Best Practices Project Team for Research on Non-Cognitive Assessment Methodologies for Military Compatibility

Brenda Ellis, PhD

Human Resources Research Organization

Briefing presented to the DACMPT

June 13, 2024

Briefing Agenda

- Best Practices Project Team (BPPT)

- Assessment Search Process

- Short-Listed Assessments

OPA
OFFICE OF PEOPLE ANALYTICS

Best Practices Project Team

OPA
OFFICE OF PEOPLE ANALYTICS

Purpose and Scope of Best Practices Project Team (BPPT)

- Offer expertise and guidance in search for existing assessments for use in assessing military compatibility

- Officer focus: differences in officer vs. enlisted population
  - Demographics: 4-year degree minimum; average age at accession = 22 (counterproductive behavior peaks in mid-20s)
  - Accessions process: Military Service Academies (MSA); ROTC, Officer Candidate School (OCS), direct commission
  - Existing testing infrastructure
    - Officers in general do not go through a MEPS, although they complete an SF 86 during screening
    - ROTC or MSA route does not require behavioral health screening (required for enlisted; may be required for OCS or direct commission)
    - TAPAS is not currently administered to officers

BPPT Members

Chad Van Iddekinge, PhD

Professor of Management, Tippie College of Business, University of Iowa
Expert and extensively published on psychological measurement and assessment selection

Mario Scalora, PhD

Professor of Psychology at University of Nebraska, Lincoln
Expert in forensic assessment, personality assessment, and the psychology of terrorism, targeted violence, and threat assessment

Deniz Ones, PhD

Distinguished Professor of I/O Psychology, University of Minnesota
Published extensively on individual difference predictors and correlates of counterproductive work behavior for over 30 years

Laura Baron, PhD

I/O Psychologist at U.S. Air Force Air Education & Training Command

David Corey, PhD, ABPP

Board-certified specialist in both forensic psychology and police and public safety psychology with American Board of Professional Psychology

Louise Fitzgerald, PhD

Professor Emeritus of Psychology
University of Illinois at Urbana-Champaign
Published extensively on sexual harassment in organizations for over 30 years

Assessment Search Process

OPA
OFFICE OF PEOPLE ANALYTICS

Process for Finding and Reviewing Potential Assessments

- Assessment search: reviewed 156 assessments
  - Two main search streams
    - Reviewed 42 **commercially available pre-hire assessments**
    - Reviewed 114 **assessments developed and used in research**
  - Information gathering—points of information gathered in initial search
    - **Criterion-related validity for criteria of interest**; test length; item content; test population; test uses; reliability; quality of criterion-related validity evidence
  - Exclusion criteria
    - Developed for diagnosing mental illness
    - Used only in children/adolescents
    - Owned by non-U.S. entity
    - Sensitive questions (e.g., questions about sex behavior or sex attitudes)
    - No evidence of criterion-related validity for criteria of interest

Defining the Criterion Space – Initial Input from BPPT

- Sexual *harassment* is distinct from sexual *assault*
  - *Harassment* consists of sexual coercion, unwanted sexual attention, and gender harassment, as described in the previous presentation
    - Gender harassment can also be divided into *sexist* hostility and *sexual* hostility
    - Most sexual harassment does not have a sexual gratification motive

- *Violent* behaviors and *extremist* behaviors
  - Include searching for predictors of violence and extremist behavior
  - These behaviors are represented in the 10 misconduct categories described in the previous presentation

- Counterproductivity peaks between ages 21 and 25; sexual harassment peaks around 25; trait levels such as conscientiousness change throughout this period
  - This will have implications for measurement—the predictor and criterion are both moving targets

Presenting Potential Assessments to the BPPT

- First round: initial list of viable assessments identified, narrowed down to 33 conditionally viable assessments
  - Information presented to BPPT
  - **Criterion-related validity for criteria of interest**; test length; item content; test population; test uses; reliability; sources of criterion-related validity evidence

