Development and Criterion-Related Validation of a Forced-Choice Soft Skills Test for 911 Dispatchers

IPAC Annual Conference 2023

Clinton Kelly, Ph.D.

Agenda

- Test Content Identification
- Item Development
- Performance Criteria & Ratings
- Validation Sample
- Construct Validity
- Test Reliability
- Test Validity
- Adverse Impact
- Test Utility
- Report Development
- Next Steps

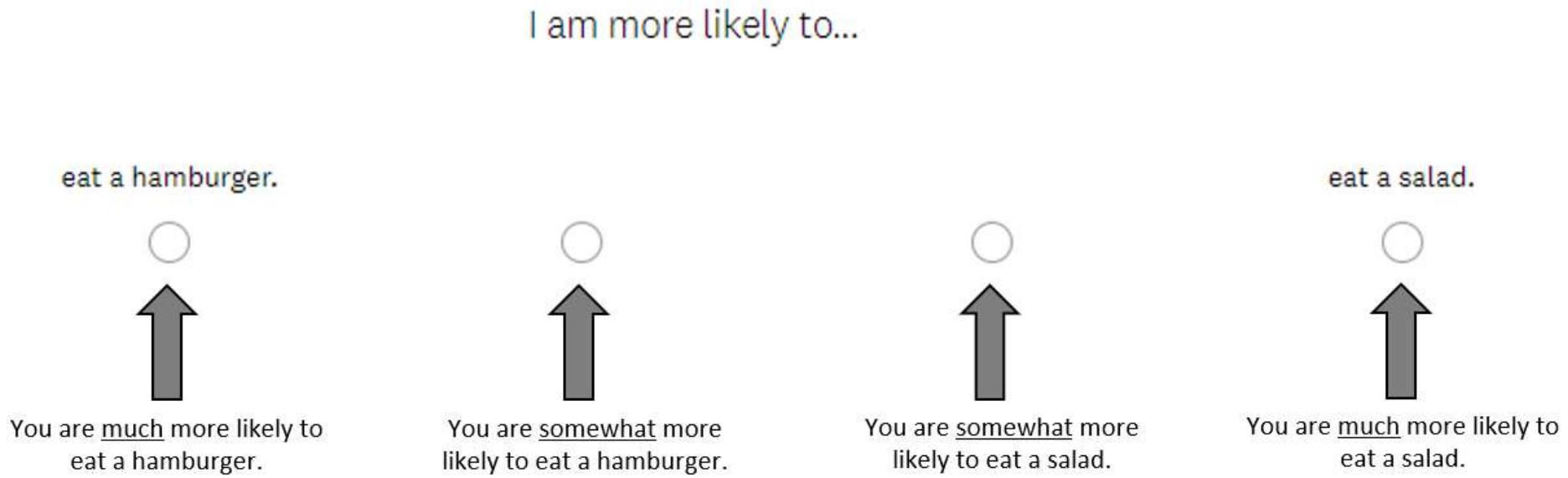
Test Content Identification

- Identified 22 KSAs from job analysis that were amenable to soft skills constructs
- Assigned 22 KSAs to 8 test constructs for item development
  - **Adaptability**
  - **Ambition**
  - **Composure**
  - **Confidence**
  - **Following Policy/Procedure**
  - **Interpersonal Perception**
  - **Multitasking**
  - **Resilience**

Item Type Determination

- Desire for resistance to faking (Cao & Drasgow, 2019; Lee & Joo, 2021)
- Increase item variance
- Ease of administration
- Minimize administration time

Sample Item


I am more likely to...
eat a hamburger.
eat a salad.
You are much more likely to eat a hamburger.
You are somewhat more likely to eat a hamburger.
You are somewhat more likely to eat a salad.
You are much more likely to eat a salad.

Sample Item


I am more likely to...
make quick, decisive decisions.
make long, thought out decisions.

