



Tailored Adaptive Personality Assessment System (TAPAS)

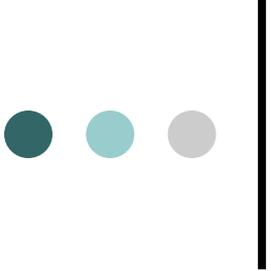
Fritz Drasgow

University of Illinois at Urbana-
Champaign



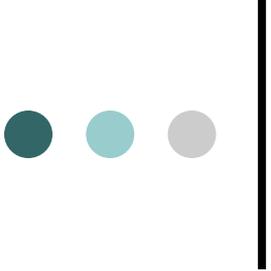
IPAC July 22, 2013





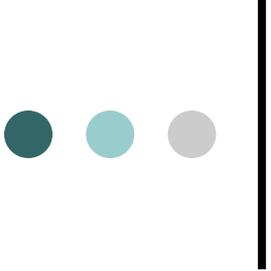
Thanks to my Colleagues:

- Sasha Chernyshenko
- Steve Stark
- Chris Nye
- Len White and Tonia Heffner, ARI
- Chris Kubisiak and Kristen Horgen, PDRI
- Deidre Knapp and our friends at HumRRO



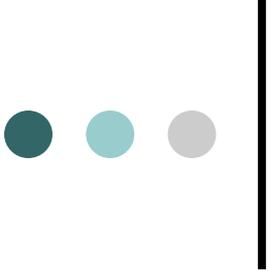
TAPAS Vision

- We wanted to build a fully customizable assessment of personality to fit an array of users' needs
- Users should be able to select:
 - any dimension from a comprehensive superset of 22 facets of the Big Five;
 - a scale length to suit their needs
 - a fake resistant response format (if faking is a problem)
 - adaptive or static
- Resulting scores can be used to predict multiple criteria or as source of feedback



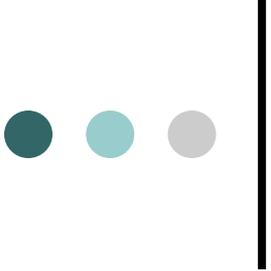
Tailored Adaptive Personality Assessment System (TAPAS)

- To this end, TAPAS incorporates recent advancements in:
 - Item response theory (IRT);
 - Models of personality; and
 - Computerized adaptive testing (CAT)
- and a fake resistant format to provide a means for operational use of personality assessment for pre-employment testing



Today,

- I'll talk about the 15 year journey that has led to today's TAPAS

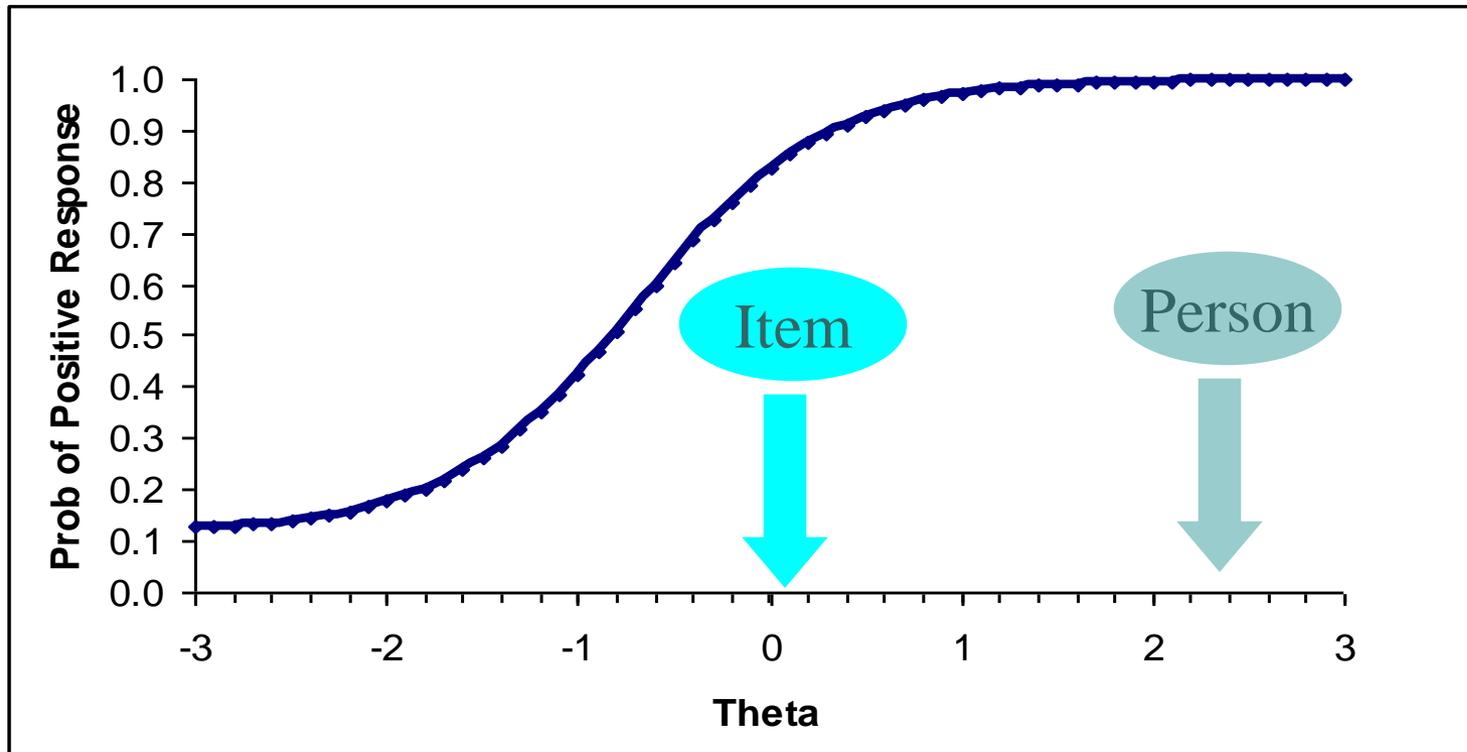


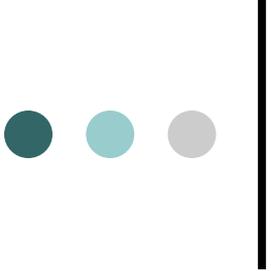
The Beginning,

- Sasha Chernyshenko and Steve Stark were doctoral students interested in fitting item response theory models to personality data
- They fit the two- and three-parameter logistic models to 16 Personality Factor (16PF) data
- The fit was not good, which was surprising because Steve Reise had already published papers about fitting IRT models to personality data

The 2PL and 3PL are Dominance Models

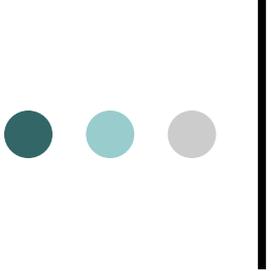
A person endorses an item if his/her standing on the latent trait, θ , is more extreme than that of the item.





Examples of Dominance Models

- Factor analysis
- Structural equations models
- Item response theory
- Classical test theory

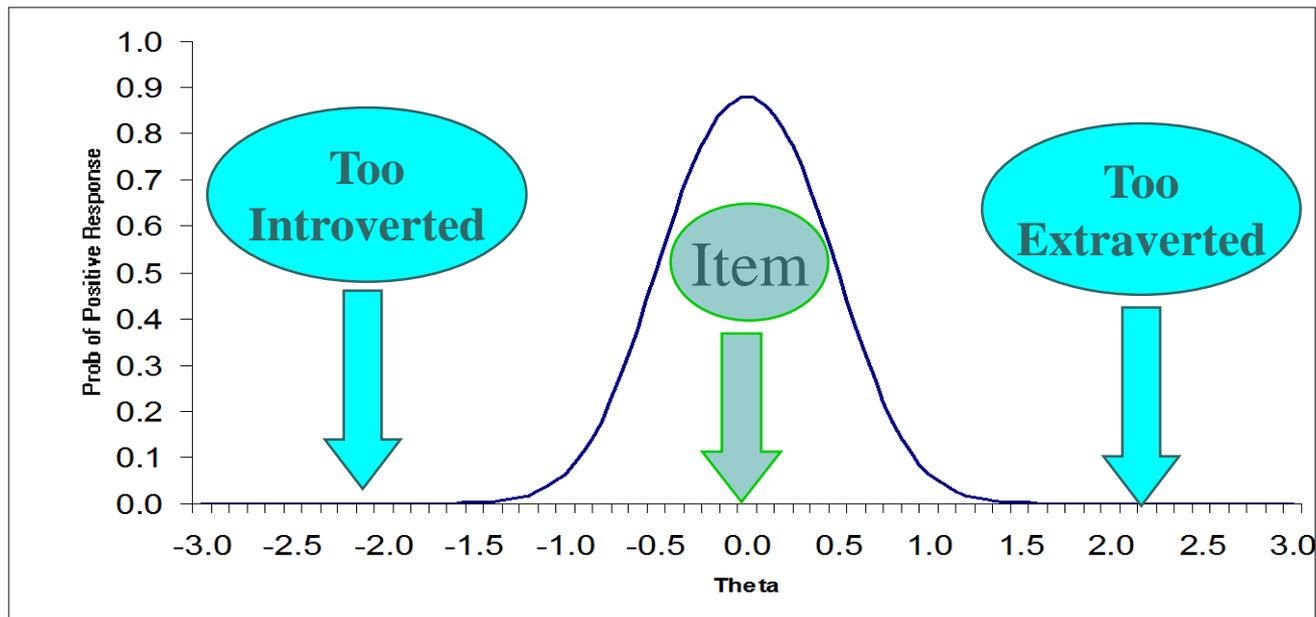


An alternative Conceptualization: Thurstone Scaling

- Thurstone assumed people endorse items reflecting attitudes close to their own feelings
- Coombs (1964) called this an **ideal point process**
- Sometimes called an **unfolding model**

