OPA

OFFICE OF PEOPLE ANALYTICS

Development of a Complex Reasoning (CR) Test

Katherine Klein

Human Resources Research Organization (HumRRO)

Briefing presented to the DACMPT

January 22, 2024

Agenda

- Background
- Development Update
- Overview of New Task Order
- Pilot Study Three

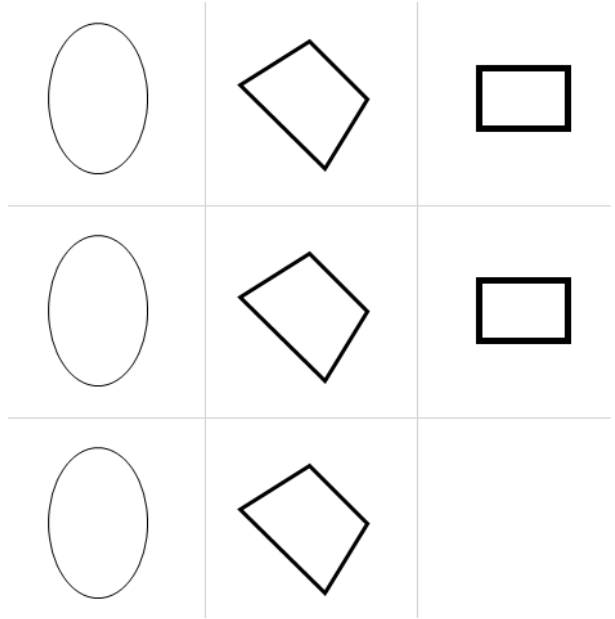
OPA
OFFICE OF PEOPLE ANALYTICS

Background

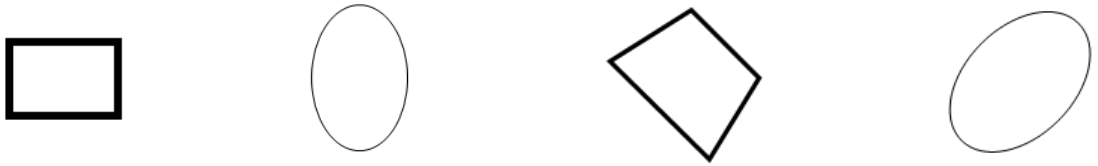
- ***What is complex reasoning?***
  - Non-verbal reasoning; ability to analyze visual information and to solve problems using visual reasoning
- ***Why a complex reasoning test?***
  - Fluid intelligence has been found to be a strong predictor of training and job success
    - Complex (non-verbal) reasoning 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

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

Development Update

Launched on the ASVAB Platform

- August 13, 2024
  - Four forms are static, and the 24 items constituting each form are administered in a specified presentation order

Available to Applicants

- September 16, 2024
  - A total of 9,837 applicants have taken the assessment between September 24 – November 4.

OPA
OFFICE OF PEOPLE ANALYTICS

CR Operational Descriptives		
	Raw	Standard Score
Mean	17.03	52.30
Standard Deviation	5.04	10.34
Min	0	17
5th Pct	7	32
25th Pct	14	46
50th Pct	18	54
75th Pct	21	60
95th Pct	23	65
Max	24	67
Correlation with ASVAB and Special Tests		
Armed Forces Qualification Test (AFQT)	.56	
Assembling Objects (AO)	.56	
Arithmetic Reasoning (AR)	.52	
Mechanical Comprehension (MC)	.51	
Math Knowledge (MK)	.49	
Cyber Test (CT)	.46	
General Science (GS)	.45	
Paragraph Comprehension (PC)	.45	
Verbal Expression (VE)	.45	
Electronics Information (EI)	.40	
Word Knowledge (WK)	.40	
Auto-Shop Information (AS)	.26	

Note. Correlations are observed and uncorrected; VE is a composite of WK and PC

Complex Reasoning (CR) Task Order

Line of Effort (LOE)
LOE 1: Design CR Items & Piloting Procedures • Dimensionality analyses and calibrations • Design CR item piloting data collection • Develop test blueprint for CAT version • Develop new CR items
LOE 2: Pilot New Items and Assemble CAT Pools • Pilot new CR items • Conduct item analysis • Develop CAT pools and conventional forms • Scale and equate scores
LOE 3: Recommend Refinements to Procedures • Identify refinements for test blueprints, item generation, and form assembly
LOE 4: Evaluate CR and CompT Scores • Create research plans to evaluate construct validity, criterion-related validity, ongoing psychometrics analysis, and coachability and practice effects
LOE 5: Document CR and CompT • Document task order efforts