Item Development

- Three test development professionals developed 187 items to the 8 constructs
- Item authors independently reviewed and ranked items within each construct
    - Jointly reviewed and selected final items for pilot
- Pilot test consisted of 118 items
    - Approximately 12 to 15 per construct

Performance Criteria

- Supervisors asked to evaluate incumbents in 10 performance areas and overall job performance
  - 11 overall job performance ratings
  - Each area was operationally defined
    - **Adaptable/Flexible**
    - **Ambitious/Motivated**
    - **Dependable**
    - **Creative Problem Solver**
    - **Perceptive/Insightful**
    - **Confident**
    - **Assertive**
    - **Resilient**
    - **Composed**
    - **Multitasker**
    - **Overall Performance**

Performance Criteria

- Example Performance Criteria
  - *Dependable:* Is reliable and can be counted on to do what they say they will do. Completes work assignments in a timely manner. Shows up for work on-time and ready to contribute. Follows rules, protocols, and procedures.
- 56 Supervisors evaluated employee job performance
  - Average of 6.04 per supervisor (St. Dev=5.49, Range=1 to 28)

Performance Ratings

Rating Dimension: Managing workflow by prioritizing responsibilities in a certain order or pattern, including safely performing more than one job-related task at a time (multi-tasking).

Employee	Rating Level									
	LOWER 1-10%	LOWER 11-20%	LOWER 21-30%	MIDDLE 31-40%	MIDDLE 41-50%	MIDDLE 51-60%	MIDDLE 61-70%	HIGHER 71-80%	HIGHER 81-90%	HIGHER 91-100%
Jane				X						
Armando	X									
Daniel			X							
Robert							X			
Mary					X					
Julie						X				X
Frank									X	
Ernesto								X		
Sarah						X				
Favian			X							
Sandra			X							
Ellie				X						
Patrick						X				
Dillion	X									
Jerome		X								X
Michael									X	
Mary					X					
Julie							X			X
Frank									X	
Ernesto								X		
RATINGS PER CATEGORY	2	1	3	2	2	3	2	2	3	3

Performance Ratings

Intercorrelations Between Job Performance Criteria

Job Performance Criteria		2	3	4	5	6	7	8	9	10	11
1. Adaptable/Flexible		.673**	.666**	.698**	.635**	.626**	.621**	.578**	.557**	.659**	.715**
	n	304	308	295	304	304	302	307	304	304	303
2. Ambitious/Motivated			.729**	.704**	.534**	.593**	.597**	.530**	.518**	.570**	.630**
		n	316	306	313	313	313	316	313	313	312
3. Dependable				.713**	.562**	.607**	.628**	.599**	.590**	.593**	.705**
			n	309	318	316	314	320	317	318	317
4. Creative Problem Solver					.644**	.744**	.754**	.610**	.603**	.681**	.712**
				n	306	304	306	307	304	305	303
5. Perceptive/Insightful						.658**	.599**	.593**	.587**	.562**	.563**
					n	313	310	317	314	313	312
6. Confident							.813**	.707**	.678**	.781**	.753**
						n	310	315	312	313	312
7. Assertive								.694**	.661**	.782**	.786**
							n	314	311	312	309
8. Resilient									.722**	.701**	.724**
								n	318	316	315
9. Composed										.709**	.702**
									n	315	313
10. Multitasker											.816**
										n	314
11. Overall Performance											
											--

Note: *Correlation significant at the 0.05 level. **Correlation significant at the 0.01 level.

Performance Ratings

- Principal Components Factor Analysis
  - A single factor explained 69.71% of the total variance
  - Consistent with meta-analytic findings (Viswesvaran, Schmidt & Ones, 2005)
- The 11 performance criteria were averaged into an overall combined average job-performance rating

Validation Sample (N=327)

Table 2: Participating Agencies

Agency	Location	Number of Participants
American Medical Response – New Haven	Connecticut	29
Ada County Sheriff	Idaho	40
Barnstable Police Department	Massachusetts	13
City of Alpharetta	Georgia	22
Florida Highway Patrol	Florida	133
Global Medical Response - Glendale	Arizona	42
Independence Police Department	Missouri	20
Windsor Police Service	Ontario, Canada	28