Example of an Ideal Point Process

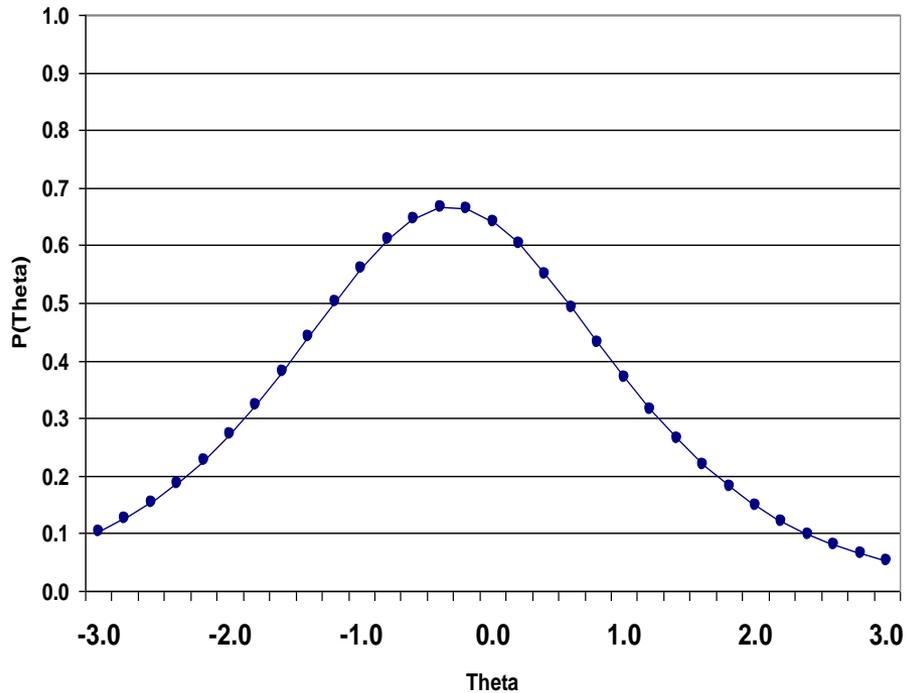
- Person endorses item if his/her standing on the latent trait is near that of the item.
 - *“I enjoy chatting quietly with a friend at a cafe.”*
 - Disagree **either** because:
 - Too **introverted** (uncomfortable in public places)
 - Too **extraverted** (chatting over coffee is boring)



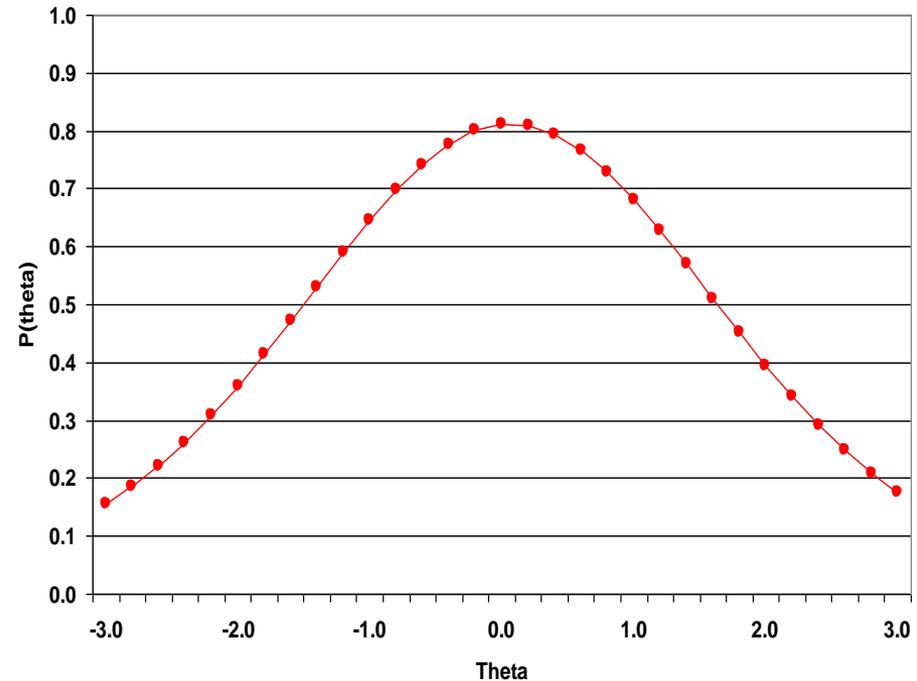


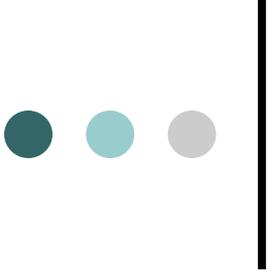
GGUM IRFs for two Personality Statements

"I enjoy chatting quietly with a friend at a café."
(Sociability)



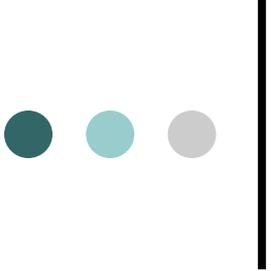
"I am about as organized as most people."
(Order)





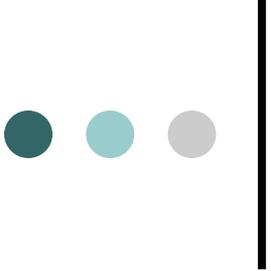
Important Point:

- The item-total correlation of intermediate ideal point items will be close to zero!
- This led Likert (1932) to assert such items were double-barreled and should be avoided



Which Process is Appropriate for Temperament Assessment?

- In a series of studies, we've
 - Examined the appropriateness of dominance process by fitting models of increasing complexity to data from several personality inventories
 - Compared the fits of dominance and ideal point models of similar complexity to several existing measures of personality
 - Compared the fits of dominance and ideal point models to sets of items not preselected to fit dominance models



Key Findings:

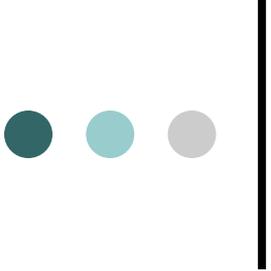
- Dominance models only fit personality data if the items are carefully pre-selected to screen out those assessing intermediate trait values
- Ideal point models fit items assessing low, intermediate, and high trait values
- For CAT to work well, we need to use a model that fits the data well and assesses trait values throughout the trait continuum → Ideal point IRT



The Generalized Graded Unfolding Model (GGUM)

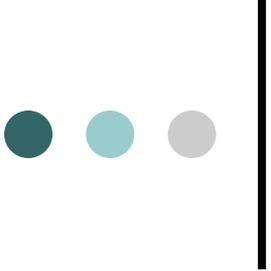
- Roberts, Donoghue, Laughlin (2000)
- Implemented in the GGUM2004 computer program
- For dichotomously scored items,

$$P[U_i = 1 | \theta_j] = \frac{\exp(\alpha_i [(\theta_j - \delta_i) - \tau_{i1}]) + \exp(\alpha_i [2(\theta_j - \delta_i) - \tau_{i1}])}{1 + \exp(\alpha_i [3(\theta_j - \delta_i)]) + \exp(\alpha_i [(\theta_j - \delta_i) - \tau_{i1}]) + \exp(\alpha_i [2(\theta_j - \delta_i) - \tau_{i1}])}$$



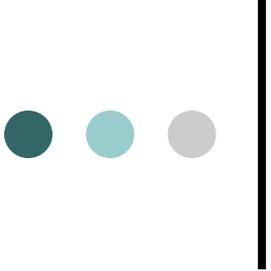
TAPAS Model of Personality

- Based on factor analysis of each of the Big Five dimensions
 - E.g., Roberts, B., Chernyshenko, O.S., Stark, S., & Goldberg, L. (2005). The structure of conscientiousness. *Personnel Psychology*
- Currently 22 facets
- Resulted from analyses of Lewis Goldberg's data set – 7 major personality inventories administered to a sample of over 700



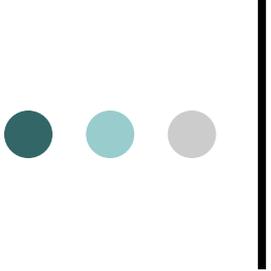
Goldberg Data Set

- A sample of 737 respondents, ranging in age from 22 to 90, all levels of education, average of 2 years of post-secondary schooling
- Over a period of 5 years, participants completed 7 personality measures



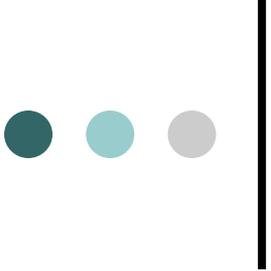
Goldberg Data Set

- Included the following scales:
 - The revised NEO Personality Inventory (NEO-PI-R), 240 items, 30 facets
 - California Psychological Inventory (CPI), 462 true-false items, 20 facets
 - Hogan Personality Inventory (HPI), 206 items, 41 “homogeneous item composites” (HICs)
 - Jackson Personality Inventory-Revised (JPI-R), 300 items, 15 scales



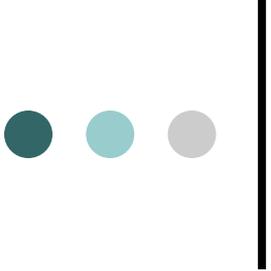
Goldberg Data Set

- Multidimensional Personality Questionnaire (MPQ), 272 items, 11 primary scales
- Abridged Big 5 Circumplex scales from the International Personality Item Pool (AB5C-IPIP), over 400 items, 45 facets
- Sixteen Personality Factor Questionnaire (16PF), 185 items, 16 primary scales



So, What Is a Comprehensive Set of Facets Underlying the Big 5?

