OFFICE OF PEOPLE ANALYTICS

Development of a Complex Reasoning (CR) Test

Katherine Klein
Human Resources Research Organization (HumRRO)

Briefing presented to the DACMPT
June 12, 2024

Briefing Agenda

- Background Information
- CR Test Development Overview
- Update on Lines of Effort 2b and 3
- Discussion

OPA
OFFICE OF PEOPLE ANALYTICS

Background Information

OPA
OFFICE OF PEOPLE ANALYTICS

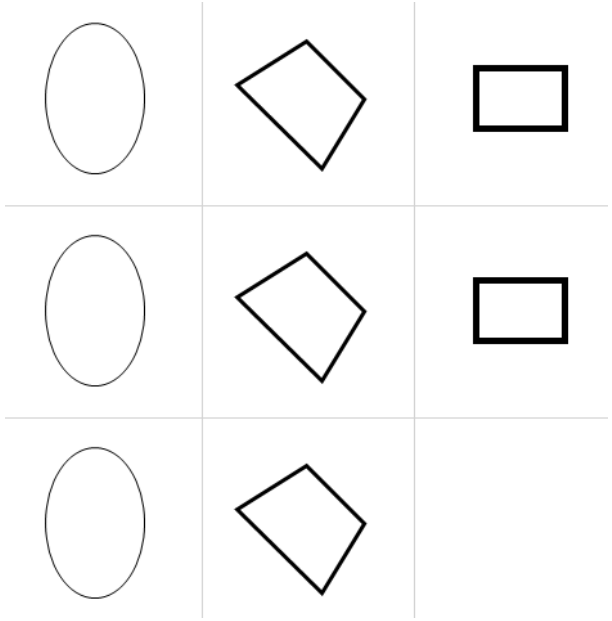
Background Information

- ***What is complex reasoning?***
  - Non-verbal reasoning; ability to analyze visual information and to solve problems using visual reasoning
- ***Why a complex reasoning test?***
  - Fluid intelligence has been found to be a strong predictor of training and job success
    - Complex (non-verbal) reasoning is one element of fluid intelligence
    - ASVAB Review Panel (2006) recommended that DoD consider adding tests of fluid intelligence to balance the ASVAB's composition (between fluid and crystalized intelligence)
  - Potential benefits to the ASVAB testing program
    - Improved prediction of training and job success in military jobs
    - Lower susceptibility to test compromise
    - Less adverse impact; increased qualification rates for non-native and non-heritage English speakers

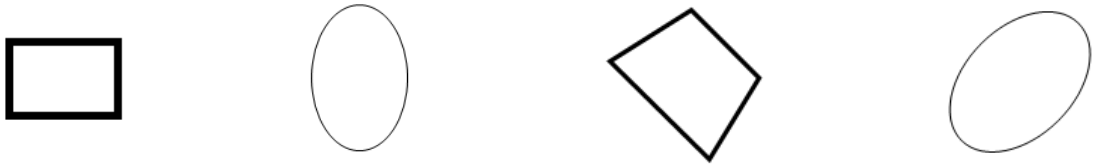
OPA
OFFICE OF PEOPLE ANALYTICS

Sample Transformation Item

Look at the 3X3 grid below. Identify the pattern(s).



Which of the following images best completes the pattern(s) in the grid?



- **Transformation item features**
  - Types of shapes
  - Orientation of shape(s)
  - Size of shape(s)
  - Number of shape(s)
  - Line weighting on shape(s)
- **Direction(s) of transformations**
  - Vertical
  - Horizontal
  - Diagonal

OPA
OFFICE OF PEOPLE ANALYTICS

CR Test Development Program Overview

OPA
OFFICE OF PEOPLE ANALYTICS

Complex Reasoning (CR) Test Development Program Overview

Line of Effort (LOE)	Progress
LOE 1: Develop, Pilot, and Evaluate Initial CR Capability	COMPLETED
LOE 2a: Develop an Improved CR Item Generation Tool	COMPLETED
LOE 2b: Pilot and Evaluate Refined CR Capability	COMPLETED
LOE 3: Develop Operational CR Test Form(s) and Future R&D/Maintenance Plans	COMPLETED

OPA
OFFICE OF PEOPLE ANALYTICS

LOE 2b: Pilot and Evaluate Refined CR Capability

OPA
OFFICE OF PEOPLE ANALYTICS

LOE 2b: CR Pilot Study 2 Overview

Objective

- Collect data on refined pool of CR items representative of the population of CR items with a military applicant representative sample

