OFFICE OF PEOPLE ANALYTICS

Development of a Complex Reasoning (CR) Test

Presentation to the Defense Advisory Committee on Military Personnel Testing (DACMPT)

Presenter: Mike Ingerick, Human Resources Research Organization (HumRRO)

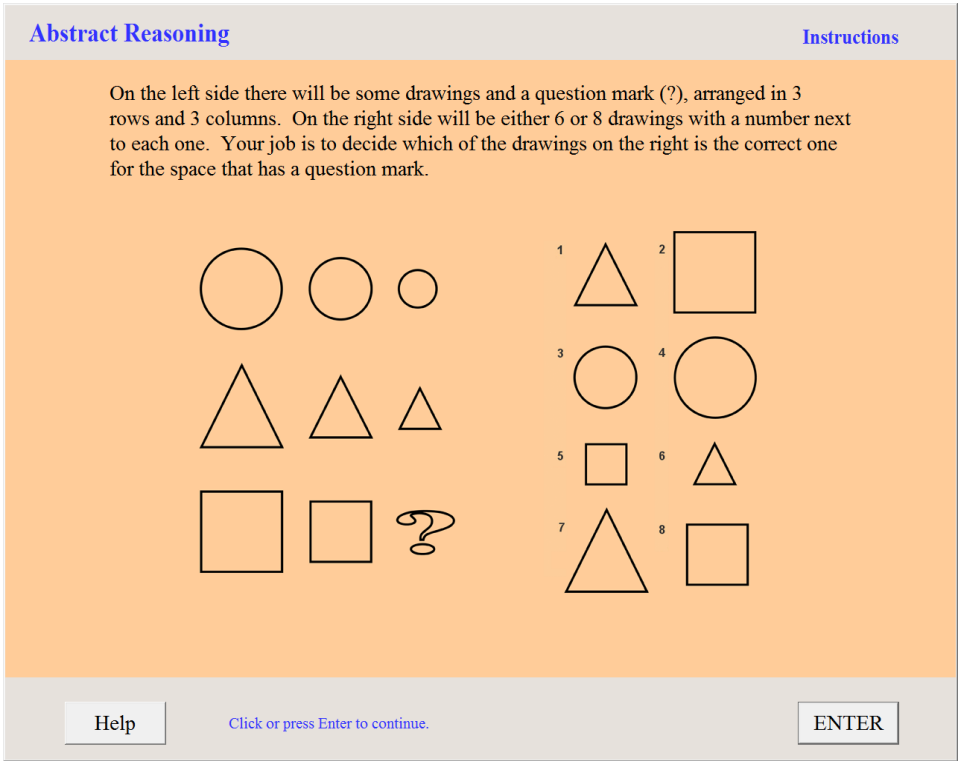
December 16, 2022

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

INITIAL DEVELOPMENT EFFORT: ABSTRACT REASONING TEST (ART)

- Developed by Susan Embretson
- Item format similar to Raven's Progressive Matrices (RPM)
  - Multiple choice, 6 or 8 response options per item
- DTAC commissioned the development of one form (30 items)
- Administered (for research purposes) to language training applicants (2017)
- Items were found to be relatively easy, time-consuming