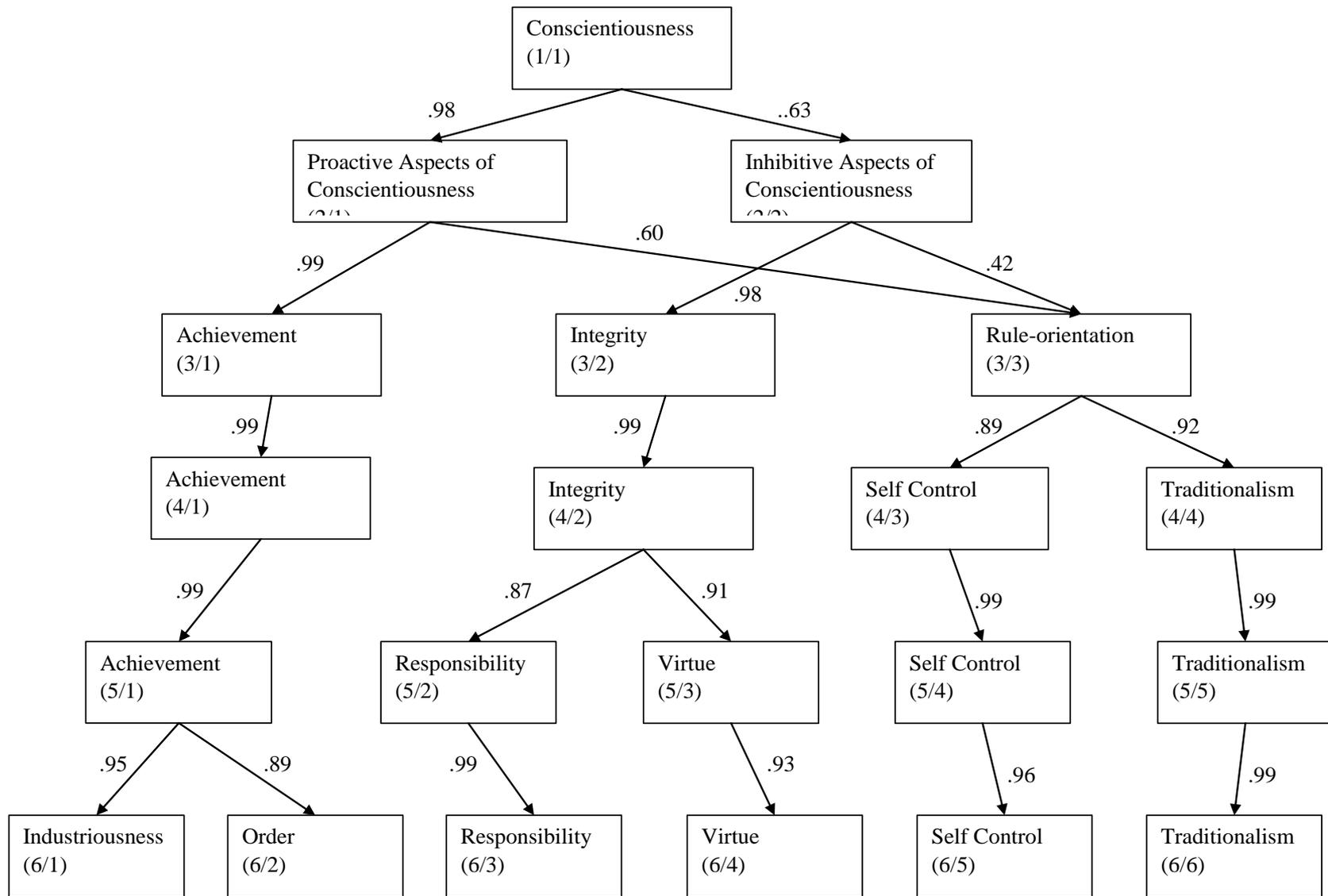
- E.g., for Conscientiousness, Roberts et al. (2005) identified all of the facets, HICs, primary scales, etc. of the seven instruments that were related to conscientiousness, ran factor analysis
- This is the method of “Standing on the shoulders of giants”...i.e., “extending science by understanding and using the research and works of great thinkers of the past”



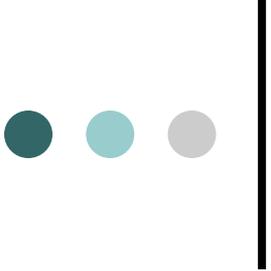
Example of TAPAS Facets

Conscientiousness

- Six facet hierarchical structure:
 - Industriousness: task- and goal-directed
 - Order: planful and organized
 - Self-control: delays gratification
 - Traditionalism: follows norms and rules
 - Social Responsibility: dependable and reliable
 - Virtue: ethical, honest, and moral

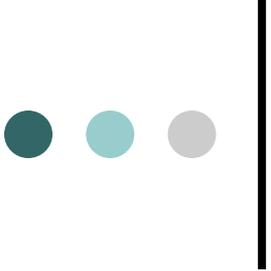


From Roberts et al. 2005



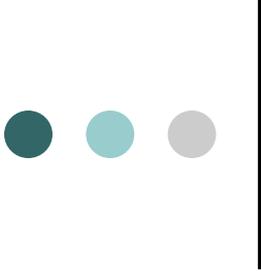
TAPAS Facets

- Conscientiousness: 6
- Emotional Stability: Adjustment, Even Tempered, Well Being
- Agreeableness: Warmth, Selflessness, Cooperation
- Extraversion: Dominance, Sociability, Excitement Seeking, Energy
- Openness: Intellectual Efficiency, Curiosity, Ingenuity, Aesthetic, Tolerance, Depth



Computerized Adaptive Testing (CAT)

- Has been used by DoD for ASVAB pre-enlistment testing for 20 years
- By selecting the next item based, in part, on the test taker's previous responses, we can adapt the difficulty level to the ability of a test taker
- We can use the same logic for personality assessment: adapt the extremity of the items administered to the trait level of the respondent



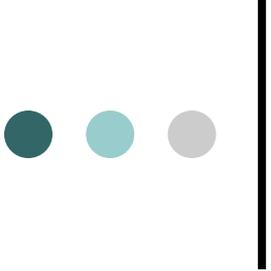
Average Correlations of True vs. Estimated Trait Values for Static vs. CAT Simulated Personality Assessments

- For 7 facets:
 - 70 item static: .84
 - 35 item CAT: .85
- For 10 facets:
 - 100 item static: .84
 - 50 item CAT: .84

Notes: items administered in MDPP format
from Stark et al., 2012, ORM



Overcoming the Faking Problem

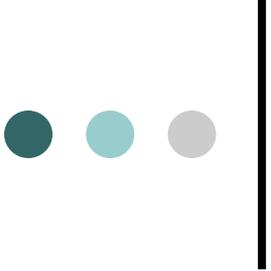


Examples of “Traditional” Items that Appear to Be Easily Faked

What is the the positively keyed response to these items? Do you “Agree” or “Disagree”?

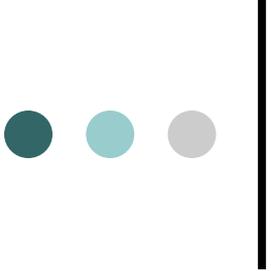
- I get along well with others. (A+)
- I try to be the best at everything I do. (C+)
- I insult people. (A-)
- My peers call me “absent minded.” (C-)

Because these items consist of individual statements, they are commonly referred to as “**single stimulus**” items.



Forced Choice Formats

- There has been a long interest in multidimensional forced choice formats:
 - Edwards (1954) Personal Preference Schedule
 - White & Young's Assessment of Individual Motivation (AIM)
 - Christiansen et al. (1998)
 - Jackson et al. (2000)
 - SHL's OPQ

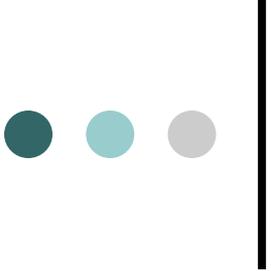


Why Forced-Choice Items?

