Computer Adaptive Assessment of Personality

FRITZ DRASGOW
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
Thanks to my Colleagues:

Sasha Chernyshenko
Steve Stark
Chris Nye
Len White and Tonia Heffner, U.S. Army Research Institute
Chris Kubisiak and Kristen Horgen, PDRI
Deidre Knapp and our friends at HumRRO
Outline

1. Latent structure of personality
2. Item response theory model for personality
3. Fake resistant response format
4. Using CAT to reduce length
5. Putting it all together: TAPAS
6. Predictive validity findings
The Beginning,

15 years ago, my doctoral students Sasha Chernyshenko and Steve Stark were 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.

What was the explanation?
Outline

1. Latent structure of personality
2. Item response theory model for personality
3. Fake resistant response format
4. Using CAT to reduce length
5. Putting it all together: TAPAS
6. Predictive validity findings
Latent Structure of Personality

The most widely used versions of item response theory assume unidimensionality: all items assess a single, common latent trait.

As applied to ability tests, what really seems to be required is a strong general trait, but some violation of unidimensionality is not a problem.
Latent Structure of Personality

We applied factor analysis to a data set collected by Lewis Goldberg’s – 7 major personality inventories administered to a sample of over 700


This resulted in 22 facets underlying the Big 5 that are more nearly unidimensional
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 major 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
- 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

Factor analysis of all these subscores yielded 6 factors...All 7 instruments assessed some of the factors, none assessed all.

Method: “Standing on the shoulders of giants”...i.e., extending science by understanding and using the research and works of great thinkers of the past
Conscientiousness Facets

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
Facets Assessed by TAPAS

*Conscientiousness*: Industriousness, Order, Self-control, Traditionalism, Social Responsibility, Virtue

*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
Outline

1. Latent structure of personality
2. Item response theory model for personality
3. Fake resistant response format
4. Using CAT to reduce length
5. Putting it all together: TAPAS
6. Predictive validity findings
Measurement Models for Personality

“Dominance” models include factor analysis and the two- and three-parameter logistic models

The key feature of “dominance” models:

- A person is increasingly likely to endorse an item (or answer it correctly) as his/her standing on the latent trait, theta, is increasingly higher than the extremity (or difficulty) parameter of the item.

For cognitive ability measurement, a smart test taker will answer easy and moderately difficult items correctly...the test taker “dominates” the items.
The 2PL and 3PL are Dominance Models
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 it’s called an unfolding model.
Example of an Ideal Point Process

Key idea: a person is most likely to endorse an item if his/her standing on the latent trait is near the extremity parameter of the item. Consider the extroversion item,

- “I enjoy chatting quietly with a friend at a cafe.”
- A person might disagree either because:
  - Too introverted (uncomfortable in public places); “disagree from below”
  - Too extroverted (chatting over coffee is boring); “disagree from above”
Example of an Ideal Point Process
Estimated Item Response Functions 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.
Outline

1. Latent structure of personality
2. Item response theory model for personality
3. Fake resistant response format
4. Using CAT to reduce length
5. Putting it all together: TAPAS
6. Predictive validity findings
Examples of “Traditional” Items that Appear to Be Easily Faked

What is 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.
Two Alternative Forced Choice Format

Create items by pairing statements:

“Which is more like you?”

◦ ___ I get along well with others. (A+)
◦ ___ I always get my work done on time. (C+)
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 based on “Fake Good” scales (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
Conflicting Earlier Research

- Jackson et al. (2000) found reasonably good fake resistance for forced-choice scales
- Heggestad (2005) found little to no fake resistance
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
So...

Create items by pairing stimuli that are similar in social desirability and trait extremity, but represent different dimensions.
Example 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)
Forced Choice Formats and Ipsativity

A 100 item forced choice instrument assessing five dimensions yields a total score (i.e., the sum across the five dimensions) of 100 for all respondents; the scores on the separate dimensions can vary, but the sum of the dimension scores is 110. This is “ipsativity”.