Table 3: Average Years of Experience of Participants

Average	Standard Deviation
8.81	8.80

Validation Sample (N=327)

Table 4: Gender of Participants

Male	Female	Declined to Answer
108	217	2

Table 5: Race/Ethnicity of Participants

White	Black/African American	Hispanic/ Latino	Asian / Pacific Islander	Native American / Alaska Native	Declined to Answer
241	43	27	3	1	12

Table 6: Age Groupings of Participants

Less than 20 years of age	20 – 29 years of age	30 – 39 years of age	40 – 49 years of age	50 – 59 years of age	60 – or more years of age	Declined to Answer
2	109	96	61	35	20	4

Construct Validity

- In addition to the 118 pilot items, participants completed the 48-item NEO-PI-3 Conscientiousness Items

Table 7: NEO Conscientiousness Facets and Test Scores

NEO Conscientiousness	Correlation
Achievement Striving	0.144* (*n* = 358)
Competence	0.201* (*n* = 358)
Deliberation	0.189* (*n* = 358)
Order	0.238* (*n* = 358)
Self-Discipline	0.165* (*n* = 358)
Dutifulness	0.202* (*n* = 358)
Overall NEO Conscientiousness	0.249* (*n* = 358)

Note: *Correlation significant at the 0.01 level.

Test-Retest Reliability

- 116 of the participants completed the test twice
  - First test contained all 118 pilot items
  - Second test contained only the 60 scored items
- Test-retest times ranged from one month to two and half months
  - Avg = 54.71 days
  - St Dev = 11.45 days
- Reliability coefficient for the final 60 items was 0.71

Minimizing Chance Results

- Uniform Guidelines Section 14B(7)
  - Users should avoid reliance upon techniques which tend to overestimate validity findings as a result of capitalization on chance unless an appropriate safeguard is taken.
  - Use of a large sample is one safeguard: cross-validation is another.

Sample Split

- Sample was randomly split in half (randomized by participating agency)
  - **Sample 1**: Validation Sample
    - Determine which responses on items were most predictive of performance
    - Items were deleted when there was no clear preference for either side of the forced choice statement with job performance ratings.
  - **Sample 2**: Holdout Sample
    - Verify that the item response pattern identified in Sample 1 was also predictive of performance in Sample 2

Table 8. Means, Standard Deviations, and Normality Statistics for Test and Criterion Measures

Test/Criteria	*n*	Mean	Std. Dev.	Skewness (1)	Kurtosis	Kolmogorov-Smirnov (2)		Shapiro-Wilk	
						Statistic	Sig.	Statistic	Sig.
Test									
Overall Test Score	367	30.49	4.40	0.17	-0.11	0.072	0.00	0.990	0.01
Criteria									
Adaptable/Flexible	310	5.67	2.53	-0.61	-0.61	0.18	0.00	0.92	0.00
Ambitious/Motivated	319	5.30	2.64	-0.46	-0.82	0.14	0.00	0.93	0.00
Dependable	324	5.98	2.52	-0.62	-0.61	0.16	0.00	0.91	0.00
Creative Problem Solver	310	5.17	2.56	-0.34	-0.90	0.12	0.00	0.94	0.00
Perceptive/Insightful	319	5.55	2.39	-0.48	-0.55	0.15	0.00	0.94	0.00
Confident	319	5.69	2.36	-0.51	-0.69	0.17	0.00	0.94	0.00
Assertive	316	5.51	2.58	-0.47	-0.86	0.16	0.00	0.93	0.00
Resilient	321	5.77	2.35	-0.51	-0.58	0.13	0.00	0.94	0.00
Composed	318	5.68	2.37	-0.44	-0.73	0.15	0.00	0.94	0.00
Multitasker	319	5.78	2.45	-0.56	-0.67	0.16	0.00	0.93	0.00
Overall Performance	319	6.16	2.24	-0.71	-0.31	0.17	0.00	0.92	0.00
Combined Average Job Performance	327	5.63	2.05	-0.36	-0.86	0.09	0.00	0.96	0.00

Note: (1) The Skewness is the Skewness Statistic / S.E. of Skewness; (2) Lilliefors Significance Correction. Non statistically significant results for the normality tests indicate the distributions are statistically normal.