- Discussion points
  - Relative strength of criterion-related validity
  - Amount of evidence for test reliability and validity
  - Ability to adapt for military use

- BPPT recommendations and other test considerations
  - Adverse impact: gender differences in misconduct predictor scores—males tend to score higher; however, males also tend to engage in more CWB
  - ADA: when in the accessions process is the test administered, post-offer or pre-offer?
  - DoD is exempt from Title VII, but military-specific legal issues may exist that cannot be anticipated without review from military legal expert

Considerations for Commercial vs. Research Assessments

Commercial Assessments

(+) Criterion-related validity evidence from applicant samples

(+) Likely have methods for deterring or accounting for faking/desirable responding

(+) Likely have developed assessments with EEOC and litigation risks in mind

(−) Difficult to obtain technical information without contacting vendors directly*

(−) Licensing and test ownership may present barriers

(−) Most vendor assessments focus on bright side traits/integrity[†]
- May have limited use for predicting sexual assault or sexual harassment
- May have substantial overlap with TAPAS facets

Assessments from Research

(+) Body of research on psychometric properties and criterion-related validity

(+) Often able to obtain and view measure/scale items Likely

(+) Often the research on the assessment is peer reviewed

(+) Often theoretical literature on measure constructs is available

(−) Assessments may need adjustments/additions to mitigate faking/desirable responding

(−) Assessments may need to be evaluated for bias or differential prediction

(−) Less likely to have been used on a large-scale basis

*Some vendors offer their technical manuals for sale without going through the sales process (e.g., CPI)
[†]Exceptions exist

Short-Listed Assessments

OPA
OFFICE OF PEOPLE ANALYTICS

Presenting Potential Assessments to the BPPT

- Second round: 13 short-listed assessments identified; brief report submitted to BPPT and DTAC

Commercial Assessments		Assessments from Research	
Automated Scoring	**Professional Judgment**	**Civilian Research**	**Military Research**
Personnel Reaction Blank	CPI	HEXACO	LBQ (Air Force)
Reid Report	MMPI	Short Dark Tetrad	NSAB (Army)
	MPQ	Dirty Dozen	CBEF/OBEF (Army)
		Moral Disengagement	ARC (Army)

Note. CPI = California Psychological Inventory; MMPI = Minnesota Multiphasic Personality Inventory; MPQ = Multidimensional Personality Questionnaire; LBQ = Lackland Behavioral Questionnaire; NSAB = Non-commissioned Officer Special Assignment Battery; CBEF/OBEF = Cadet/Officer Background and Experience Form; ARC = Assessment of Right Conduct. "Professional Judgment" assessments require scoring or score interpretation by a qualified individual.

Recommendations from BBPT

- For literature review, more weight should be placed on criterion-related validity evidence in <u>applicant</u> samples (vs. research or incumbent samples)

- Map assessment scales to their underlying constructs
  - The initial effort searched for measures with evidence of criterion-related validity
    - Several assessments appear to measure similar constructs
  - Mapping to constructs will facilitate examining content and construct validity
    - Conduct SME exercises for validity inferences (paths 6, 7, 8)
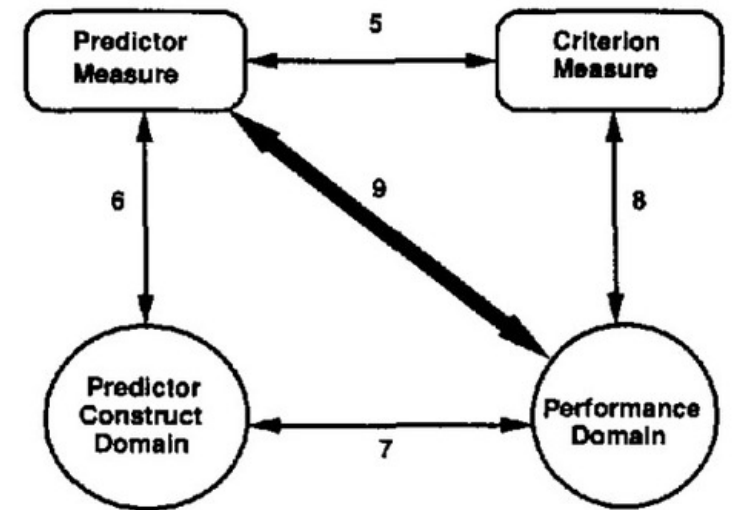
Figure 2. A common conception of the inferences for personnel selection.