OPA
OFFICE OF PEOPLE ANALYTICS

Pilot Study Three

OPA
OFFICE OF PEOPLE ANALYTICS

Wave 1

MEPS CR Form Version 1 (24 items)

- MEPS CR Form 1a w/CAT-like Order (24 items)
- MEPS CR Form 1b w/CAT-like Order (24 items)
- MEPS CR Form 1c w/CAT-like Order (24 items)
- MEPS CR Form 1d w/CAT-like Order (24 items)

Wave 2

MEPS CR Form Version 1 (24 items)

- New CR Tryout Set A (24 new items)
- New CR Tryout Set B (24 new items)
- New CR Tryout Set C (24 new items)
- New CR Tryout Set D (24 new items)

Wave 3

MEPS CR Form Version 1 (24 items)

- New CR Tryout Set E (24 new items)
- New CR Tryout Set F (24 new items)
- New CR Tryout Set G (24 new items)
- New CR Tryout Set H (24 new items)

Wave 4

MEPS CR Form Version 1 (24 items)

- New CR Tryout Set I (24 new items)
- New CR Tryout Set J (24 new items)
- New CR Tryout Set K (24 new items)
- New CR Tryout Set L (24 new items)

OPA
OFFICE OF PEOPLE ANALYTICS

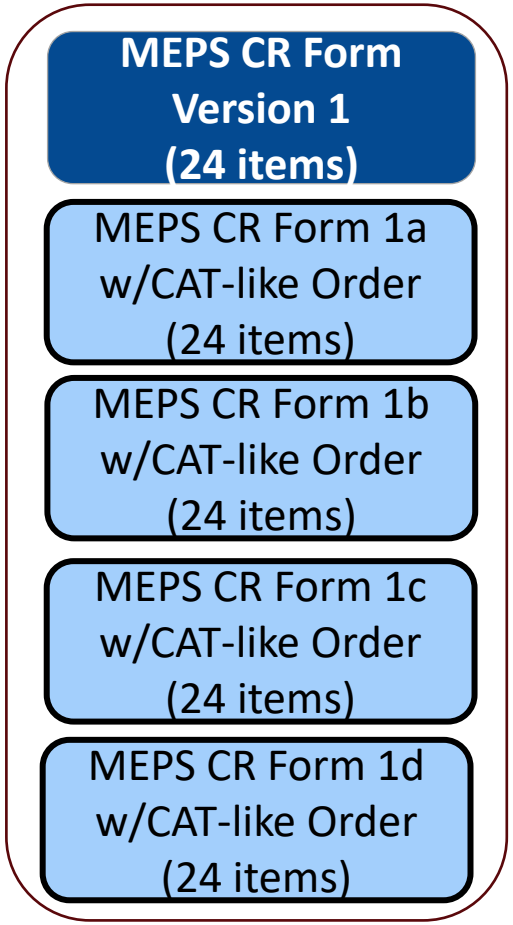
Wave 1 Overview

Objective

- Determine whether non-progressive item order impacts item functioning and test performance
  - Findings influence the feasibility of a CAT CR

Design and Measures

- 24 CR items, 5 static forms
- 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* = 5,250 participants (~1,050 participants per form)

Method

- Administered on Qualtrics platform
- Participants randomly assigned to one CR form
- 35-minute fixed time limit
- Record time to completion
- Desktop or laptop only

OPA
OFFICE OF PEOPLE ANALYTICS

Wave 1 Data Collection (as of 1 November)

Group	Pilot 3 (as of 1 November 2024)					
	MEPS Version	Form 1a	Form 1b	Form 1c	Form 1d	All Forms (Combined)
Total	109	107	109	94	101	502
Female	67	63	64	58	59	311
Asian	2	2	6	2	5	17
Black	31	30	18	25	29	133
Hispanic	27	28	21	17	19	112

OPA
OFFICE OF PEOPLE ANALYTICS

Waves 2 – 4 Overview

Objective

- Pilot test 288 new CR items for potential inclusion on the ASVAB platform
- Evaluate, calibrate, and link new CR items to new base IRT scale (estimated with operational CR data)

Sample

- Non-military sample representative of military applicants, ages 18–35, U.S. citizen, HS degree/ GED/<1 year of college
- Targeted *N* = 5,250 participants (~525 participants per form; 1,050 responses per item)

Design and Measures

- 24 CR items per examinee, multiple static forms with overlapping items
- Pre- and post-test questionnaire
- Two CR attention check items + insufficient effort

Method

- Administered on Qualtrics platform
- Within each wave, participants randomly assigned to one CR form
- 35-minute fixed time limit
- Record time to completion
- Desktop or laptop only