Ipsative measures have some unfortunate statistical properties, e.g., dimension scores tend to be negatively correlated (because a high score on one dimension means other dimension scores have to be low). A person’s score is only high relative to his/her other scores; don’t know if it’s high relative to other people ➔ ideographic measurement.

We were able to obtain “normative” scores by means of item response theory scoring. Typically, we see positive correlations across personality dimensions, but not as large as with single statement, Likert format, items.
Outline

1. Latent structure of personality
2. Item response theory model for personality
3. Fake resistant response format
4. Using CAT to reduce length
5. Putting it all together: TAPAS
6. Predictive validity findings
Computerized Adaptive Testing (CAT)

Has been used by the U. S. Department of Defense for pre-enlistment testing for 25 years

By selecting the next item based, in part, on the test taker’s previous responses, a CAT adapts the difficulty level to the ability of a test taker. The computer branches to an easier item after an incorrect response and a more difficult item after a correct response.

Item response theory can be used to optimally select items.

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

Note: Stark et al., 2012, *Organizational Research Methods*
Outline

1. Latent structure of personality
2. Item response theory model for personality
3. Fake resistant response format
4. Using CAT to reduce length
5. Putting it all together: TAPAS
6. Predictive validity findings
Putting It All Together: The Tailored Adaptive Personality Assessment System (TAPAS)

My colleagues and I developed an instrument we call TAPAS.

It uses:
- The 22 dimension model of personality (typically, 9 to 15 dimensions are selected by a user)
- The two-alternative forced choice format
- An ideal point item response theory model with Bayesian trait estimates
- Adaptive item selection
Outline

1. Latent structure of personality
2. Item response theory model for personality
3. Fake resistant response format
4. Using CAT to reduce length
5. Putting it all together: TAPAS
6. Predictive validity findings
So, Does it Work?
TAPAS Research

Development began in 2004

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

13 to 15 facets, 120 items, median response time of about 22 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?
Fake Resistance

Is there score inflation for Army applicants?
Descriptive Statistics for TAPAS CAT Scores in Regular Army and Air Force Samples

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

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

Is there adverse impact?
### Female-Male Comparisons of TAPAS Scale Scores among U.S. Army Applicants at MEPS

<table>
<thead>
<tr>
<th>TAPAS Facet</th>
<th>Females</th>
<th>Males</th>
<th>F - M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Achievement</td>
<td>0.17</td>
<td>0.46</td>
<td>0.15</td>
</tr>
<tr>
<td>Adjustment</td>
<td>-0.14</td>
<td>0.56</td>
<td>0.02</td>
</tr>
<tr>
<td>Cooperation</td>
<td>-0.08</td>
<td>0.37</td>
<td>-0.06</td>
</tr>
<tr>
<td>Dominance</td>
<td>-0.02</td>
<td>0.56</td>
<td>0.05</td>
</tr>
<tr>
<td>Even Tempered</td>
<td>0.11</td>
<td>0.47</td>
<td>0.16</td>
</tr>
<tr>
<td>Attention Seeking</td>
<td>-0.24</td>
<td>0.51</td>
<td>-0.19</td>
</tr>
<tr>
<td>Selflessness</td>
<td>-0.06</td>
<td>0.43</td>
<td>-0.23</td>
</tr>
<tr>
<td>Intellectual Efficiency</td>
<td>-0.12</td>
<td>0.54</td>
<td>0.00</td>
</tr>
<tr>
<td>Non-delinquency</td>
<td>0.13</td>
<td>0.44</td>
<td>0.06</td>
</tr>
<tr>
<td>Order</td>
<td>-0.33</td>
<td>0.55</td>
<td>-0.41</td>
</tr>
<tr>
<td>Physical Conditioning</td>
<td>-0.16</td>
<td>0.59</td>
<td>0.08</td>
</tr>
<tr>
<td>Self Control</td>
<td>0.05</td>
<td>0.54</td>
<td>0.05</td>
</tr>
<tr>
<td>Sociability</td>
<td>-0.03</td>
<td>0.58</td>
<td>-0.04</td>
</tr>
<tr>
<td>Tolerance</td>
<td>-0.07</td>
<td>0.53</td>
<td>-0.25</td>
</tr>
<tr>
<td>Optimism</td>
<td>0.12</td>
<td>0.46</td>
<td>0.13</td>
</tr>
</tbody>
</table>