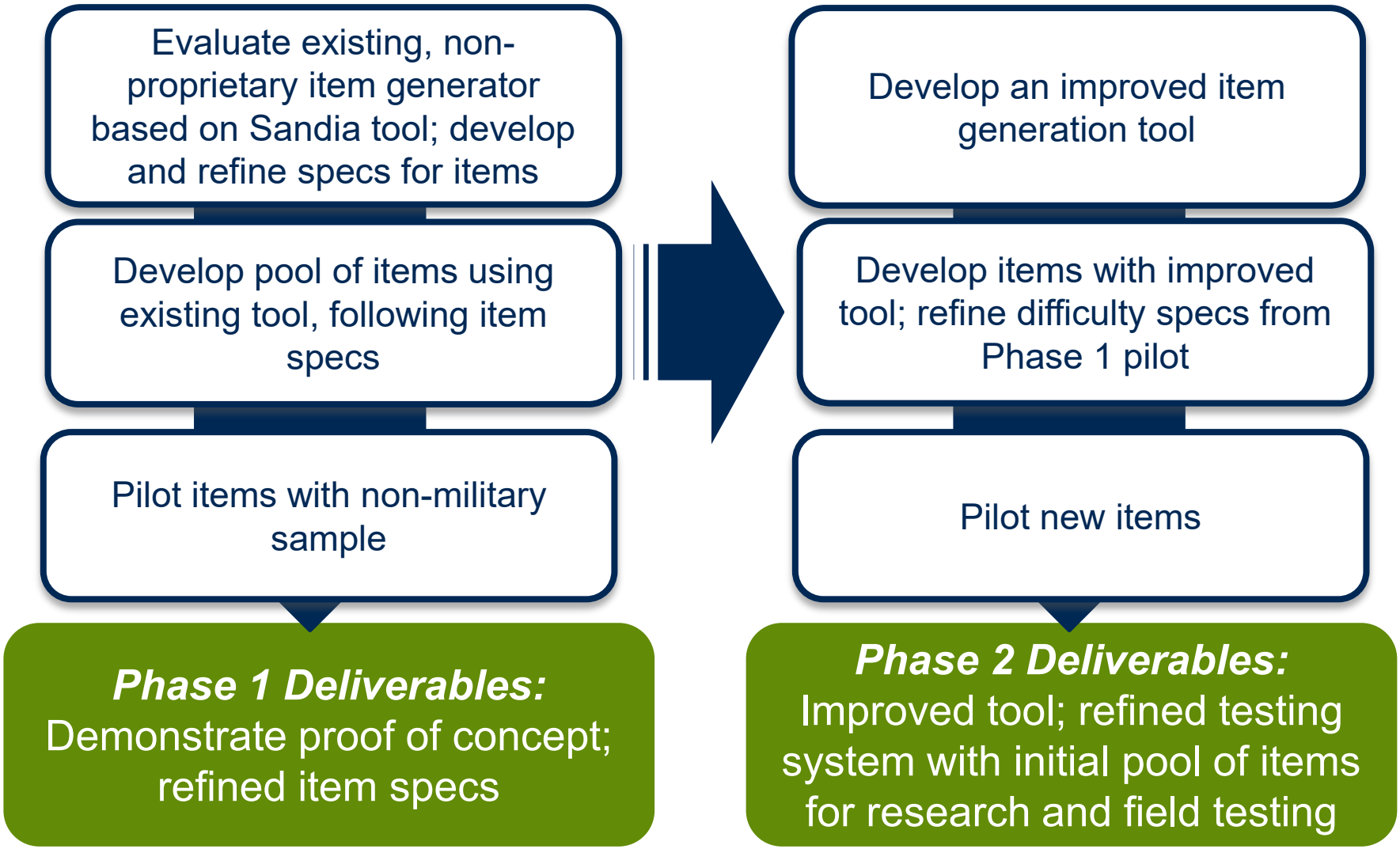
Sample ART Item

CURRENT DEVELOPMENT EFFORT

Objective: **Develop a complex (non-verbal) reasoning testing system to generate items for potential inclusion on ASVAB**

- **Employ non-proprietary Automated Item Generation (AIG) capability**
  - Improve item development efficiency
  - Reduce or eliminate field-testing requirements
- **Generate items with targeted properties**
  - Items similar to Raven's Progressive Matrices (RPM) items
  - Items at appropriate difficulty for qualifying military applicants into jobs of varying complexity

CURRENT DEVELOPMENT EFFORT (CONTINUED)