OPA
OFFICE OF PEOPLE ANALYTICS

Challenge & Methodology

- Determine how to calibrate and link the new CR items to base scale estimated from operational data on applicants
  - Conducted a simulation study (100 replications) to evaluate the three data collection designs and the four calibration designs to determine which resulted in the best psychometric solution

Data Collection Design Options Included:

1. Gold Standard—Operational + randomly seeded new items*
2. Fully Crossed—Every combination of evens and odds of new item sets with operational (e.g., even A, odd B)
3. Daisy Chain—Chained combinations of even and odd new item sets with operational
4. Random groups—Randomly assign one of five intact item sets (operational or one of four new item sets)

Scaling Method Options Included:

1. BILOG-Scaled Params*
2. True-Scaled Params*
3. Fixed OP Params
4. Fixed OP Params (Rescaled)
5. Latent Mu-Sigma Scaled
6. Stocking-Lord Equated

OPA
OFFICE OF PEOPLE ANALYTICS

*Comparison group only; option not being considered

Solution

Daisy Chain Design: 10 combinations of even-odd item sets across the operational form and four experimental item sets

Reasons for Recommendation:

1. All designs performed very similarly on psychometric metrics
2. Allows for common items, guards against deviations from randomly equivalent groups
3. Less intensive effort compared to fully crossed design

	OP Even	A Even	B Even	C Even	D Even
OP Odd	X	X			
A Odd		X	X		
B Odd			X	X	
C Odd				X	X
D Odd	X				X

OP = Operational Form

Note. Results can be reviewed in the back-up slides.

OPA
OFFICE OF PEOPLE ANALYTICS

Steps

1. Collect sufficient data at MEPS from military applicants on operational CR form (4 versions, same 24 items). MEPS military applicant sample and CR form to establish the new IRT base scale ***completed**
2. Calibrate operational CR form (24 items), derive new base scale using operational data on MEPS military applicant sample (Step 1) ***completed**
3. Pilot 288 new CR items (96 items per wave) using the daisy-chain design with non-military sample
4. Calibrate 288 new CR items (96 items per wave) using data collected (Step 3) and link to the new base scale (Step 2), scaling approach TBD (e.g., fixing parameters to operational MEPS sample, scaling to latent mu-sigma of operational MEPS sample, Stockard-Lord equating)

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC

- Does the DAC have any feedback on the Daisy-Chain design and plan for scaling and linking new CR items to the new base scale in Waves 2–4?
- Are there are any analyses we should consider for evaluating the feasibility of an adaptive CR version from the Wave 1 data?
- Are there any thoughts on creating an adaptive version of CR?

