5930181	53-7062.00	2007-12-01	2018-03-14
5930181	55-3019.00	1996-03-01	2005-09-01
5930181	53-2011.00	1983-01-01	1986-01-01
5930181	43-5111.00	<NA>	<NA>
5930181	15-1199.01	2010-02-01	2018-03-14
5930181	15-1199.09	2007-11-01	2010-02-01
5930181	15-1199.02	2001-10-01	2007-08-01
5930181	15-1121.00	1997-08-01	2000-10-01
5930181	11-9041.00	1995-01-01	1997-09-01

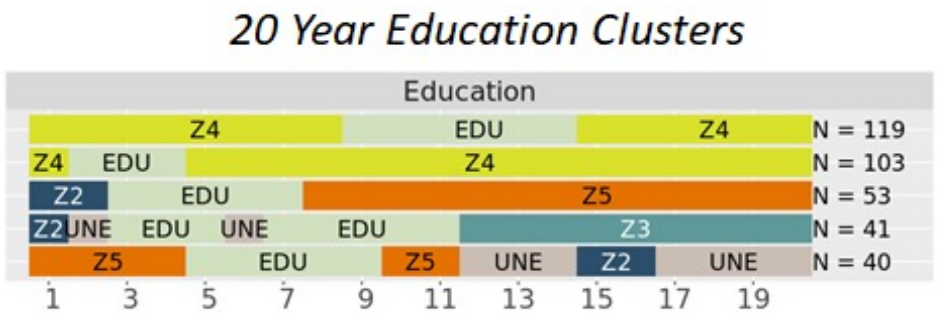
Data after cleaning and merging with job zones, education

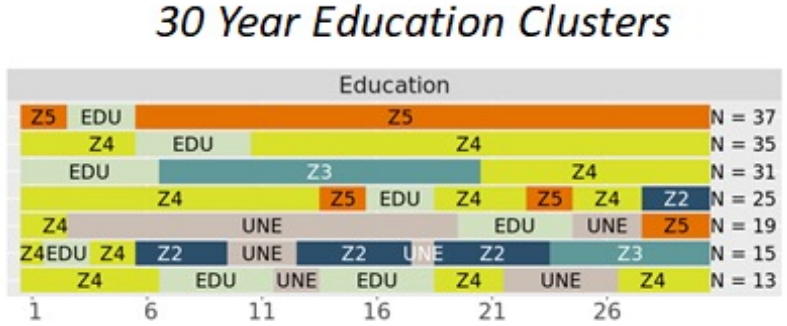
id	onet_job_zone	start_year	end_year
5930181	4	1983	1986
5930181	EDU	1991	1995
5930181	5	1995	1997
5930181	Military	1996	2005
5930181	4	1997	2000
5930181	EDU	1999	2001
5930181	4	2001	2007
5930181	2	2007	2018
5930181	4	2007	2010
5930181	EDU	2009	2011
5930181	4	2010	2018

Assign highest job zone to given year

10 Year Clusters for Veterans







- Grouped into 5 categories: Education, Promotion, No Change, Demotion, and Unemployment
- Zone 4 No Change cluster has the highest membership in multiple clustering runs
- Clusters are more complex in 20 and 30-year solutions
  - Characterized by more states
  - Featuring late/early career transitions

Conclusions from Sequence Analysis

- Overall, types of career trajectory clusters (Education, Promotion, etc.) were not as influential as individual job zones,
  - e.g., Officers were more likely to be in a cluster containing Zone 5 regardless if it was an Education, Promotion, No Change, Demotion, or Unemployment cluster
- Zones 4 and 5 clusters
  - Individuals earning Masters, PhDs, and Officers
  - Individuals skilled in Economics, Policy, Social Studies, Science & Research, and Analysis
- Unemployment Cluster
  - Individuals majoring in Computer Science less likely to be in this cluster