Design and Measures

- 24 CR items, 3 static forms, same 24 items on each form but in a different fixed order (spiraled by estimated difficulty)
- Pre- and post-test questionnaire
- Two CR attention-check items + insufficient effort

Sample

- Non-military sample representative of military applicants, ages 18–35, U.S. citizen, HS degree/GED/<1 year of college
- Targeted *N* = 2,600 participants
  - ~866 participants per form

Method

- Administered on Qualtrics platform
- Participants randomly assigned to one CR form
- No fixed time limit; record time to completion
- Desktop or laptop only

OPA
OFFICE OF PEOPLE ANALYTICS

Psychometric Summary

Internal Consistency

	Form A	Form B	Form C
N	838	853	842
α	.86	.86	.87
SEM	1.96	1.93	1.97
M CITC	.45	.44	.47
Min CITC	.20	.18	.18
5th Pct	.22	.19	.26
25th Pct	.42	.41	.44
50th Pct	.47	.46	.48
75th Pct	.52	.51	.56
95th Pct	.56	.56	.58
Max CITC	.61	.61	.63

Difficulty

	Form A	Form B	Form C
N	838	853	842
Average *p*	.63	.63	.63
Min *p*	.22	.18	.21
5th Pct	.37	.29	.37
25th Pct	.52	.53	.52
50th Pct	.64	.64	.63
75th Pct	.72	.72	.74
95th Pct	.92	.94	.92
Max *p*	.98	.99	.99

Dimensionality

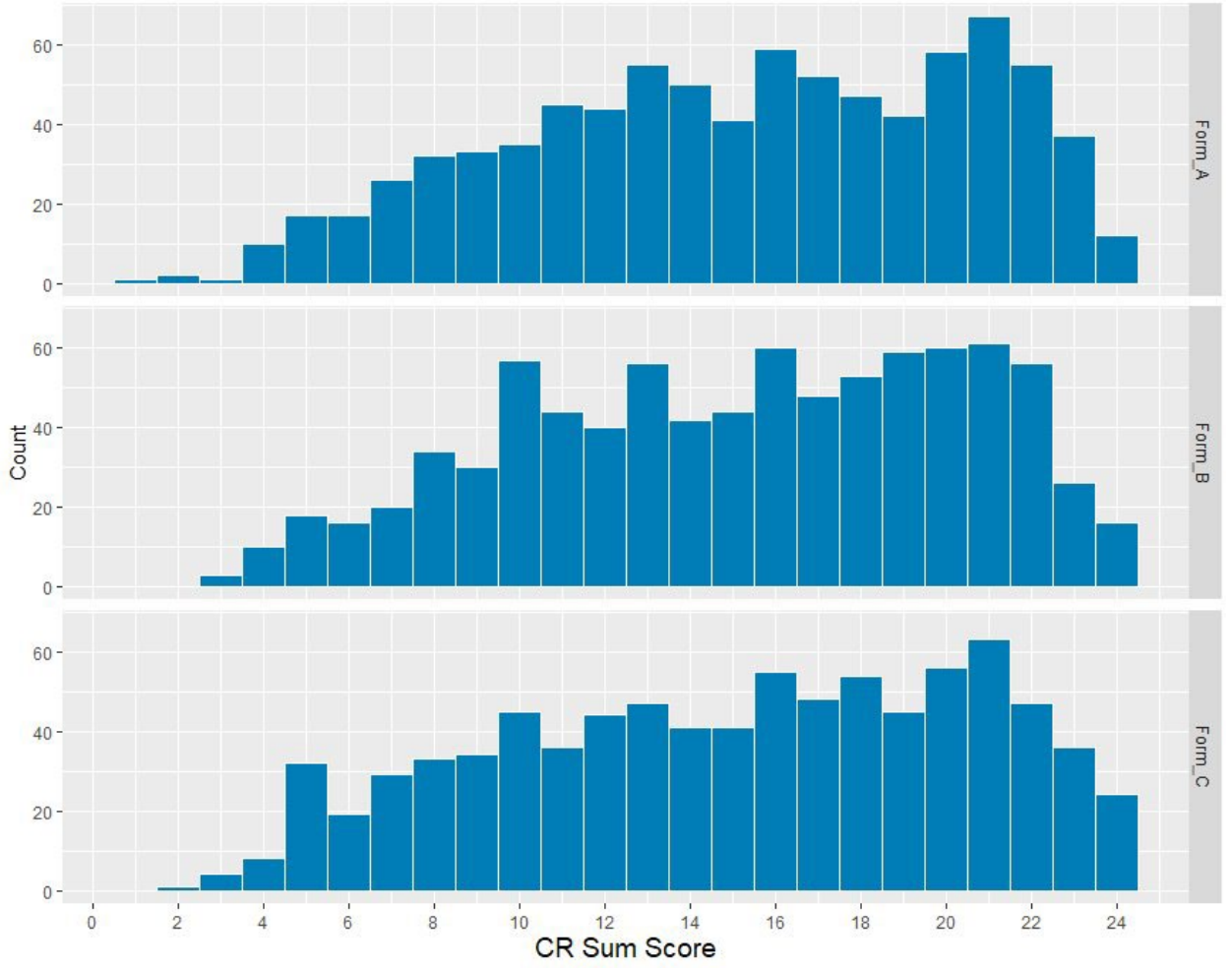
- We tested for unidimensionality by conducting a modified parallel analysis
- We observed a weak eigenvalue for a second factor in each of our three CR forms
  - Only Form B failed to reject our hypothesis test at $p < .05$
  - First factor eigenvalues ranged from 8.56 to 9.52
- Exploratory factor analysis results suggest a weak second factor consisting of the six, single-layer CR items (items 2–7)
- Single-layer items were easier (mean $p = .73$) relative to two-layer items (mean $p = .59$)