Validity Coefficients

- Calculated uncorrected and corrected validity coefficients
  - Corrected Coefficients: job performance ratings attenuated (i.e., corrected) for the unreliability of the criteria (with an assumed reliability of rxx = 0.70)
- Uniform Guidelines Section 14B(6)
  - *“The appropriateness of a selection procedure is best evaluated in each particular situation and there are no minimum correlation coefficients applicable to all employment situations.”*

Validity Coefficients

- Statistically significant correlation of r > 0.20 is generally considered the minimum that should be considered acceptable for making hiring decisions

U. S. Department of Labor's Guidelines for Interpreting Correlations

Coefficient Value	Interpretation
Above 0.35	Very beneficial
0.21 - 0.35	Likely to be useful
0.11 - 0.20	Depends on circumstances
Below 0.11	Unlikely to be useful

Performance Criteria	Uncorrected Validity Coefficient			Corrected Validity Coefficient		
	Sample 1	Sample 2	Combined Sample	Sample 1	Sample 2	Combined Sample
Adaptable/Flexible	0.194* (*n* = 157)	0.251** (n = 153)	0.221** (n = 310)	0.232* (*n* = 157)	0.300** (*n* = 153)	0.264** (n = 310)
Ambitious/Motivated	0.217** (*n* = 160)	0.172* (*n* = 159)	0.193** (*n* = 319)	0.259** (*n* = 160)	0.206* (*n* = 159)	0.231** (*n* = 319)
Dependable	0.243** (*n* = 163)	0.295** (*n* = 161)	0.270** (*n* = 324)	0.290** (*n* = 163)	0.353** (*n* = 161)	0.323** (*n* = 324)
Creative Problem Solver	0.289** (*n* = 155)	0.316** (*n* = 155)	0.306** (*n* = 310)	0.345** (*n* = 155)	0.378** (*n* = 155)	0.366** (*n* = 310)
Perceptive/Insightful	0.135 (*n* = 163)	0.243** (*n* = 156)	0.185** (*n* = 319)	NS	0.290** (*n* = 156)	0.221** (*n* = 319)
Confident	0.325** (*n* = 159)	0.376** (*n* = 160)	0.354** (*n* = 319)	0.388** (*n* = 159)	0.449** (*n* = 160)	0.423** (*n* = 319)
Assertive	0.279** (*n* = 158)	0.348** (*n* = 158)	0.323** (*n* = 316)	0.333** (*n* = 158)	0.416** (*n* = 158)	0.386** (*n* = 316)
Resilient	0.199* (*n* = 161)	0.258** (*n* = 160)	0.234** (*n* = 321)	0.238* (*n* = 161)	0.308** (*n* = 160)	0.280** (*n* = 321)
Composed	0.276** (*n* = 159)	0.354** (*n* = 159)	0.317** (*n* = 318)	0.330** (*n* = 159)	0.423** (*n* = 159)	0.379** (*n* = 318)
Multitasker	0.252** (*n* = 159)	0.375** (*n* = 160)	0.321** (*n* = 319)	0.301** (*n* = 159)	0.448** (*n* = 160)	0.384** (*n* = 319)
Overall Performance	0.334** (*n* = 158)	0.355** (*n* = 161)	0.351** (*n* = 319)	0.399** (*n* = 158)	0.424** (*n* = 161)	0.420** (*n* = 319)

Note: *Correlation significant at the 0.05 level. **Correlation significant at the 0.01 level.