OPA
OFFICE OF PEOPLE ANALYTICS

Acknowledgments

Matthew Brown, *HumRRO*

Mike Ingerick, *HumRRO*

Scott Oppler, *HumRRO*

Sergio Marquez, *HumRRO*

Nathanial Voss, *HumRRO*

Alex Burgoyne, *HumRRO*

Leilani Seged, *HumRRO*

Robert Wellman, *HumRRO*

Sachi Phillips, *HumRRO*

Furong Gao, *HumRRO*

Jeff Dahlke, *HumRRO*

Mary Pommerich, *DTAC*

Matt Trippe, *DTAC*

Tia Fechter, *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

Back-up Slides

OPA
OFFICE OF PEOPLE ANALYTICS

Simulation Results — Bias

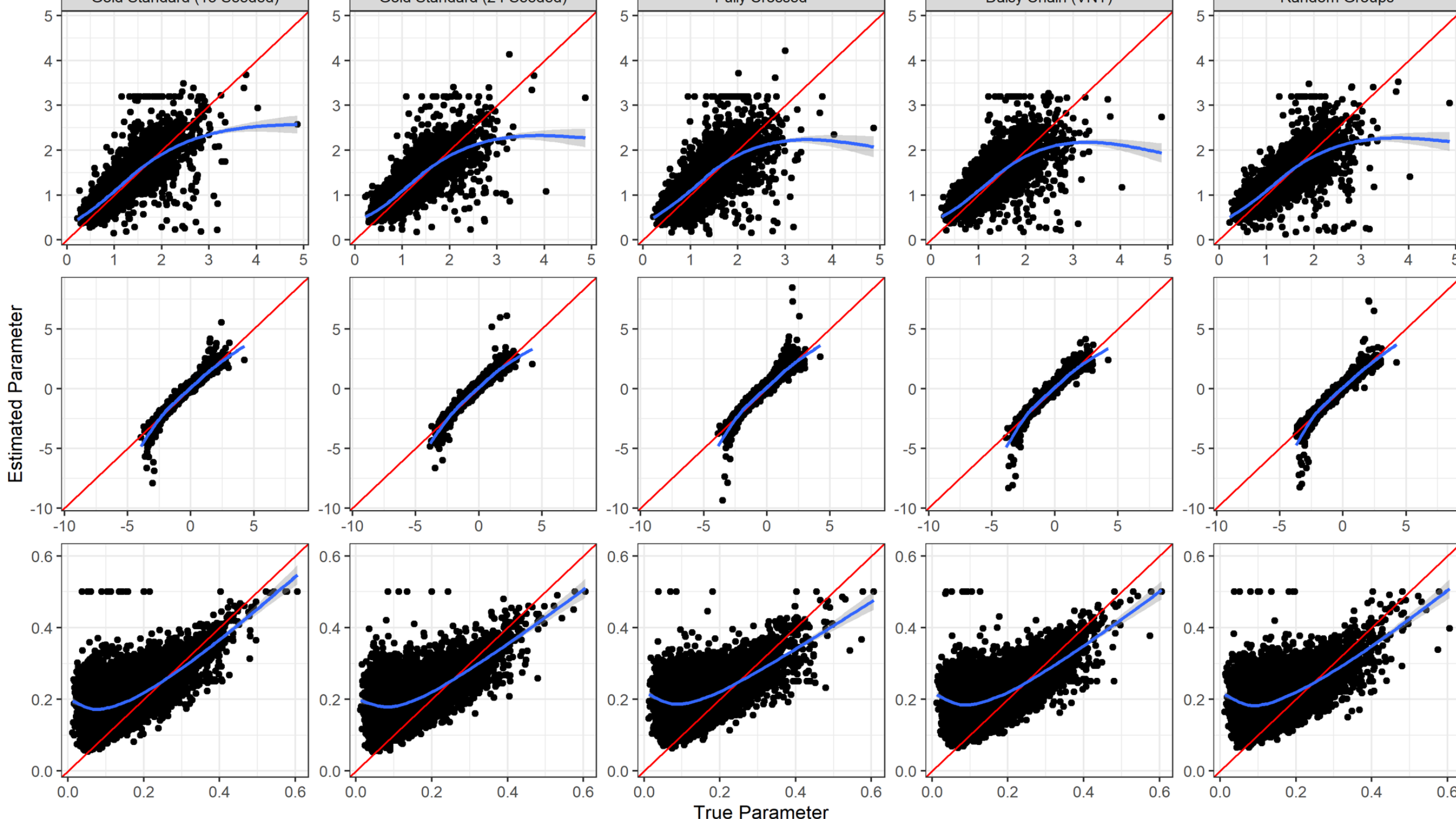
Evaluation	Scaling Method	Gold Standard (15 Seeded)	Gold Standard (24 Seeded)	Fully Crossed	Daisy Chain (VNT)	Random Groups
ICC	BILOG-Scaled Params	-0.047	-0.047	-0.047	-0.047	-0.047
ICC	True-Scaled Params	0.000	0.000	0.000	0.000	0.000
ICC	Fixed OP Params	-0.008	-0.009	-0.034	-0.044	-0.052
ICC	Fixed OP Params (Rescaled)	-0.003	-0.004	-0.003	-0.003	-0.002
ICC	Latent Mu-Sigma Scaled	0.000	0.000	-0.003	0.002	0.000
ICC	Stocking-Lord Equated	0.001	0.001	0.001	0.002	0.001
a	BILOG-Scaled Params	0.013	0.018	0.025	0.019	0.017
a	True-Scaled Params	0.067	0.073	0.081	0.074	0.072
a	Fixed OP Params	0.028	0.033	0.011	0.008	0.010
a	Fixed OP Params (Rescaled)	0.049	0.053	0.065	0.054	0.050
a	Latent Mu-Sigma Scaled	0.062	0.066	0.091	0.073	0.065
a	Stocking-Lord Equated	0.060	0.066	0.082	0.068	0.064
b	BILOG-Scaled Params	0.336	0.343	0.359	0.349	0.343
b	True-Scaled Params	0.081	0.088	0.103	0.093	0.087
b	Fixed OP Params	0.101	0.108	0.208	0.245	0.284
b	Fixed OP Params (Rescaled)	0.076	0.081	0.086	0.076	0.078
b	Latent Mu-Sigma Scaled	0.081	0.087	0.121	0.083	0.087
b	Stocking-Lord Equated	0.075	0.080	0.094	0.081	0.081
c	BILOG-Scaled Params	0.055	0.059	0.063	0.062	0.061
c	True-Scaled Params	0.055	0.059	0.063	0.062	0.061
c	Fixed OP Params	0.050	0.052	0.049	0.046	0.048
c	Fixed OP Params (Rescaled)	0.046	0.048	0.051	0.046	0.051
c	Latent Mu-Sigma Scaled	0.055	0.059	0.063	0.062	0.061
c	Stocking-Lord Equated	0.055	0.059	0.063	0.062	0.061