MODELING STRATEGIES

UNIVERSITY of VIRGINIA

BIOCOMPLEXITY INSTITUTE

Data Sources

- Burning Glass Technologies (BGT)
  - Proprietary data collected from publicly available sites between 2016 and 2018
  - Includes anonymized information on education, certifications, skills, geographic location, job history
- Occupational Information Network (O*NET)
- Standard Occupation Classification (O*NET-SOC) codes classify jobs
  - O*NET-SOC codes also correspond with O*NET job zones, which are grouped based on the amount of preparation needed for a position (zone 1-least; zone 5- most preparation)

Geographic Level	Unit of Analysis	Time Period Used in Analysis
DC, MD, VA, (**DMV**) and Texas	Individual (Resume)	Resumes collected from 2016-2018

BGT resume data quick facts

Proposed Methods

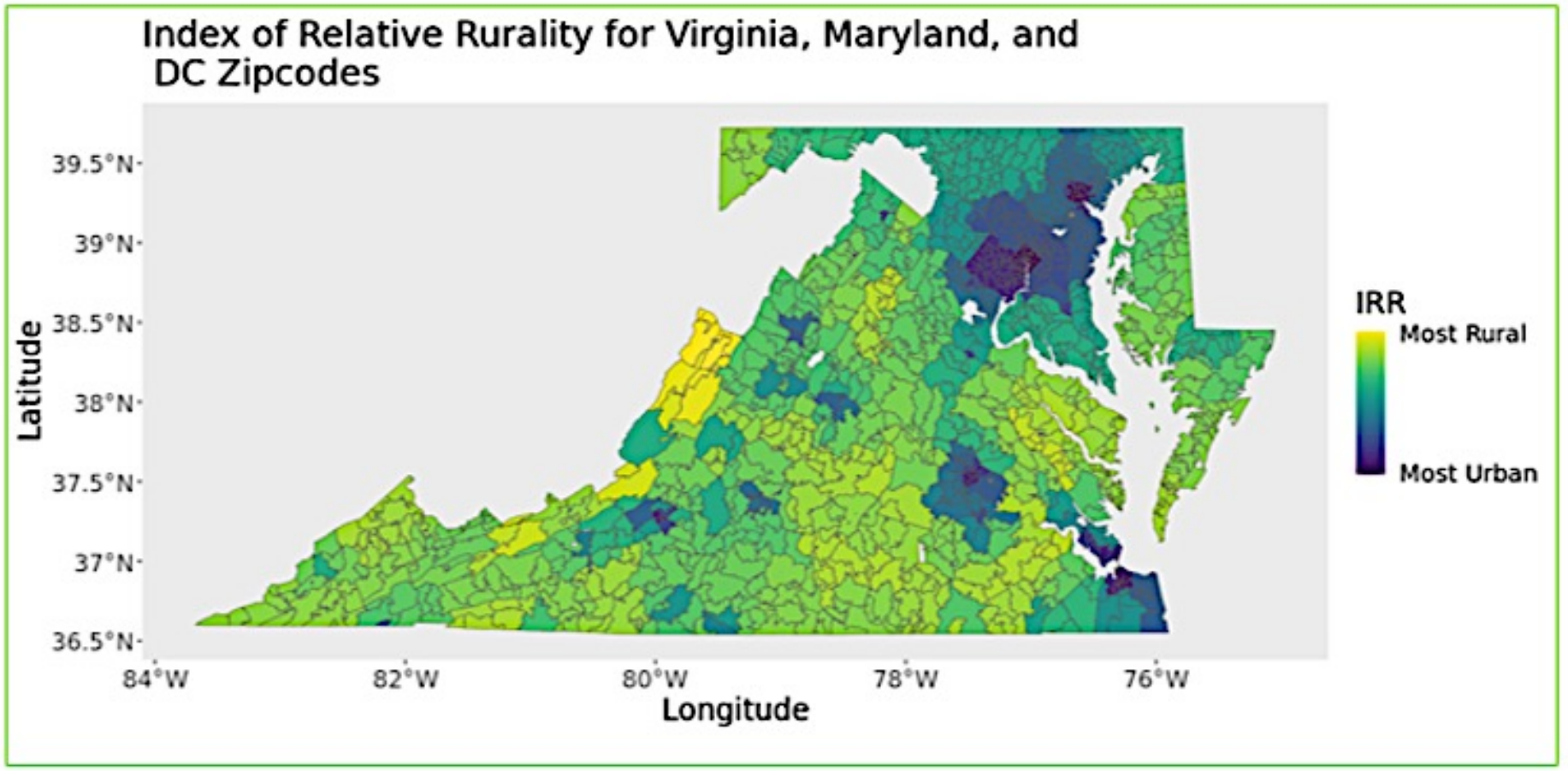
- Develop framework for modeling career pathways that allows for quantifying spatial and temporal variation
- Conduct synthetic experiments establish the accuracy of the model setup
- Include covariates to build a predictive modeling framework
- Apply framework using the Burning Glass Technologies resume data to provide insights into factors affecting workforce progression