Validity Coefficients

- Sample 1
  - Uncorrected: 8 out of 11 coefficients > 0.20
  - Corrected: 10 out of 11 coefficients > 0.20
- Sample 2
  - Uncorrected: 10 out of 11 coefficients > 0.20
  - Corrected: 11 out of 11 coefficients > 0.20
- Combined
  - Uncorrected: 9 out of 11 coefficients > 0.20
  - Corrected: 11 out of 11 coefficients > 0.20

Validity Coefficients

Table 10: Relationship Between Combined Average Job Performance and Test Scores

Performance Criteria	Uncorrected Validity Coefficient			Corrected Validity Coefficient		
	Sample 1	Sample 2	Combined Sample	Sample 1	Sample 2	Combined Sample
Combined Average Job Performance	0.328** (*n* = 164)	0.355** (*n* = 163)	0.343** (*n* = 327)	0.392** (*n* = 164)	0.424** (*n* = 163)	0.410** (*n* = 327)

Note: *Correlation significant at the 0.05 level. **Correlation significant at the 0.01 level.

Validity Coefficients - Subscales

Table 11: Relationship Between Job Performance Domains and Sub Test Scores

Test Sub Scales/Performance Domains	Uncorrected Validity Coefficient			Corrected Validity Coefficient		
	Sample 1	Sample 2	Combined Sample	Sample 1	Sample 2	Combined Sample
Dependable	0.344** (*n* = 163)	0.407** (*n* = 161)	0.375** (*n* = 324)	0.411** (*n* = 163)	0.486** (*n* = 161)	0.448** (*n* = 324)
Confident/Assertive[13]	0.389** (*n* = 162)	0.365** (*n* = 163)	0.381** (*n* = 325)	0.465** (*n* = 162)	0.436** (*n* = 163)	0.455** (*n* = 324)
Resilient/Composed[14]	0.306** (*n* = 161)	0.339** (*n* = 160)	0.323** (*n* = 321)	0.366** (*n* = 161)	0.405** (*n* = 160)	0.386** (*n* = 324)
Multitasker	0.306** (*n* = 159)	0.364** (*n* = 160)	0.341** (*n* = 319)	0.366** (*n* = 159)	0.435** (*n* = 160)	0.408** (*n* = 324)

Note: *Correlation significant at the 0.05 level. **Correlation significant at the 0.01 level.

[13] Combined average of the Confident and Assertive job performance criteria.

[14] Combined average of the Resilient and Composed job performance criteria.

Test Fairness

- Black/African American, Hispanic/Latino, Asian/Pacific Islander, and Native American/Alaska Native were combined into a grouping labeled “non-white”
- Correlation for non-white group just narrowly (by 1%) missed being statistically significant (p = 0.06)

Table 12: Relationship Between Combined Average Job Performance and Test Scores

Performance Criteria	Uncorrected Validity Coefficient		Uncorrected Validity Coefficient	
	Males	Females	Whites	Non-Whites
Combined Average Job Performance	0.426** (*n* = 108)	0.282** (*n* = 217)	0.364** (*n* = 241)	0.220 (*n* = 74)

Note: *Correlation significant at the 0.05 level. **Correlation significant at the 0.01 level.

Test Bias

- Hierarchical moderated multiple regression analyses were conducted following the step-down procedure described by Lautenschlager and Mendoza (1986)
  - Began with the test scores alone being used to predict the job performance criterion
  - Next the protected class variable (e.g., race/ethnicity, gender) term and the interaction of this term with the test score were added to the model to simultaneously evaluate potential slope and/or intercept differences for the protected class being evaluated.

Test Bias: Race/Ethnicity

- Test scores contributed significantly to the regression model
  - F (1,313) = 38.47, p < 0.001
  - Accounted for 10.9% of variance in combined average job performance
- Adding the slope and intercept bias terms (i.e., white vs. non-white) does NOT contribute to the model
  - F (2,311) = 0.413, p = 0.66
  - Explained an additional 0.2% of variance in combined average job performance

Test Bias: Gender

- Test scores contributed significantly to the regression model
  - F (1,323) = 42.49, p < 0.001
  - Accounted for 11.6% of variance in combined average job performance
- Adding the slope and intercept bias terms (i.e., male vs. female) does NOT contribute to the model
  - F (2,321) = 0.455, p = 0.64
  - Explained an additional 0.2% of variance in combined average job performance