Note. Figure from Binning & Barrett (1989).

Additional Points to Consider

Faking on Non-Cognitive Assessments

- On average, applicant scores are higher than non-applicant scores
  - Important to have correct normative data for personality measures
  - Base norms on operational use

- Ordinary impression management has little influence on factor structure or normal personality measures

- Hogan et al. (2007): applicants re-applying for a job did not "improve" their personality test scores

- Social Desirability Scales and Corrections do not improve criterion-related validity or content validity
  - Race differences in social desirability can lead to adverse impact

- Warnings *do* appear to decrease faking
  - Warnings can be given regardless of detected faking

- Less evaluative measures/items can decrease fakeability but slightly decrease criterion-related validity

- Forced-choice (FC) measures
  - Faking correlated with g in FC measures (Vasilopoulos et al., 2006)
  - IRT scoring of FC does not improve criterion-related validity

Narrowing the List of Assessments:
Comparing Short-Listed Assessments with Each Other

- Goal: obtain final list of assessments to recommend for further research
  - SME judgment exercise—using psychometric checklist developed by DTAC for ASVAB evaluation
    - Identified test evaluation factors that could be judged via SME
      - Step 1: SMEs assign weights to each evaluation factor
      - Step 2: debrief to achieve consensus on weights
      - Step 3: SMEs rate each assessment on each evaluation factor
      - Step 4: assess interrater agreement for each assessment on each evaluation factor; calculate final score for each assessment

Assessment Evaluation Factors

Evaluation Factor	Description
Construct definition(s)	Domains/subdomains measured; construct(s) related to other constructs/factors (or confounds)
Item taxonomy and test scoring	Content taxonomy/types/categories of items; how the test is scored (usability of scores for selection/screening)
Content validity	The extent to which the test content aligns with the domains being assessed and predicted
Convergent/divergent validity	Correlation with other measures or scales
Subgroup differences	Potential for (or evidence for) differential subgroup performance due to item type, content, or other
Criterion-related validity	Relationship(s) between the construct(s) measured and criteria of interest
Reliability	Internal consistency reliability; evidence on other forms of reliability (test-retest; alternate forms)
Incremental validity	Potential for (or evidence for) incremental validity beyond current screening and selection procedures
Administration mode	Proctored/unproctored; options for administration suitable for use in officer candidates
Testing time	Expected distribution of testing times (for differing administration modes)
Test range	Potential for floor or ceiling effects in officer population
Test population	How similar are previously tested population(s) to military officer applicants
Test durability	Susceptibility to effects of faking, compromise, construct irrelevant responses, coaching, context effects; content or item obsolescence over time
Motivation (applicant vs. non-applicant)	Whether previous research findings from a possibly non-motivated sample would apply to the applicant population
Applicant reactions	Level of enjoyment, effort, fatigue, boredom, confusion, face validity
Equipment requirements	Specific software, systems, hardware, equipment required

Next Steps

- Analyze and interpret results of SME judgment exercise; provide final list of assessments to recommend for validation research

- Research plan: proposed methods for examining assessment validity
  - Content and construct validity exercises previously discussed will be conducted during this contract period
  - Criterion-related validity or further validation research will be proposed for future

- Final report with recommendations: will include results from SME linkage exercises for content/construct validity

Conclusion

- The BPPT is currently in the process of evaluating the short-listed assessments

- We look forward to feedback on our decisions and next steps when we have further information to report at future meetings

Thank you!

For more information please contact:

Brenda Ellis
bellis@humrro.org
571.290.5301

OPA
OFFICE OF PEOPLE ANALYTICS