Modeling Career Pathways

- Model career pathways in the Burning Glass Technologies resume data
- Bayesian Multinomial Logistic Regression **(MNL)** model with increasing complexity to capture
  - Impact of covariates featuring individual level history/experience on career transitions
  - Quantify spatial and temporal variation arising from location of individual
- Synthetic experiments showcasing accuracy and time complexity of proposed models
- Applications to Burning Glass Technologies Career Pathways
- Spatiotemporal Extensions

Linking Career Pathways and Veteran History

- Index of Relative Rurality (see Figure)
- Gender: “male”, “female” and “unspecified”
- Officer Status: “officer” and ”not officer”



MNL Model with Covariates

- Estimated transition probability matrix for the BGT career pathways for year-to-year transitions
- Model contains covariates

	Education	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Education	0.8213	0.0000	0.0128	0.0313	0.0951	0.0395
Zone 1	0.0000	0.6825	0.0635	0.1429	0.1111	0.0000
Zone 2	0.0318	0.0003	0.8260	0.0430	0.0763	0.0226
Zone 3	0.0240	0.0002	0.0245	0.8222	0.0949	0.0342
Zone 4	0.0310	0.0004	0.0118	0.0212	0.9022	0.0334
Zone 5	0.0292	0.0002	0.0128	0.0237	0.1023	0.8319

Modeling with covariates recovers the same transition probability matrix but allows for exploring effects of different covariates on the probabilities

MNL Model with Covariates

- Estimated coefficients for BGT data
- Interpret in the multinomial logistic scale for judging contributions to transition probability

Starts From	Transitioned to	Effect	Estimate	lower HPD	upper HPD
Education	Zone 1	IRR	-11.1948	-25.4981	3.9297
	Zone 2		0.0993	-2.0956	2.1368
	Zone 3		-0.5143	-2.1062	0.8902
	Zone 4		**-1.4429**	-2.3585	-0.6226
	Zone 5		**-3.6931**	-4.9662	-2.3775
	Zone 1	Officer Status	-1.6442	-4.4635	0.8010
	Zone 2		**-1.2539**	-1.6443	-0.9088
	Zone 3		**-0.7149**	-0.9616	-0.4338
	Zone 4		**-0.3845**	-0.5434	-0.2315
	Zone 5		**-0.3289**	-0.5731	-0.0864
	Zone 1	Gender	-1.9402	-4.3747	0.0268
	Zone 2		**-1.1355**	-1.4763	-0.7955
	Zone 3		**-1.1911**	-1.4393	-0.8800
	Zone 4		**-0.5870**	-0.7378	-0.4234
	Zone 5		**-0.8544**	-1.0884	-0.6342