- Correcting or detecting faking doesn't seem to work well:

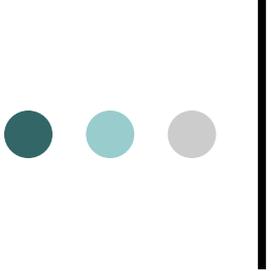
- Validity doesn't increase after corrections (Schmitt & Oswald, 2006)
- Scales to detect faking are nowhere close to 100% effective and it is not clear what to do with “disqualified” applicants
- Warnings may not be very effective in settings with coached applicants

- **Solution** – Discourage faking through the use of forced-choice response formats



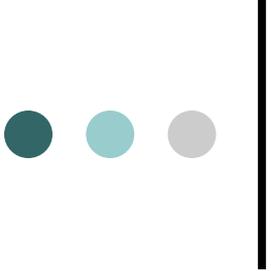
Jackson et al. (2000)

- Compared traditional “single stimulus” personality items to “quartets” formed by:
 - First placing pairs of statements from different dimensions into dyads...statements in dyads had similar endorsement rates (as single stimulus items) and social desirability ratings
 - Then combining high-desirability dyads with low-desirability dyads to form a quartet
 - Respondents chose the statement “Most characteristic of me” and “Least characteristic of me” from each quartet



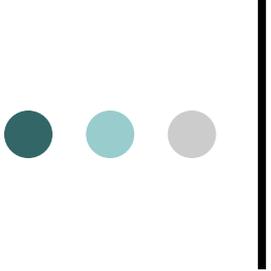
Jackson et al. (2000)

- Respondents were given the quartets under two conditions:
 - Answer honestly
 - Imagine you're a job applicant who really wants to get hired
- Mean scores were higher in the job applicant condition for the quartet format by .30 SD but were .95 SD higher in the applicant condition for the single stimulus items
- → Fake Resistant (but not fake-proof)



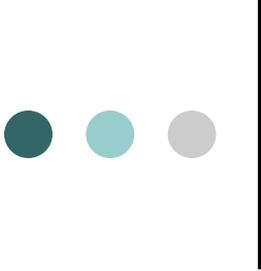
Heggestad et al. (2005)

- Also examined the multidimensional forced-choice (MFC) format as a way to combat faking
- Compared an MFC format to two Likert-type measures (NEO, IPIP) under Honest and Fake Good conditions
- Also used “Most like me” and “Least like me” ratings
- Created quartets by matching on statement extremity on the dimension it assesses, but not social desirability



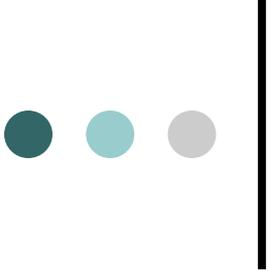
Heggestad et al. (2005)

- Effect sizes for Fake Good vs. Honest conditions were generally larger for the single stimulus format
- But, for Conscientiousness, the effect size was 1.23 for the single stimulus format vs. 1.20 for the MFC format
- → Not too much Fake Resistance



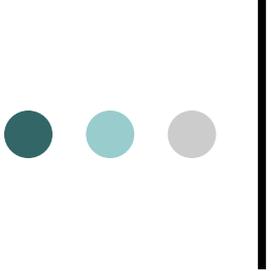
TAPAS Multidimensional Pairwise Preference (MDPP) Format

- Create items by pairing stimuli that are similar in social desirability, but represent different dimensions
- “Which is more like you?”
 - ___ I get along well with others. (A+)
 - ___ I always get my work done on time. (C+)



Our Experience with Faking

- First study at recruit training centers:
 - Matched statements on social desirability
 - Found score inflation for 2AFC just as large as single statements
- Second study:
 - Matched statements on social desirability and their IRT extremity parameters
 - Found greatly improved resistance to faking for 2AFC



Example MDPP Items

- *“For each of the following pairs, select the statement that is more like you.”*
 - ___1a) People come to me when they want fresh ideas. (+Ingenuity)
___1b) Most people would say that I’m a “good listener”. (+Warmth)
 - ___2a) I almost always complete assignments on time. (+Industrious)
___2b) I generally perform well under pressure. (+Adjustment)
 - ___3a) I set high goals and work to meet them. (+ Industrious)
___3b) I get along well with other people. (+Cooperation)

IRT Model for Scoring MDPP Tests

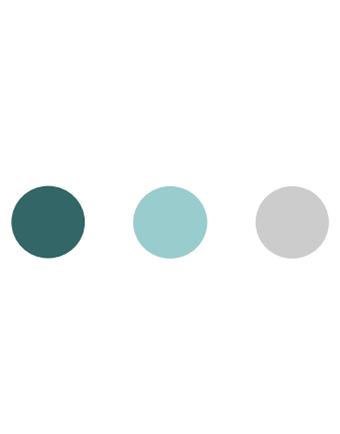
(Stark, 2002; Stark, Chernyshenko, & Drasgow, 2005)

s = 1st statement
t = 2nd statement

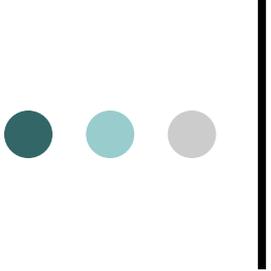
1 = Agree
0 = Disagree

$$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) = \frac{P_{st}\{1,0\}}{P_{st}\{1,0\} + P_{st}\{0,1\}} = \frac{P_s\{1\}P_t\{0\}}{P_s\{1\}P_t\{0\} + P_s\{0\}P_t\{1\}}$$

- Respondent evaluates each statement in pair *separately* and makes independent decisions about endorsement.
- Statement endorsement probabilities $P\{0\}$ and $P\{1\}$ computed using the GGUM model
- **Trait scores** are obtained via Bayes modal estimation involving k-dimensional minimization

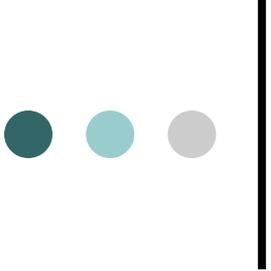


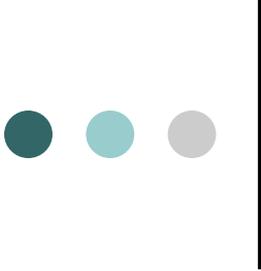
So, Does it Work?



TAPAS Research

- US Army and Air Force began implementation of TAPAS for enlistment screening at six Military Enlistment Processing Stations (MEPS) on June 8, 2009 and at all MEPS in September 2009
- 15 facets, 120 items, median response time of about 20 minutes
- Army applicants were told that their scores might affect their enlistment eligibility
- Air Force given “for research only” instructions
- Will TAPAS predict attrition and “will-do” behaviors?

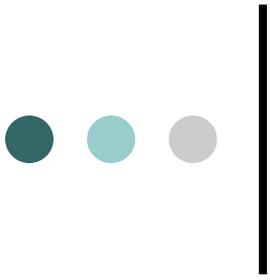
- 
- Is there score inflation for Army applicants?



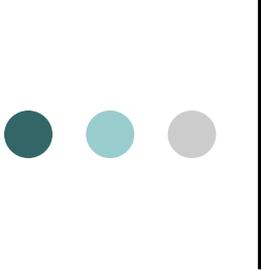
Descriptive Statistics for TAPAS CAT Scores in Regular Army and Air Force Samples

TAPAS Facet	Army		Air Force		Army - Air Force
	Mean	SD	Mean	SD	d
Achievement	0.16	0.49	0.13	0.50	0.07
Adjustment	0.01	0.57	-0.04	0.58	0.08
Cooperation	-0.07	0.38	-0.04	0.38	-0.07
Dominance	0.03	0.57	-0.05	0.59	0.13
Even Tempered	0.15	0.46	0.19	0.46	-0.08
Attention Seeking	-0.20	0.52	-0.19	0.52	-0.01
Selflessness	-0.21	0.44	-0.22	0.45	0.02
Intellectual Efficiency	-0.02	0.59	0.01	0.60	-0.05
Non-delinquency	0.07	0.47	0.14	0.46	-0.14
Order	-0.40	0.54	-0.43	0.56	0.06
Physical Conditioning	0.03	0.61	0.06	0.64	-0.04
Self Control	0.05	0.54	0.03	0.54	0.04
Sociability	-0.05	0.58	-0.06	0.59	0.01
Tolerance	-0.21	0.55	-0.24	0.56	0.05
Optimism	0.12	0.45	0.14	0.45	-0.04

Note. Sample Sizes: Regular Army = 86,962; Air Force = 30,658



- Is there adverse impact?



Female-Male Comparisons of TAPAS Scale Scores among U.S. Army Applicants at MEPS

TAPAS Facet	Females		Males		F - M
	Mean	SD	Mean	SD	d
Achievement	0.17	0.46	0.15	0.48	0.04
Adjustment	-0.14	0.56	0.02	0.57	-0.29
Cooperation	-0.08	0.37	-0.06	0.38	-0.03
Dominance	-0.02	0.56	0.05	0.58	-0.12
Even Tempered	0.11	0.47	0.16	0.46	-0.11
Attention Seeking	-0.24	0.51	-0.19	0.52	-0.11
Selflessness	-0.06	0.43	-0.23	0.43	0.37
Intellectual Efficiency	-0.12	0.54	0.00	0.59	-0.21
Non-delinquency	0.13	0.44	0.06	0.46	0.15
Order	-0.33	0.55	-0.41	0.53	0.15
Physical Conditioning	-0.16	0.59	0.08	0.61	-0.40
Self Control	0.05	0.54	0.05	0.53	0.00
Sociability	-0.03	0.58	-0.04	0.58	0.02
Tolerance	-0.07	0.53	-0.25	0.56	0.34
Optimism	0.12	0.46	0.13	0.45	-0.04

Note. F = Female (N = 23,170); M = Male (N = 97,165); d = mean difference (F-M). Sample includes applicants for Regular Army, U. S. Army National Guard, and U. S. Army Reserve.