*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.
Black-White Comparisons of TAPAS Scale Scores among U.S. Army Applicants at MEPS

### White-Black Comparisons for U.S. Army

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

*Note. W = White (N = 97,202); B = Black (N = 19,945).*
Criterion-related Validity

Does TAPAS predict performance?
MEPS TAPAS Results for Army IMT Outcomes

**APFT (n = 4382)**

<table>
<thead>
<tr>
<th>TAPAS Will-Do Composite Quintiles</th>
<th>APFT Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Low)</td>
<td>238</td>
</tr>
<tr>
<td>2</td>
<td>243</td>
</tr>
<tr>
<td>3</td>
<td>250</td>
</tr>
<tr>
<td>4</td>
<td>254</td>
</tr>
<tr>
<td>5 (High)</td>
<td>260</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>TAPAS Attrition Composite Quintiles</th>
<th>6-Month Attrition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Low)</td>
<td>12%</td>
</tr>
<tr>
<td>2</td>
<td>11%</td>
</tr>
<tr>
<td>3</td>
<td>9%</td>
</tr>
<tr>
<td>4</td>
<td>8%</td>
</tr>
<tr>
<td>5 (High)</td>
<td>7%</td>
</tr>
</tbody>
</table>

**Disciplinary Incidents (n = 2885)**

<table>
<thead>
<tr>
<th>TAPAS Will-Do Composite Quintiles</th>
<th>% With Incidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Low)</td>
<td>26%</td>
</tr>
<tr>
<td>2</td>
<td>21%</td>
</tr>
<tr>
<td>3</td>
<td>20%</td>
</tr>
<tr>
<td>4</td>
<td>17%</td>
</tr>
<tr>
<td>5 (High)</td>
<td>14%</td>
</tr>
</tbody>
</table>

**Self-Reported Adjustment (n = 4332)**

<table>
<thead>
<tr>
<th>TAPAS Will-Do Composite Quintiles</th>
<th>Adjustment (1-5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Low)</td>
<td>3.86</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4.09</td>
</tr>
<tr>
<td>4</td>
<td>4.11</td>
</tr>
<tr>
<td>5 (High)</td>
<td>4.23</td>
</tr>
</tbody>
</table>
TAPAS Results for Army IMT Outcomes For Infantry (MOS 11B)
TAPAS Results for Army IMT Outcomes For Military Police (MOS 31B)
TAPAS Results for Army IMT Outcomes For Medics (MOS 68W)
TAPAS Results for Army IMT Outcomes For Recruiters

**Army Commitment (n = 808)**

- Overall Performance Composite Quintiles:
  - 1 (Low): 2.79
  - 2: 3.15
  - 3: 3.51
  - 4: 3.47
  - 5 (High): 4

**Recruiting Fit (n = 813)**

- Overall Performance Composite Quintiles:
  - 1 (Low): 2.56
  - 2: 2.99
  - 3: 3.35
  - 4: 3.46
  - 5 (High): 3.94

**Training and Development Satisfaction (n = 816)**

- Overall Performance Composite Quintiles:
  - 1 (Low): 2.95
  - 2: 3.28
  - 3: 3.56
  - 4: 3.54
  - 5 (High): 3.87

**Performance Ratings (n = 609)**

- Mean Performance Ratings:
  - 1 (Low): 61.87
  - 2: 65.62
  - 3: 69.57
  - 4: 68.12
  - 5 (High): 71.71
In Sum,

TAPAS was designed as an easily customizable assessment tool to meet the needs of diverse users and researchers.

To this end, we 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

Results for “can-do” aspect of performance have been weaker than “will-do”
Thank you for the opportunity to talk about our work!
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