Intercept Only MNL Model

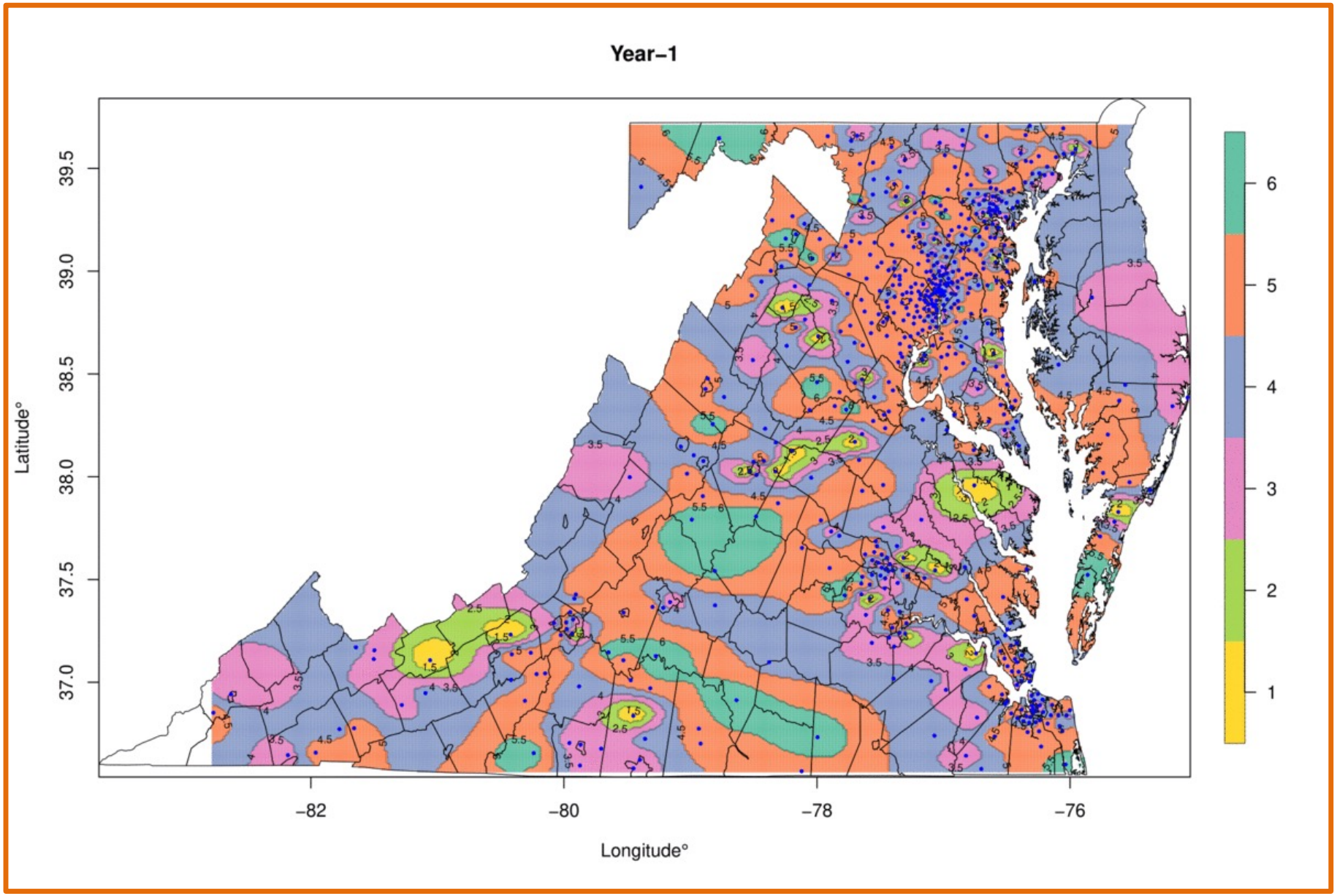
- Estimated intercepts β_0 for the BGT career pathways accompanied by 95% HPD intervals
- Posterior credible intervals (estimates) containing 0 are marked