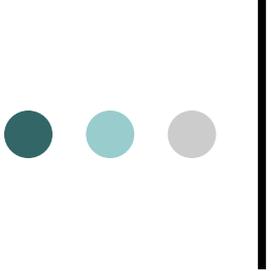
Female-Male Comparisons of TAPAS Scale Scores among U.S. Army Applicants at MEPS

White-Black Comparisons for U.S. Army

TAPAS Facet	White		Black		W - B
	Mean	SD	Mean	SD	d
Achievement	0.16	0.48	0.13	0.46	0.07
Adjustment	0.00	0.57	-0.05	0.55	0.09
Cooperation	-0.07	0.38	-0.07	0.37	0.01
Dominance	0.04	0.58	0.05	0.51	-0.02
Even Tempered	0.15	0.46	0.15	0.46	0.01
Attention Seeking	-0.19	0.53	-0.23	0.49	0.08
Selflessness	-0.20	0.44	-0.15	0.43	-0.12
Intellectual Efficiency	-0.01	0.59	-0.07	0.54	0.10
Non-delinquency	0.08	0.46	0.09	0.45	-0.03
Order	-0.42	0.54	-0.29	0.51	-0.24
Physical Conditioning	0.06	0.61	-0.06	0.57	0.19
Self Control	0.03	0.54	0.12	0.53	-0.18
Sociability	-0.03	0.58	-0.06	0.54	0.05
Tolerance	-0.24	0.56	-0.12	0.51	-0.22
Optimism	0.13	0.45	0.13	0.43	0.00

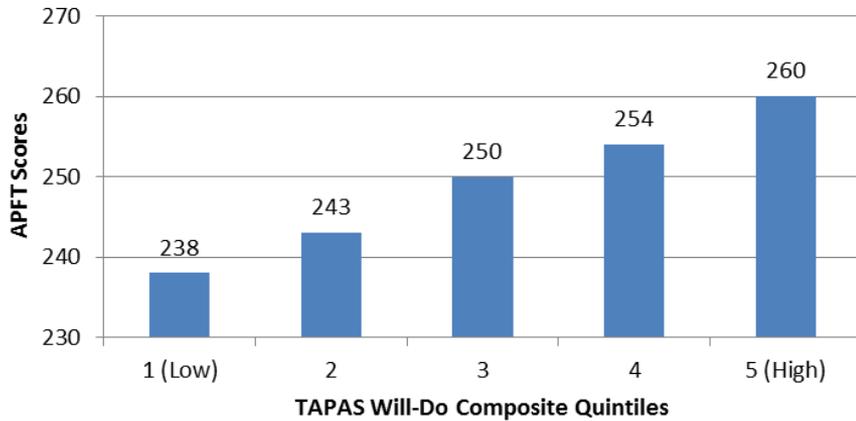
Note. W = White (N = 97,202); B = Black (N = 19,945).

15

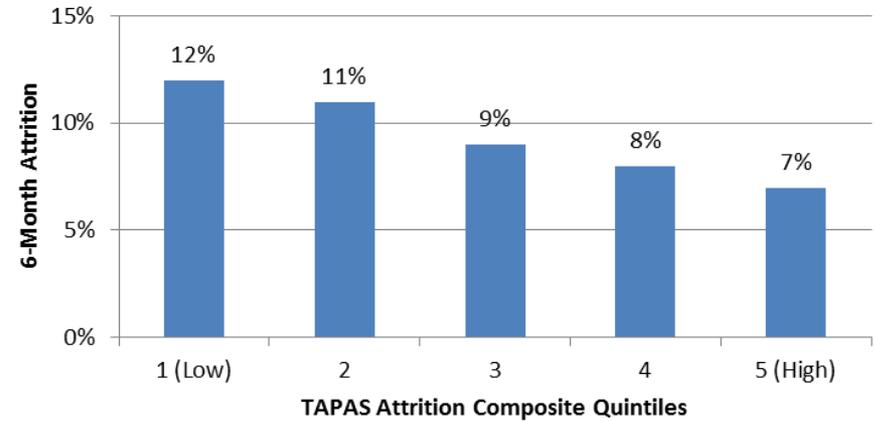
- 
- Does TAPAS predict performance?

MEPS TAPAS Results for Army IMT Outcomes

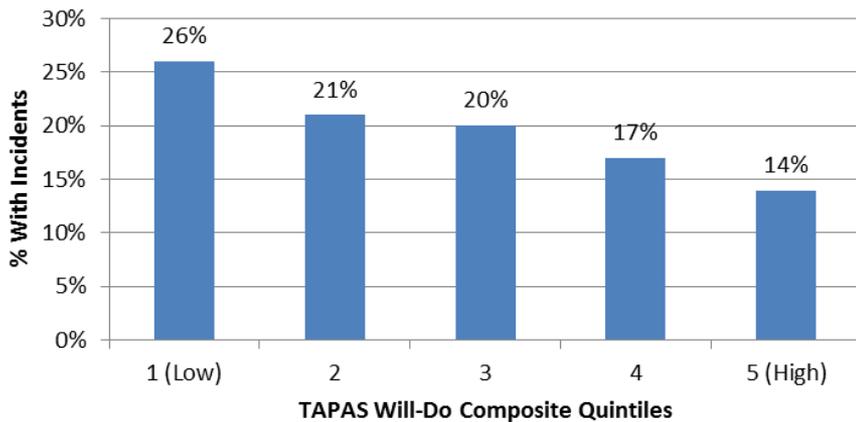
APFT (n =4382)



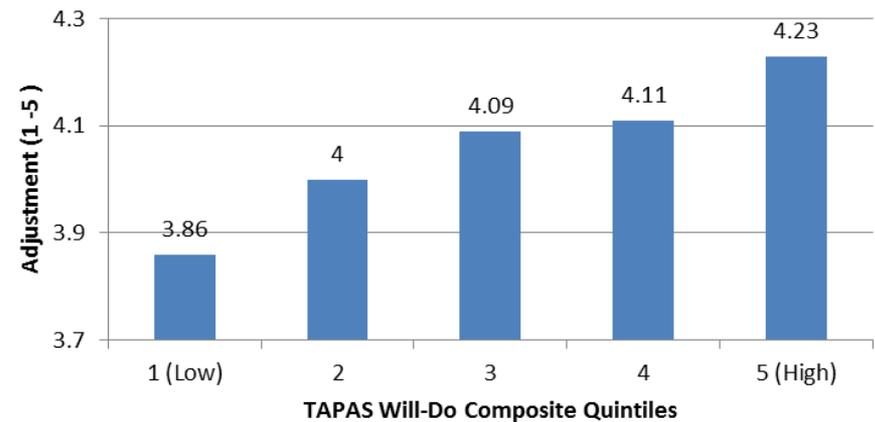
6-Month Attrition (n =16,458 RA)