Mean Score Differences

Table 13: Average Overall Test Score by Gender

Subgroup	Avg	SD	*d*-value
Male (n = 115)	31.64	4.65	
Female (n = 249)	29.93	4.20	-0.38

Table 14: Average Overall Test Score by Race/Ethnicity

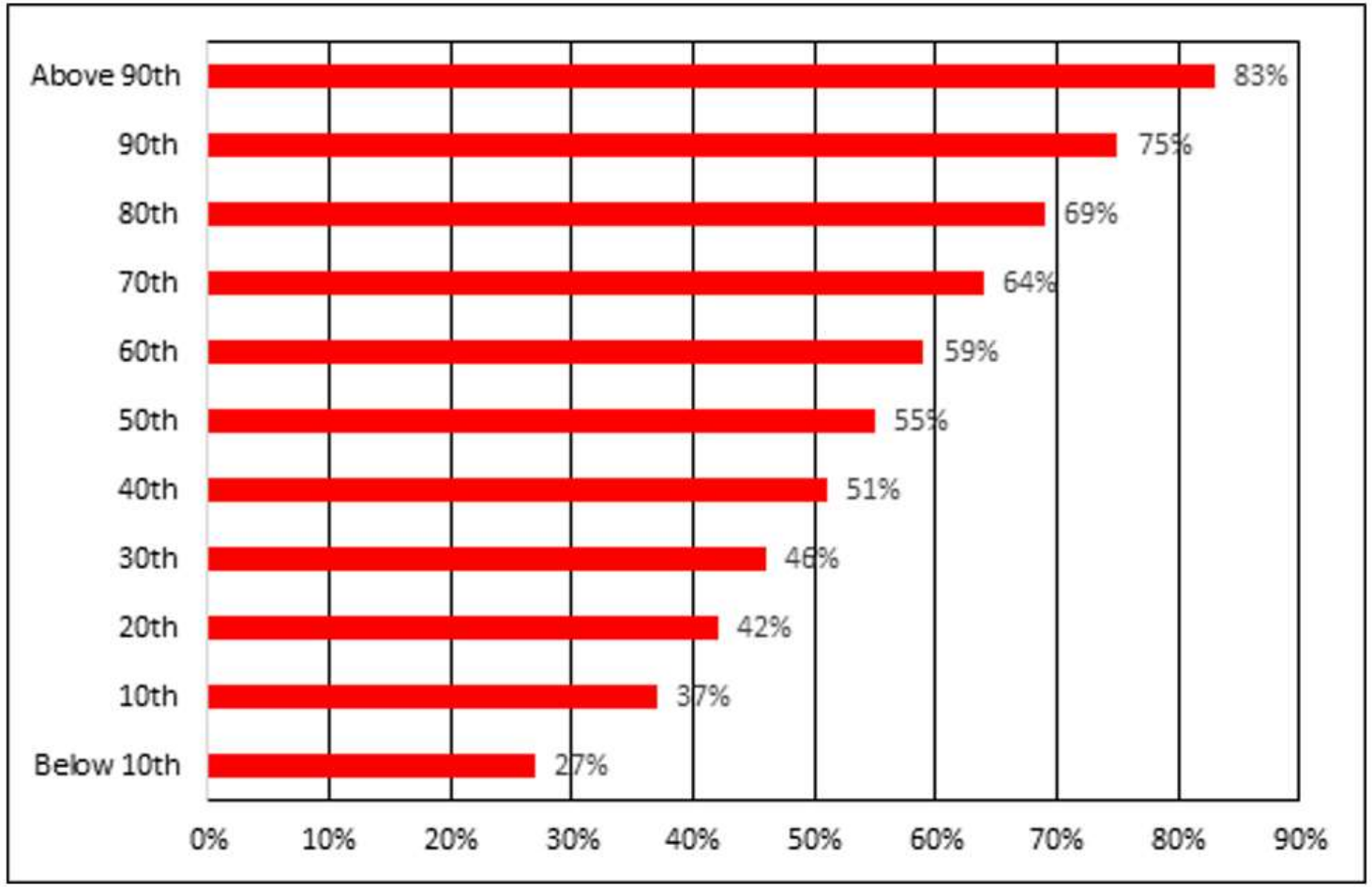
Subgroup	Avg	SD	*d*-value
White (n = 269)	30.51	4.48	
Black (n = 46)	29.54	4.18	-0.22
Hispanic (n = 32)	30.99	4.22	0.11
All Non-White (n = 82)	30.25	4.26	-0.06