(Starts From) Transitioned to	β_{est}	lower	upper
(Education) Zone 1	-7.9135	-9.1426	-6.6479
(Education) Zone 2	-4.0585	-4.2683	-3.8162
(Education) Zone 3	-3.4939	-3.6530	-3.3359
(Education) Zone 4	-2.1633	-2.2559	-2.0676
(Education) Zone 5	-3.0259	-3.1637	-2.8904
(Zone 1) Zone 1	4.2882	2.7214	6.8140
(Zone 1) Zone 2	**2.0150**	-0.2410	4.0741
(Zone 1) Zone 3	**2.1934**	-0.2844	4.2092
(Zone 1) Zone 4	2.3638	0.4578	4.7732
(Zone 1) Zone 5	**1.4222**	-0.7732	3.7038
(Zone 2) Zone 1	-4.1816	-5.4422	-2.9722
(Zone 2) Zone 2	3.4324	3.2478	3.6112
(Zone 2) Zone 3	0.5266	0.3179	0.7709
(Zone 2) Zone 4	0.9841	0.7829	1.2392
(Zone 2) Zone 5	-0.3005	-0.5743	-0.0395
(Zone 3) Zone 1	-3.8279	-4.7752	-2.9215
(Zone 3) Zone 2	-0.2616	-0.4687	-0.0296
(Zone 3) Zone 3	3.4761	3.3429	3.6350
(Zone 3) Zone 4	1.3020	1.1199	1.4575
(Zone 3) Zone 5	**0.0693**	-0.1303	0.2619
(Zone 4) Zone 1	-5.0204	-5.9505	-4.1426
(Zone 4) Zone 2	-0.8879	-1.0374	-0.7368
(Zone 4) Zone 3	-0.3238	-0.4512	-0.2099
(Zone 4) Zone 4	3.4963	3.3996	3.5750
(Zone 4) Zone 5	0.1446	0.0263	0.2512
(Zone 5) Zone 1	-5.4961	-7.3651	-3.7027
(Zone 5) Zone 2	-1.0530	-1.3235	-0.7617
(Zone 5) Zone 3	-0.2056	-0.4142	-0.0059
(Zone 5) Zone 4	1.1944	1.0027	1.3609
(Zone 5) Zone 5	3.4027	3.2438	3.5504

Spatiotemporal Variation

Spatiotemporal variation in the BGT data for year-to-year transitions within the various job zones:

- Zone 3(4) and Zone 4(5) being advanced job categories dominate the NOVA region
- Lower job categories/transitions appear as we start moving south



Probit Models: Continuous Modeling of Career Transitions

- Uses a probit link to connect the probabilities to the covariates
- Computationally more feasible
- Allows for a continuous modeling of job states and transition probabilities
- Sample size is no longer an issue

Creating Synthetic Career Pathways

Benefits of developing methods for generating a synthetic workforce

- Ability to estimate models without constraints of small sample sizes
  - Relevant in analysis of Army officers
- Analysis of workforce outcomes with different Army compositions, e.g., What would the officer corps look like in 10 years if a greater proportion of commissions came from Officer Candidate School?

Step 1: Using transition probabilities and intercepts to create synthetic pathways

- Measuring error using root mean squared error, maximum error, and sequence measures of error

Step 2: Using covariates

- Starting with gender, officer status, and IRR

Current/Next Steps

- Apply models to administrative data on Army Soldiers and Officers in the Person-Event Data Environment (PDE) data enclave
  - Datasets about military personnel including deployments, demographics, promotions, accessions/attrition, awards, and training records
- Capture progression of Soldier careers, transition in outcomes from state-to-state e.g., Rank, MOS
- Portray how transitions are driven by
  - Demographics
  - Military History
  - External Effects
- Use models to generate synthetic Army data

Thoughts and Discussion