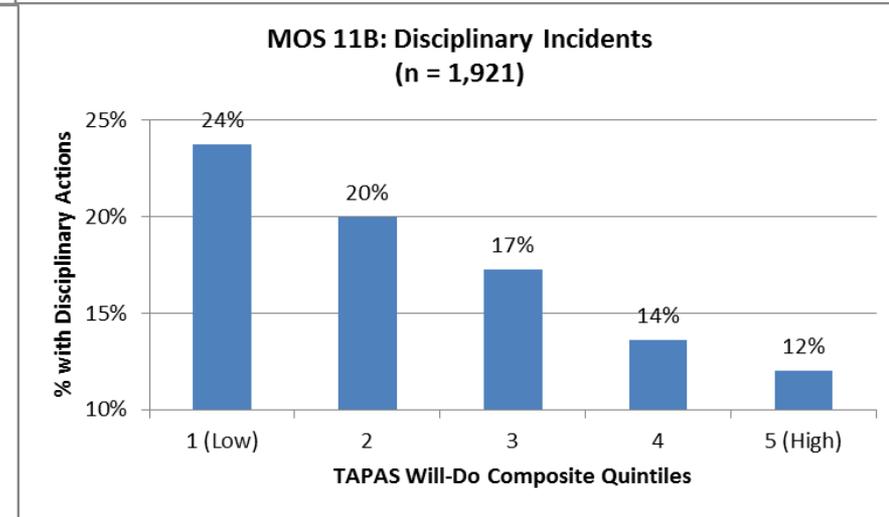
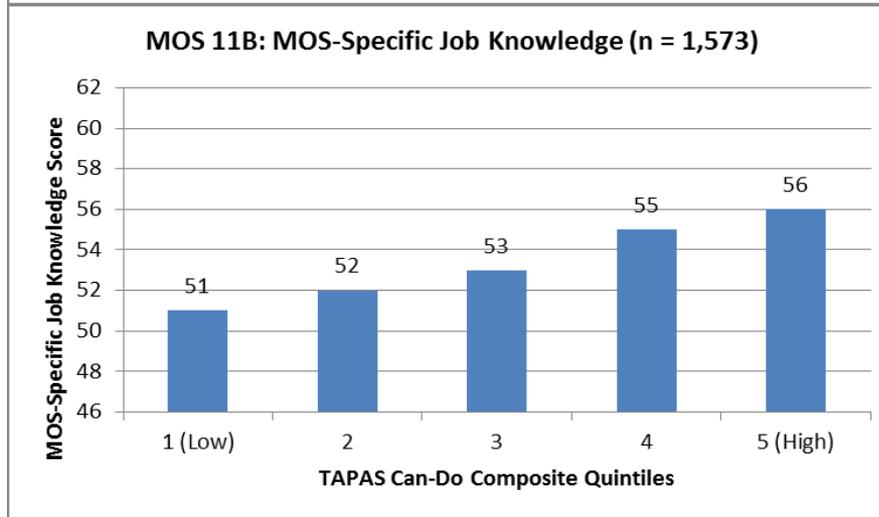
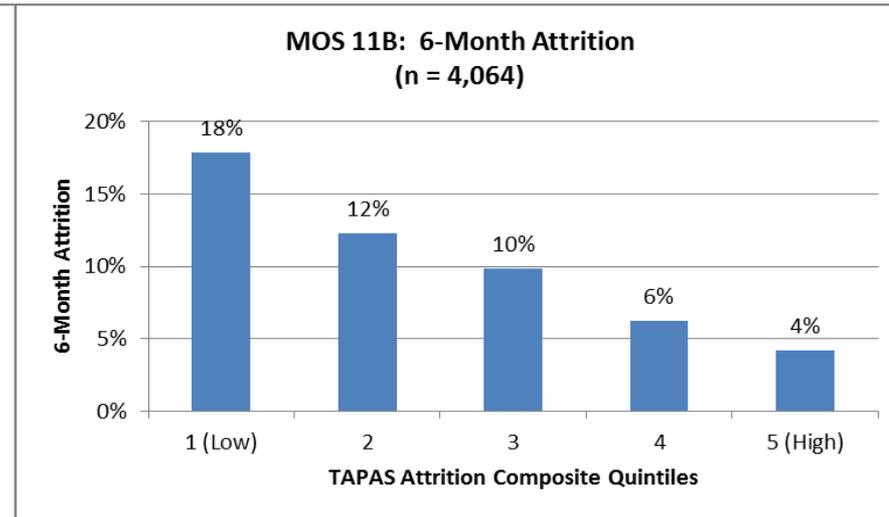
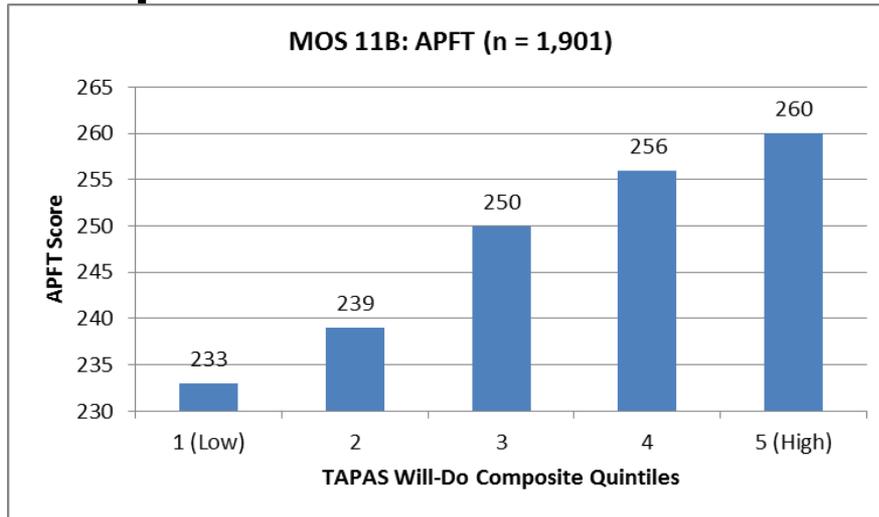
Disciplinary Incidents (n=2885)



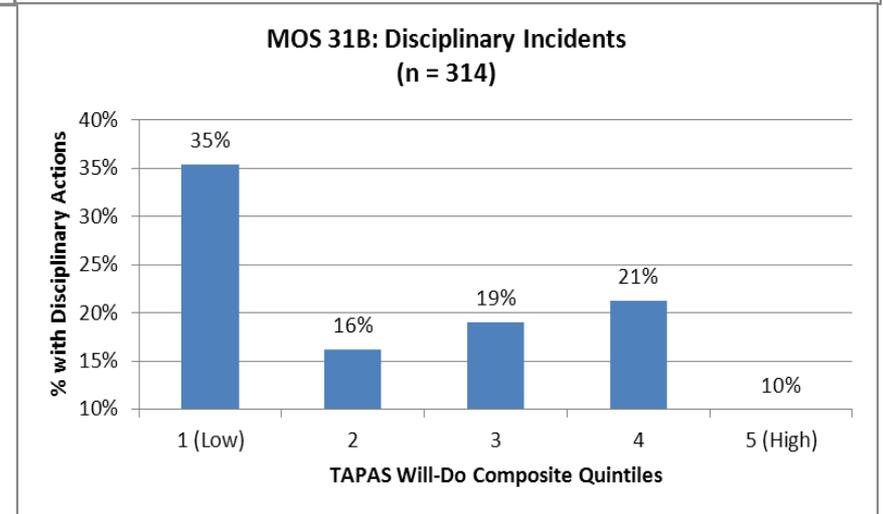
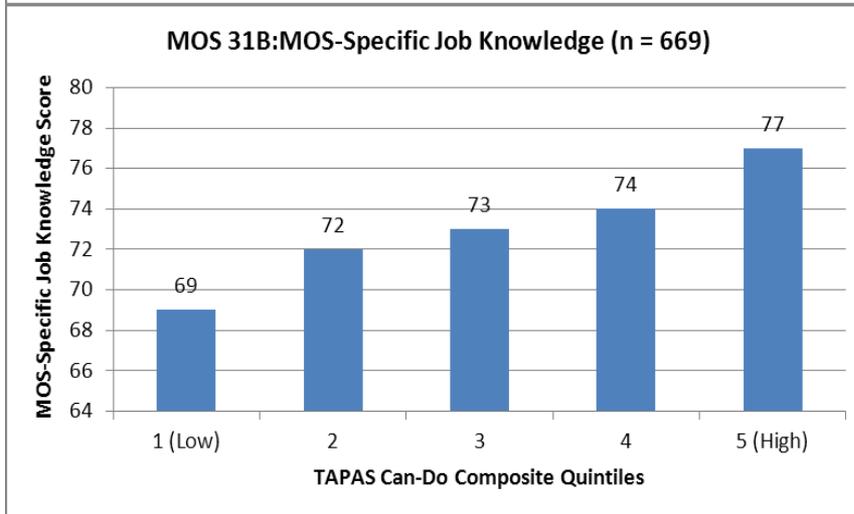
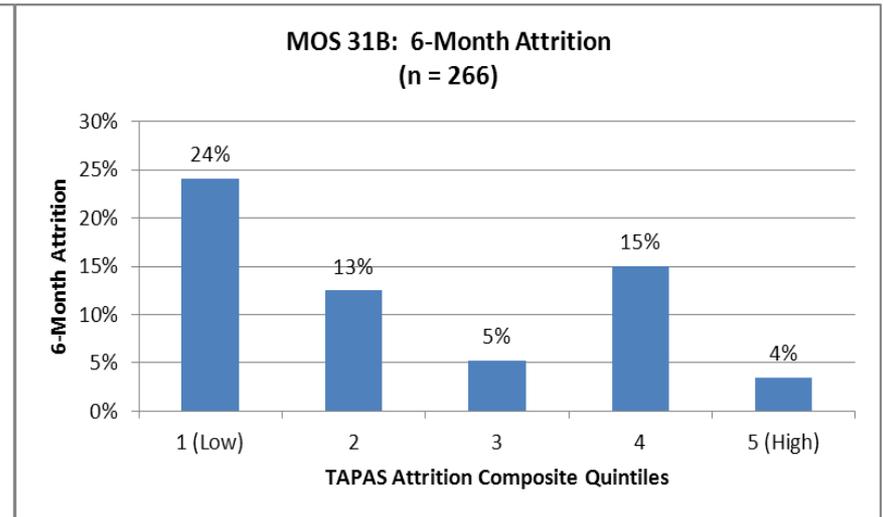
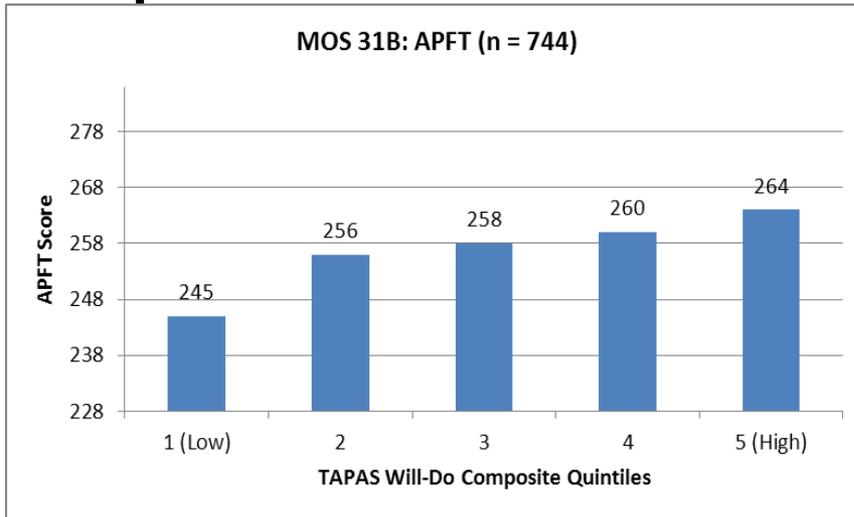
Self-Reported Adjustment (n=4332)