[15] Score differences are evaluated with Cohen's *d* effect size using the pooled standard deviation to account for unequal group sizes. Positive values indicate that the traditionally underrepresented group (i.e., females, underrepresented minorities) obtain an average score that is higher than the male or White group. Negative values indicate that the traditionally underrepresented group (i.e., females, underrepresented minorities) obtain an average score that is lower than the male or White group.

Test Utility

- See Lawshe & Bolda, 1958 and Myors, 1994 for information about the expectancy table process

Figure 6: Likelihood of Success on the Job Based on Test Score



Test Utility

Table 15: Personality Test Score Bands

Overall Score Group	Likelihood of Being a Successful Performer on the Job	Amount of Validation Sample in Group	
		N	%
Highly Recommended	77%	75	23%
Recommended	53%	150	46%
Somewhat Recommended	31%	102	31%

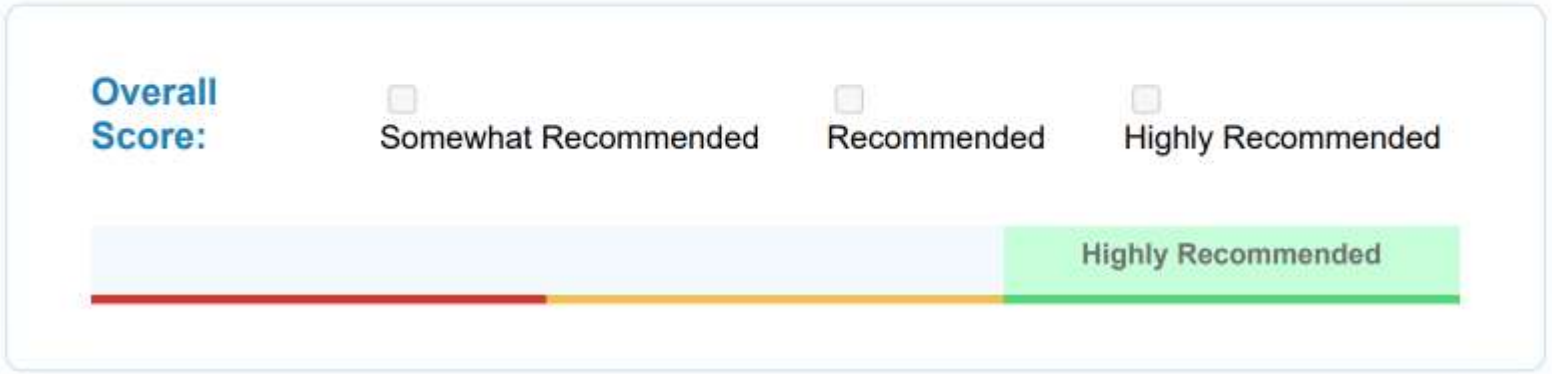
[17] Defined as receiving an above average combined job performance rating from a supervisor.

Report Development

- Educate and Provide Guidance
- Easy to Understand
- Useful for Hiring Managers

Report: Recommendation

Both CritiCall and the CritiCall Personality Test have been criterion validated. Criterion validation shows that there is a statistically-significant positive relationship between higher test scores and higher job performance. By measuring a wider variety of candidate attributes, one is able to predict a greater degree of their future job performance. For this reason, we highly recommend using these results in concert with the CritiCall test results. If an applicant possesses the skills and abilities necessary for success on the job, understanding how their behavioral attributes compare to that of high performers can greatly assist a hiring manager in making good hiring decisions.