Simulation Results — RMSE

Evaluation	Scaling Method	Gold Standard (15 Seeded)	Gold Standard (24 Seeded)	Fully Crossed	Daisy Chain (VNT)	Random Groups
ICC	BILOG-Scaled Params	0.061	0.062	0.063	0.062	0.062
ICC	True-Scaled Params	0.022	0.024	0.023	0.023	0.023
ICC	Fixed OP Params	0.024	0.029	0.048	0.061	0.069
ICC	Fixed OP Params (Rescaled)	0.029	0.033	0.029	0.035	0.028
ICC	Latent Mu-Sigma Scaled	0.022	0.024	0.029	0.025	0.024
ICC	Stocking-Lord Equated	0.022	0.024	0.024	0.024	0.025
a	BILOG-Scaled Params	0.213	0.223	0.245	0.234	0.230
a	True-Scaled Params	0.226	0.236	0.261	0.247	0.243
a	Fixed OP Params	0.193	0.212	0.211	0.216	0.214
a	Fixed OP Params (Rescaled)	0.210	0.227	0.231	0.227	0.221
a	Latent Mu-Sigma Scaled	0.225	0.233	0.284	0.251	0.242
a	Stocking-Lord Equated	0.229	0.240	0.273	0.250	0.246
b	BILOG-Scaled Params	0.390	0.399	0.435	0.411	0.417
b	True-Scaled Params	0.202	0.211	0.251	0.225	0.240
b	Fixed OP Params	0.214	0.245	0.327	0.385	0.399
b	Fixed OP Params (Rescaled)	0.256	0.274	0.237	0.288	0.234
b	Latent Mu-Sigma Scaled	0.204	0.213	0.271	0.231	0.244
b	Stocking-Lord Equated	0.200	0.204	0.237	0.221	0.238
c	BILOG-Scaled Params	0.083	0.087	0.092	0.090	0.089
c	True-Scaled Params	0.083	0.087	0.092	0.090	0.089
c	Fixed OP Params	0.081	0.086	0.086	0.088	0.087
c	Fixed OP Params (Rescaled)	0.085	0.090	0.086	0.088	0.085
c	Latent Mu-Sigma Scaled	0.083	0.087	0.092	0.090	0.089
c	Stocking-Lord Equated	0.083	0.087	0.092	0.090	0.089

Simulation Results — r

Evaluation	Scaling Method	Gold Standard (15 Seeded)	Gold Standard (24 Seeded)	Fully Crossed	Daisy Chain (VNT)	Random Groups
a	BILOG-Scaled Params	0.862	0.847	0.817	0.831	0.836
a	True-Scaled Params	0.862	0.847	0.817	0.831	0.836
a	Fixed OP Params	0.889	0.865	0.863	0.856	0.858
a	Fixed OP Params (Rescaled)	0.873	0.850	0.851	0.850	0.858
a	Latent Mu-Sigma Scaled	0.861	0.848	0.793	0.825	0.833
a	Stocking-Lord Equated	0.856	0.839	0.802	0.824	0.828
b	BILOG-Scaled Params	0.983	0.982	0.975	0.980	0.976
b	True-Scaled Params	0.983	0.982	0.975	0.980	0.976
b	Fixed OP Params	0.983	0.977	0.970	0.958	0.963
b	Fixed OP Params (Rescaled)	0.971	0.966	0.976	0.962	0.976
b	Latent Mu-Sigma Scaled	0.983	0.982	0.972	0.978	0.975
b	Stocking-Lord Equated	0.984	0.983	0.978	0.980	0.976
c	BILOG-Scaled Params	0.586	0.554	0.500	0.518	0.519
c	True-Scaled Params	0.586	0.554	0.500	0.518	0.519
c	Fixed OP Params	0.569	0.508	0.499	0.457	0.476
c	Fixed OP Params (Rescaled)	0.507	0.447	0.506	0.457	0.515
c	Latent Mu-Sigma Scaled	0.586	0.554	0.500	0.518	0.519
c	Stocking-Lord Equated	0.586	0.554	0.500	0.518	0.519


Gold Standard (10 Seeded)
Gold Standard (24 Seeded)
Fully Crossed
Baby Shaln (VNT)
Random Groups
Estimated Parameter
True Parameter