Form	*n*	Observed Eigenvalue	Simulated Eigenvalue	*p*
Form A	838	1.46	0.63	.009
Form B	853	1.24	0.73	.089
Form C	842	1.41	0.61	.009

OPA
OFFICE OF PEOPLE ANALYTICS

Test Score Distribution by Form

- We observed similar distributions of total sum scores between the three CR test conditions
  - Matching scores at the 25th, 50th, and 75th percentiles across forms
  - Kolmogorov-Smirnov test was conducted and the distribution across forms are not significantly different (D = .02–.04, p = .63–.99)
- We did not observe a significant difference in sum scores between the three CR form conditions ($F(2) = 0.29$, p = .75)



Form	*n*	*M*	*SD*	*Min*	5th Pct	25th Pct	50th Pct	75th Pct	95th Pct	*Max*
Form A	838	15.23	5.23	1.00	6.00	11.00	16.00	20.00	23.00	24.00
Form B	853	15.21	5.15	3.00	6.00	11.00	16.00	20.00	22.00	24.00
Form C	842	15.05	5.47	2.00	5.00	11.00	16.00	20.00	23.00	24.00
Total	2,533	15.16	5.28	1.00	6.00	11.00	16.00	20.00	23.00	24.00

OPA
OFFICE OF PEOPLE ANALYTICS

Group Score Differences

- We observed no statistically significant difference in overall CR sum scores based on participant gender ($F(1) = 1.58$, $p = .21$) or between racial and ethnic groups ($F(3) = 2.44$, $p = .06$)
- We did not detect any differences in group score differences between the three forms
- We did not detect any interaction effects between group and test form condition
- We did not observe any difference in CR scores between individuals who speak English only and those who speak a language other than English at home ($F(2) = 1.41$, $p = .24$)

Gender	*n*	*M*	*SD*	Observed *d*	Corrected *d*
Male	1,248	15.39	5.36	---	---
Female	1,258	14.95	5.18	−0.08	−0.09

Race-Ethnicity	*n*	*M*	*SD*	*Observed d*	*Corrected d*
White, non-Hispanic	1,136	15.53	5.24	---	---
Asian, non-Hispanic	110	16.73	4.90	0.23	0.26
Black, non-Hispanic	336	14.40	5.43	−0.21	−0.23
White, Hispanic	403	15.47	5.19	−0.01	−0.01