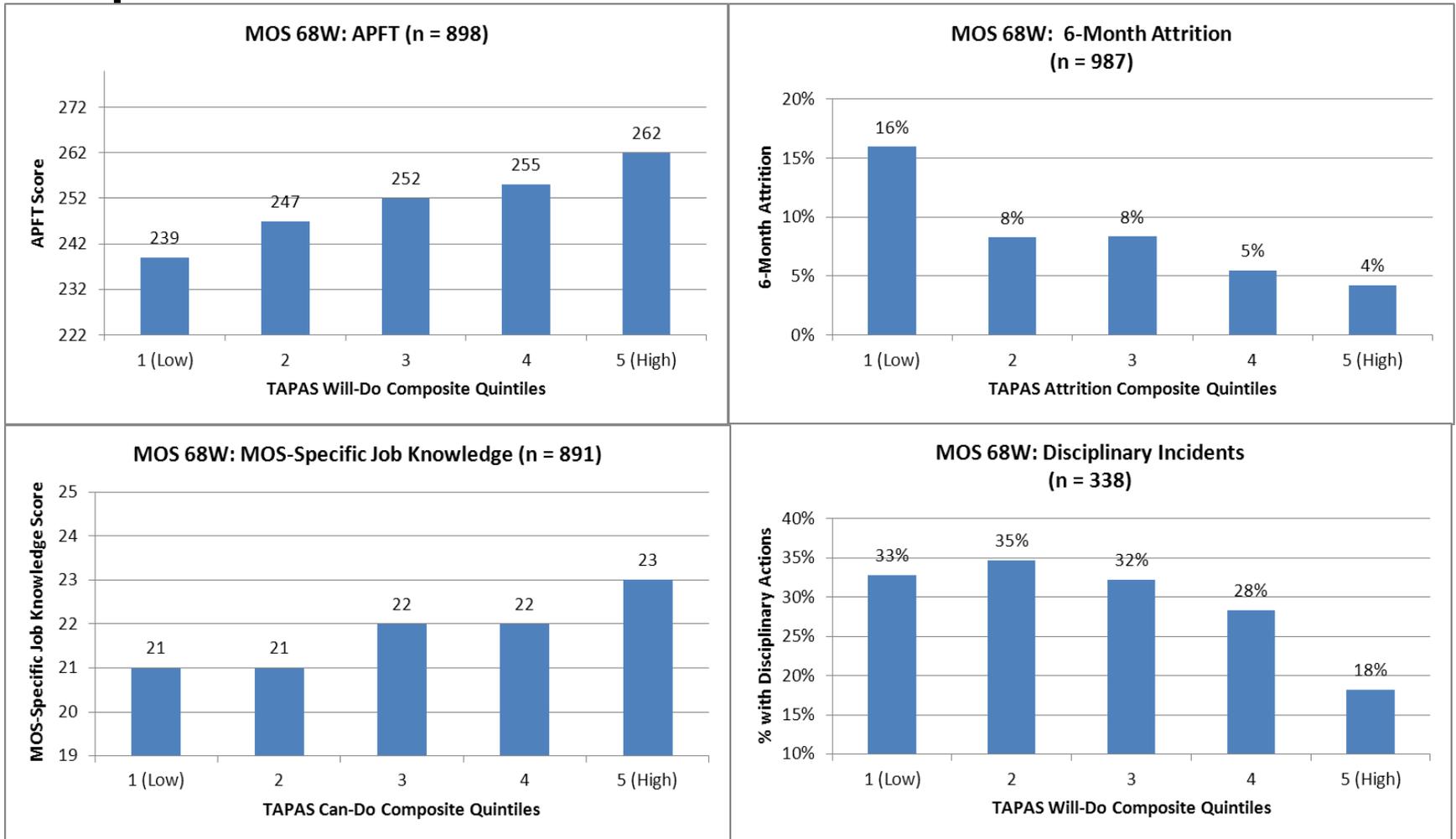
TAPAS Composite Quintile Plots for APFT scores, 6-Month Attrition, MOS-Specific Job Knowledge Scores, and Disciplinary incidents in MOS 11B (Infantry).



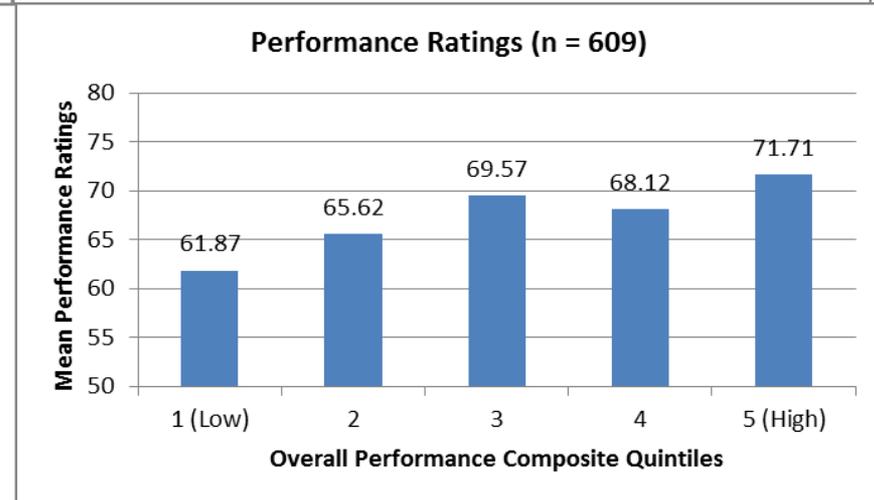
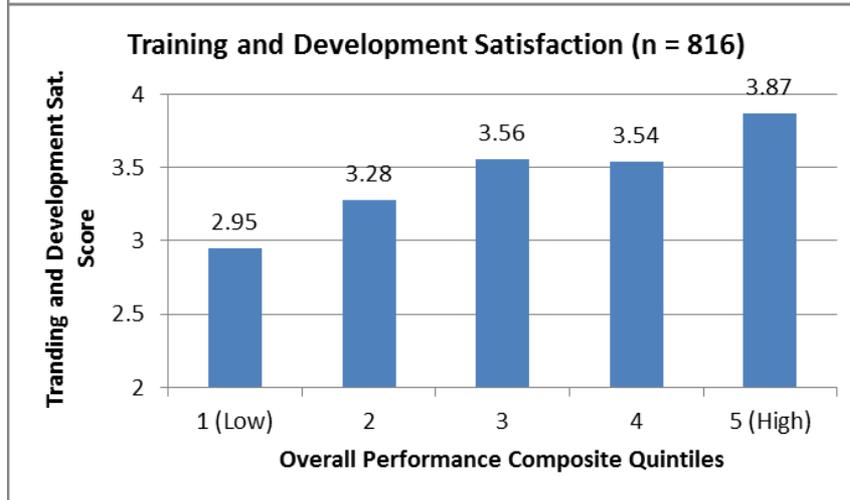
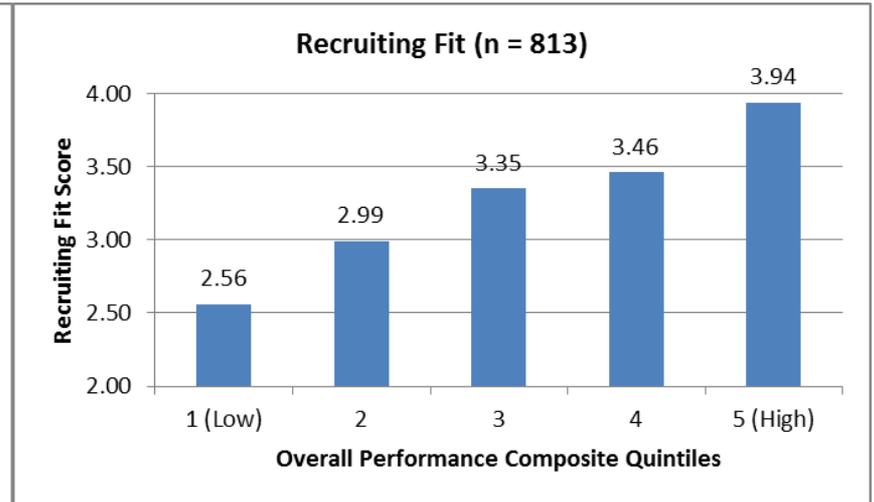
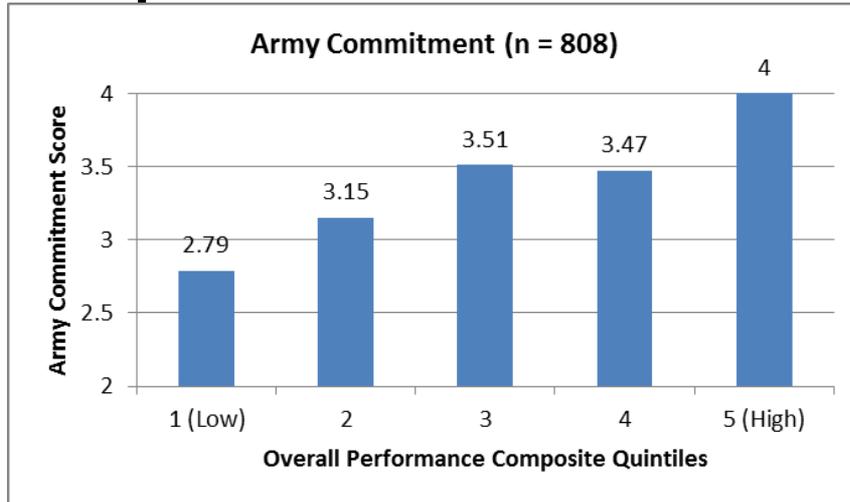
TAPAS Composite Quintile Plots for APFT scores, 6-Month Attrition, MOS-Specific Job Knowledge Scores, and Disciplinary incidents in MOS 31B (Military Police).