Report: Explanation

PART 1: Overall Score Interpretation

Use Part 1 in combination with other skill and ability tests, like CritiCall tests, and any other assessments administered, to determine which of your job applicants should be advanced to the next phase of the recruitment Do not refer to Part 2 when making this decision. Part 2 is only to be used during the interview phase of the recruitment process.

This test was keyed and validated using incumbent employees, all of whom were performing, at least, at an acceptable level at the time of the study. Overall scores that receive a "Somewhat Recommended" rating align with the lower 31% of rated employees, scores receiving a "Recommended" rating align with the middle 46% of rated employees, and scores receiving a "Highly Recommended" rating align with the top 23% of rated employees. No recommendation can guarantee that an applicant will be successful on the job; however, the higher the overall score, the greater the likelihood of success on the job.

Based upon the Overall Score Recommendation, this test-taker is, at minimum, 77% likely to be a successful performer on the job.

The percentage value is defined as the likelihood of receiving an above average job performance rating from a supervisor. This estimate is based on test score and job performance evaluations of actual employees and represents the percentage of dispatchers, telecommunicators, and call-takers with a similar test score as your job applicant who obtained an above average job performance rating from their supervisor.

Report: Subscales

- Four Subscale Scores
  - (1) Dependability, (2) Confident/Assertive, (3) Resilience/Composure, (4) Task Management/Prioritization
- Built to maximize prediction for the performance criteria
  - Not factory analytically derived
- Provide Percentile Score
- Structured Interview Questions

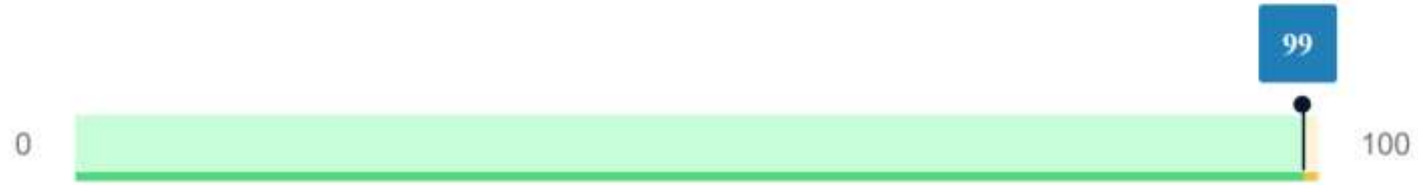
Report: Subscales

PART2:Job Performance Domain Subscore Interpretations

NOTE: The subscores are NOT a subset of the Overall Score, but rather standalone scales. Each of the four subscores presents a percentile score which can range from 0 to 100 for this particular job performance area. The percentile score indicates how the applicant compares to those who participated in the validation study for this particular job performance area.

Subscores are only intended to provide additional information that you may wish to explore during the interview or other stage of the hiring process. We strongly recommend that the overall score recommendation be the primary driver of your decision of whether or not to each applicant forward, and the subscores be used to drive conversations that will potentially uncover more valuable information about a candidate

Subscore 1: Dependability



High Scorers: Tend to be reliable, follow protocol and procedure, and can be counted on to follow through on their commitments.

Low Scorers: Tend to be less focused upon procedures, procrastinate when given an assignment, show up late or complete an assignment late, and bend the rules and/or justify not following protocol.

Suggested Interview Questions:

- Give an example of a time when you were asked to follow a rule even though you disagreed with it. How did you respond in this situation and why did you choose to respond the way you did?
- Give an example of a time when you committed to doing something and later had to back out of your commitment. How did you handle the situation and what was the outcome?
- Describe a time when you did not follow an established law, rule or procedure. Why did you decide to not follow it and what was the outcome?
- Describe a time when you were asked at work to do something outside of your normal responsibilities or that you knew would be difficult to complete by the given deadline. How did you respond to this situation and what was the outcome?

Next Steps

- Evaluate factor structure with larger sample of applicant data
- Compare item/test data from concurrent sample with that of job applicants
- Predictive criterion-related validation

Questions

ckelly@iopredict.com