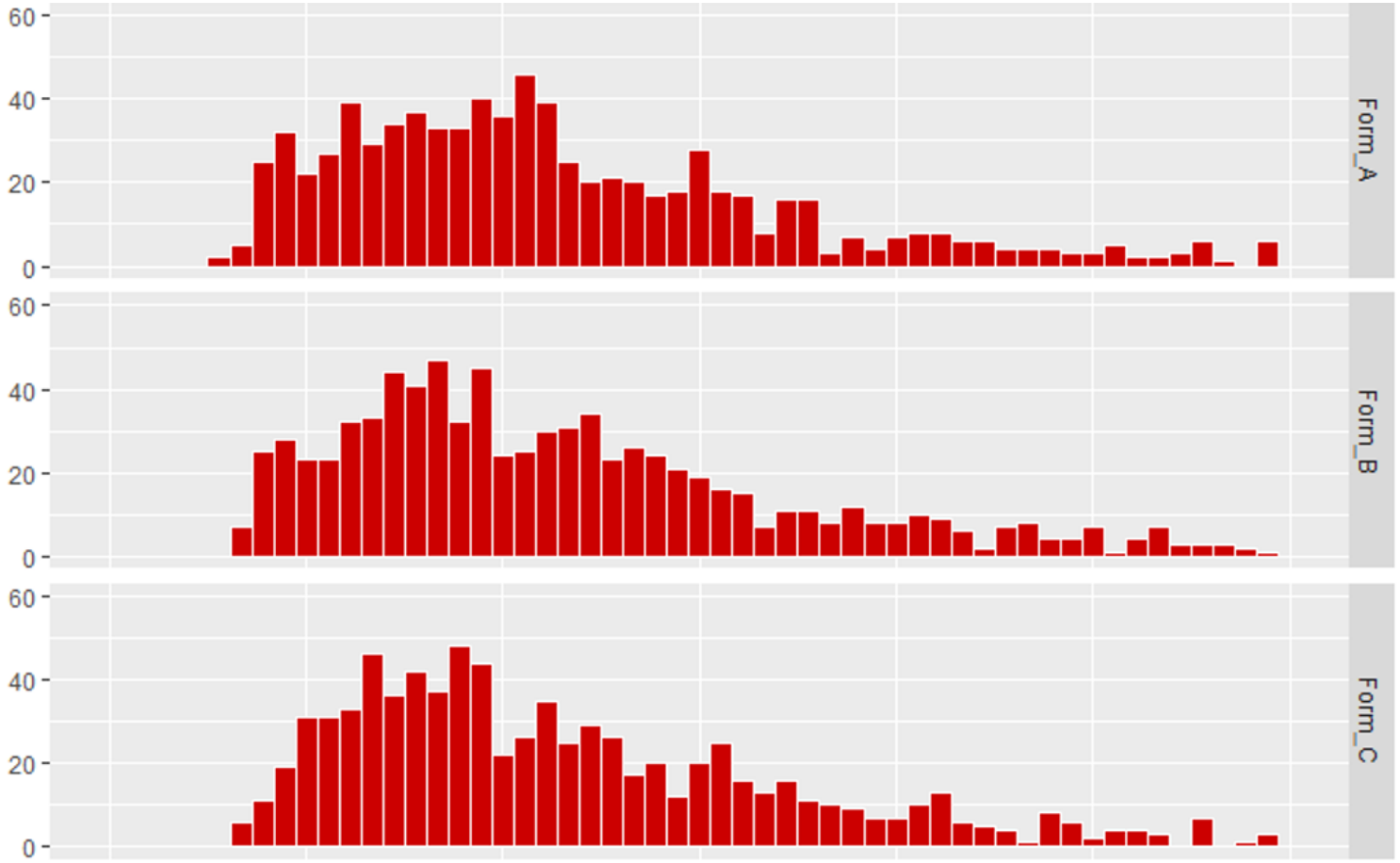
Language	*n*	*M*	*SD*	Observed *d*	Corrected *d*
English only	1,787	15.24	5.20	---	---
Other than English	727	15.02	5.44	−0.04	−0.05

Corrected *d*s were corrected for attenuation (i.e., unreliability) by dividing the observed *d* by the square root of the reliability (internal consistency) of the assessment

OPA
OFFICE OF PEOPLE ANALYTICS

Completion Time

- We observed similar distributions across the three forms
  - Similar across the various percentiles
- We did not find a significant difference in completion time between the forms ($F(2) = .23$, $p = .79$)
- Females took less time to complete the assessment compared to males (observed *d*s = −0.02)
- White, non-Hispanic participants took the least amount of time to complete the assessment compared to the other Race-Ethnicity groups (observed *d*s = −0.001 to 0.40)



Form	*M*	*SD*	Min	5th	25th	50th	75th	97.5th	99th	Max
Total	13.37	7.82	3.65	5.19	7.91	11.17	16.58	29.15	41.93	50.57

Note. Completion time = total time for reading the instructions, answering the CR items, and completing the debrief questions. Participants whose completion time was more than 60 minutes were removed from the sample.

OPA
OFFICE OF PEOPLE ANALYTICS

LOE 3: Develop Operational CR Test Form(s) and Future R&D/Maintenance Plans

OPA
OFFICE OF PEOPLE ANALYTICS

LOE 3: Operational Form Assembly

Item Scoring and Number Correct Test Score

- Four forms are static, and the 24 items constituting each form are administered in a specified presentation order
- Accordingly, all forms of the CR test are presently scored using Classical Test Theory (CTT).
  - CR items are dichotomously scored as correct or incorrect based on the scoring key (0,1).
  - An initial total score is calculated by summing the correct responses across all 24 items.