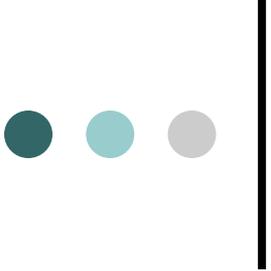


TAPAS Composite Quintile Plots for APFT scores, 6-Month Attrition, MOS-Specific Job Knowledge Scores, and Disciplinary incidents in MOS 68W (Combat Medics).



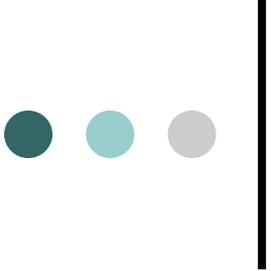
Quintile Plots of the Relationships between the Overall Performance Composite and Army Commitment, Recruiting Fit, Training and Development Satisfaction, and Performance Ratings for Recruiters





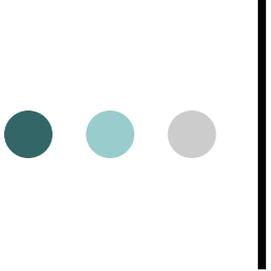
In Sum,

- Our goal has been to produce an easily customizable assessment tool to meet the needs of diverse users and researchers
- To this end, we've used the latest in
 - Psychometric theory
 - Computer technology
 - Personality theory



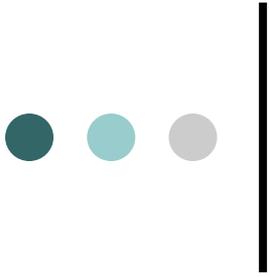
In Sum,

- Our findings to date have been positive: we are able to use operationally administered scores to predict
 - Attrition
 - Motivationally driven aspects of performance, e.g., commitment, person-job fit, physical fitness, disciplinary incidents, well being



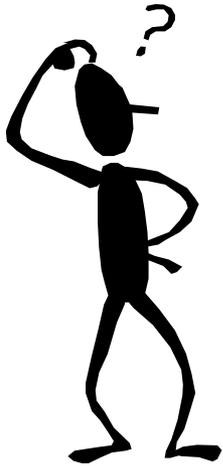
Limitations

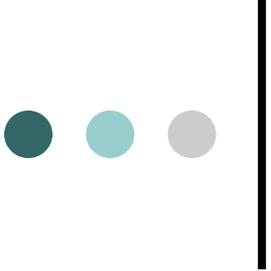
- Our validation work has been limited to the Army, no work yet in the civilian world...but...
- Results for “can-do” aspect of performance have been weaker than “will-do”



Questions?

Thank you for the opportunity to talk about our work!



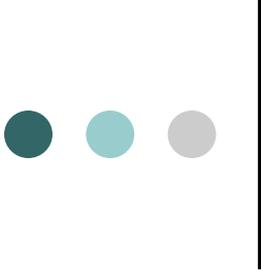


The Big Five Defined

Extraversion – tendency to be sociable, assertive, active, upbeat, talkative

“Meeting new people is enjoyable to me”

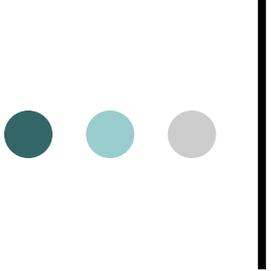
“I am a ‘take charge’ type of person”
(urgency)



TAPAS Facet Dimensions

Extraversion

- Dominance - Dominant, leading, commanding, authoritative, influential vs. weak, follower, feeble
- Sociability - friendly, outgoing, companionable, talkative, chatty, conversational
- Excitement seeking - fun seeking, entertaining, loud, flamboyant, showy vs. boring, dull, unexciting, uninteresting, shy restrained, undemonstrative

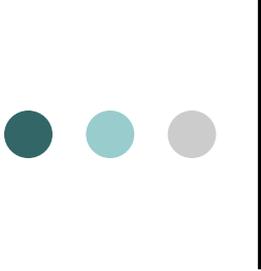


The Big Five Defined

Agreeableness – tendency to be altruistic, trusting, sympathetic, and cooperative

“I usually see the good side of people”

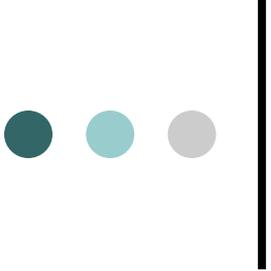
“I forgive others easily”



TAPAS Facet Dimensions

Agreeableness

- Warmth - Kind, tender, affectionate, compassionate, warm, positive toward others, encouraging
- Selflessness - Generous, giving, charitable, helpful, ready to lend a hand vs. tightfisted, stingy, cheap, frugal, thrifty
- Cooperation - accommodating, supporting, compliant vs. resistant, uncooperative, stubborn, inflexible



The Big Five Defined

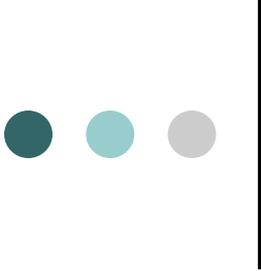
Emotional Stability (Neuroticism) -

disposition to be calm, optimistic, and well adjusted

“I can become annoyed at people quite easily”

“I worry a lot” (anxiety)

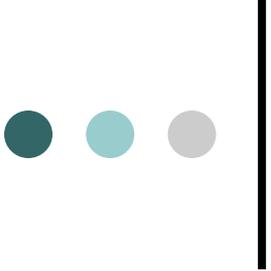
“I often feel blue” (depression)



TAPAS Facet Dimensions

Emotional Stability

- Adjustment - Confident, self-assured, no doubts vs. anxious, nervous, worried, fearful, distressed
- Even tempered - Calm, composed, poised vs. aggressive, antagonistic, hot-headed, quarrelsome, irritable
- Well being - Happy, joyful, cheerful, positive, joyful, optimistic vs. depressed, miserable, dejected, unhappy, sad

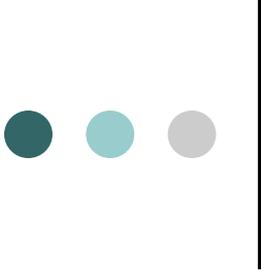


The Big Five Defined

Openness to Experience – tendency to be imaginative, attentive to inner feelings, have intellectual curiosity, and independence of judgment

“I like to work with difficult concepts and ideas”

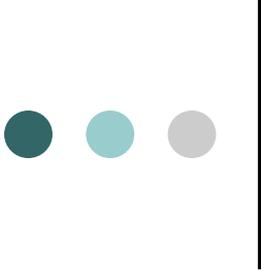
“I enjoy trying new and different things”



TAPAS Facet Dimensions

Openness to Experience

- Intellectual efficiency - able to process information quickly, knowledgeable, astute
- Curiosity - inquisitive, perceptive, questioning, learning
- Ingenuity - creative, inventive, clever, innovative
- Aesthetic - enjoy observing or creating various forms of artistic, musical, or architecture
- Tolerance - interested in travel and learning about different cultures, often attend cultural events or meet and befriend people from around the world
- Depth – seek to understand the meaning of one's life, improve oneself



Performance of MDPP CAT Algorithm in Simulation Studies

Average Correlation Across Dimensions

			Nonadaptive					Adaptive				
ρ_{gen}	% Unidim.	Items Per Dimension	3-d	5-d	7-d	10-d	25-d	3-d	5-d	7-d	10-d	25-d
0	5	5	.72	.68	.72	.73	.75	.84	.83	.85	.84	.84
		10	.83	.82	.84	.84	.85	.90	.90	.90	.90	.89
		20	.91	.90	.91	.92	.92	.94	.94	.94	.94	.94
	10	5	.71	.68	.72	.72	.75	.85	.84	.85	.84	.83
		10	.83	.82	.84	.84	.84	.90	.90	.90	.90	.89
		20	.91	.90	.91	.92	.92	.93	.93	.94	.94	.95
		5	.71	.69	.70	.71	.73	.85	.84	.84	.84	.84
		20	.82	.80	.83	.83	.84	.90	.90	.91	.91	.89
	20	.90	.90	.91	.91	.92	.94	.93	.94	.94	.95	

With tests up to 25d, very good rank order recovery of trait scores with 5% to 10% unidimensional pairings and 10 “items per dimension”