Transformation to the ASVAB Standard Score Metric

- Once the number correct score is calculated, the scores will be transformed into a T-distribution with a mean of 50 and a standard deviation of 10.
  - The scores will be rounded to the nearest whole number.

Illustrative Example with Two Forms	
Form X	Form Y
1	3
2	2
3	1
4	6
5	5
6	4
7	9
8	8
9	7
10	10
11	11
12	12
13	15
14	14
15	13
16	18
17	17
18	16
19	21
20	20
21	19
22	22
23	23
24	24

Low – Medium Difficulty

Medium Difficulty

Medium – High Difficulty

OPA
OFFICE OF PEOPLE ANALYTICS

Platform Development

Important Milestones	Current Target Dates
CR forms and CompT scores implemented in development environment	April 2024
CR forms and CompT scores implemented in pre-production	May 2024
CR forms and CompT scores implemented in production	August 2024
MEPCOM receives 4 new CR and CompT scores	August 2024

OPA
OFFICE OF PEOPLE ANALYTICS

LOE 3: R&D/Maintenance Plans (Near Term – 24 Months)

LOE 3.1: Design CR Piloting Methods and Develop New Items	LOE 3.2: Pilot New Items and Assemble Pools for Adaptive Version
3.1.1: Dimensionality analyses and item calibrations	**3.2.1:** Pilot new CR items
3.1.2: Design CR item piloting data collection	**3.2.2:** Conduct item analyses
3.1.3: Develop blueprint for a CAT version of CR	**3.2.3:** Assemble adaptive (CAT) forms/pools for initial evaluation
3.1.4: Develop new CR items	**3.2.4:** Develop conventional forms
3.1.5: Develop IT requirements	**3.2.5:** Scale and equate new CR test scores

OPA
OFFICE OF PEOPLE ANALYTICS

LOE 3: R&D/Maintenance Plans (Near Term – 24 Months)

LOE 3.3: Refinement of test-item specification, item generation, and form assembly for future adaptive pools	LOE 3.4: Design studies for ongoing psychometric evaluation and validation of CR and CompT
3.3.1: Identify refinements to test blueprints	**3.4.1:** Evaluate construct validity
3.3.2: Identify refinements to item generation	**3.4.2:** Evaluate criterion-related validity
3.3.3: Identify refinements to form assembly	**3.4.3:** Conduct ongoing psychometric analyses of CR forms, items, and CompT composite scores
	3.4.4: Evaluate coachability and practice effects

OPA
OFFICE OF PEOPLE ANALYTICS

Discussion

OPA
OFFICE OF PEOPLE ANALYTICS

Previous DAC Feedback

Recommendation 1: Examine Test Score Mode (11)

- The distribution changed once we finished data collection
- Most likely due to a different sub-sample at the beginning of data collection

Recommendation 2: Examine Nomological Net of CR

- CompT collecting data on CR, CT, and AR in Spring 2024
- CR will expand the validation evidence once we have enough operational data

Recommendation 3: Conduct Validation Work

- Part of future CR R&D plans when operational data are available
- Next Generation ASVAB effort compiled data on tests like CR (RPM) that may generalize (Adams et al., 2022) as well as other published work on the Abstract Reasoning Test (ART) (Embretson, 1998)

Recommendation 4: Examine Utility for Classification

- Future R&D plans include review of differential validity across occupations requiring differing levels of CR when operational data are available

OPA
OFFICE OF PEOPLE ANALYTICS

Guidance from the DAC

- Thoughts on the dimensionality results?
  - There was evidence for substantive psychological differences between one-layer vs. two-layer items.
- What implications does the progressive item difficulty that currently characterizes CR have for a CAT version of CR?
  - How should that be handled in a CAT version?

OPA
OFFICE OF PEOPLE ANALYTICS

Acknowledgments

Matthew Brown, *HumRRO*

Furong Gao, *HumRRO*

Mike Ingerick, *HumRRO*

Sergio Marquez, *HumRRO*

Scott Oppler, *HumRRO*

Sachi Phillips, *HumRRO*

Mary Pommerich, *DTAC*

Tia Fechter, *DTAC*

Matt Trippe, *DTAC*

Jeff Harber, *DTAC*

Ping Yin, *DTAC*

OPA
OFFICE OF PEOPLE ANALYTICS

Thank you!

For more information,
please contact:

Katherine Klein
KKlein@HumRRO.org
651.370.210

OPA
OFFICE OF PEOPLE ANALYTICS