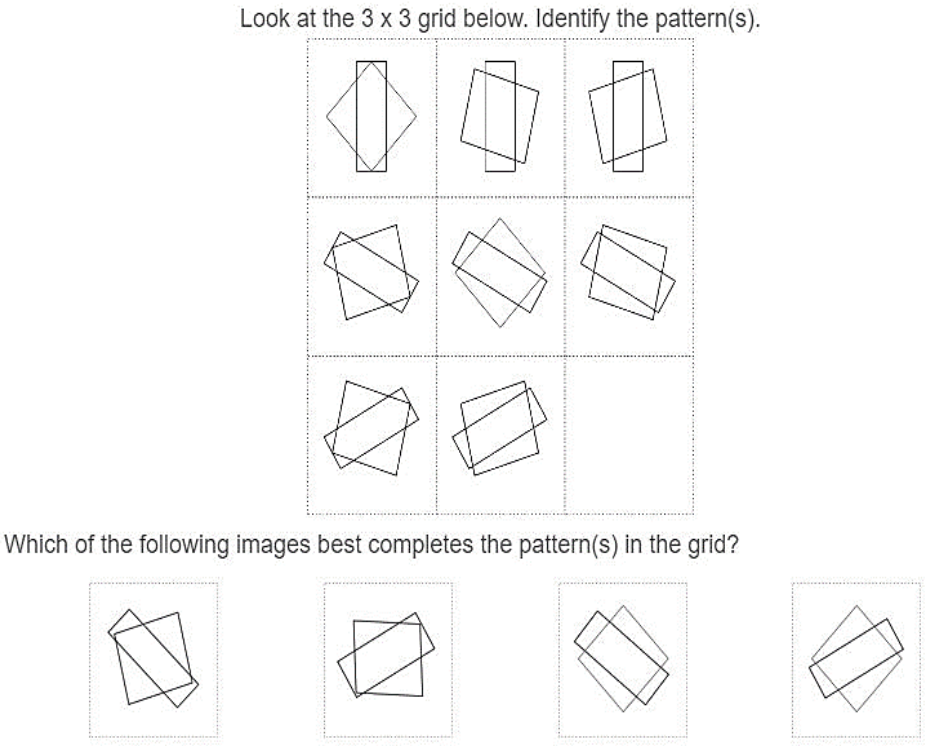


PHASE 1 PILOT: DESIGN

Pilot intended to answer the following questions:

- Does performance differ by item type (transformation vs. logic)? Does performance align with item difficulty specs?
- Does the number of response options impact performance? Does including a “None of these are correct” (NOTAC) option impact performance?
- Does performance differ by gender, race, or ethnicity?
- How much time is needed to complete the items? Does completion time differ by item type or response option set? Does completion time differ by gender, race, or ethnicity?

PHASE 1 PILOT: DESIGN (CONTINUED)



Sample Transformation Item

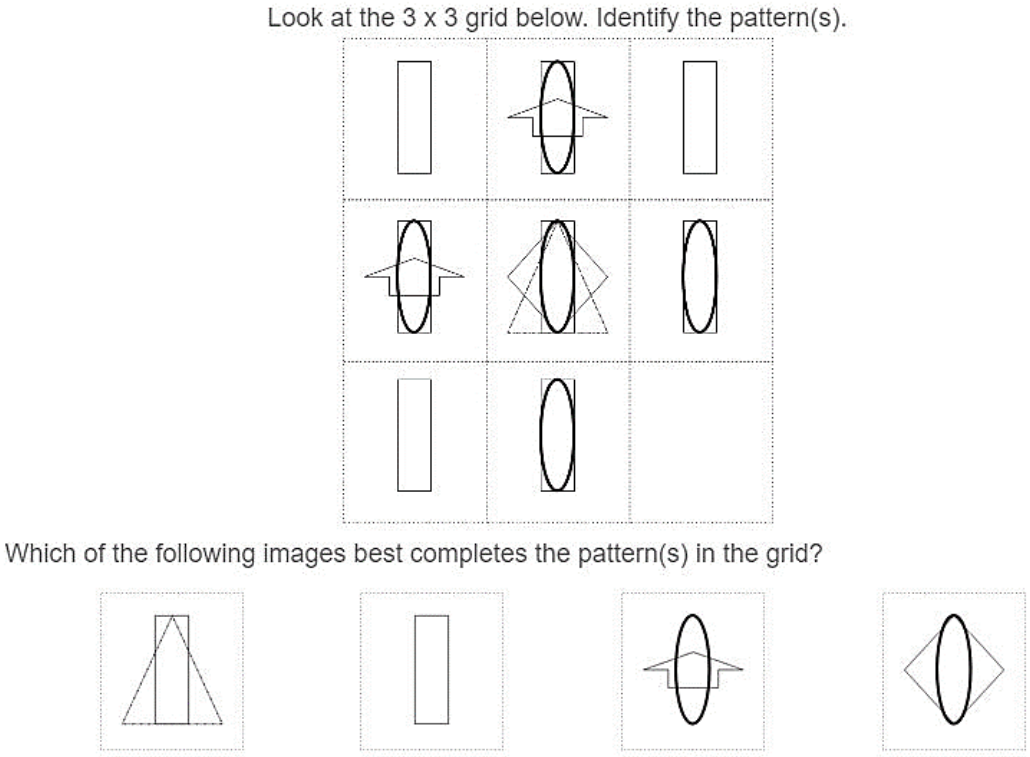
Transformation item features:

- **Number of transformations**
  - 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

PHASE 1 PILOT: DESIGN (CONTINUED)

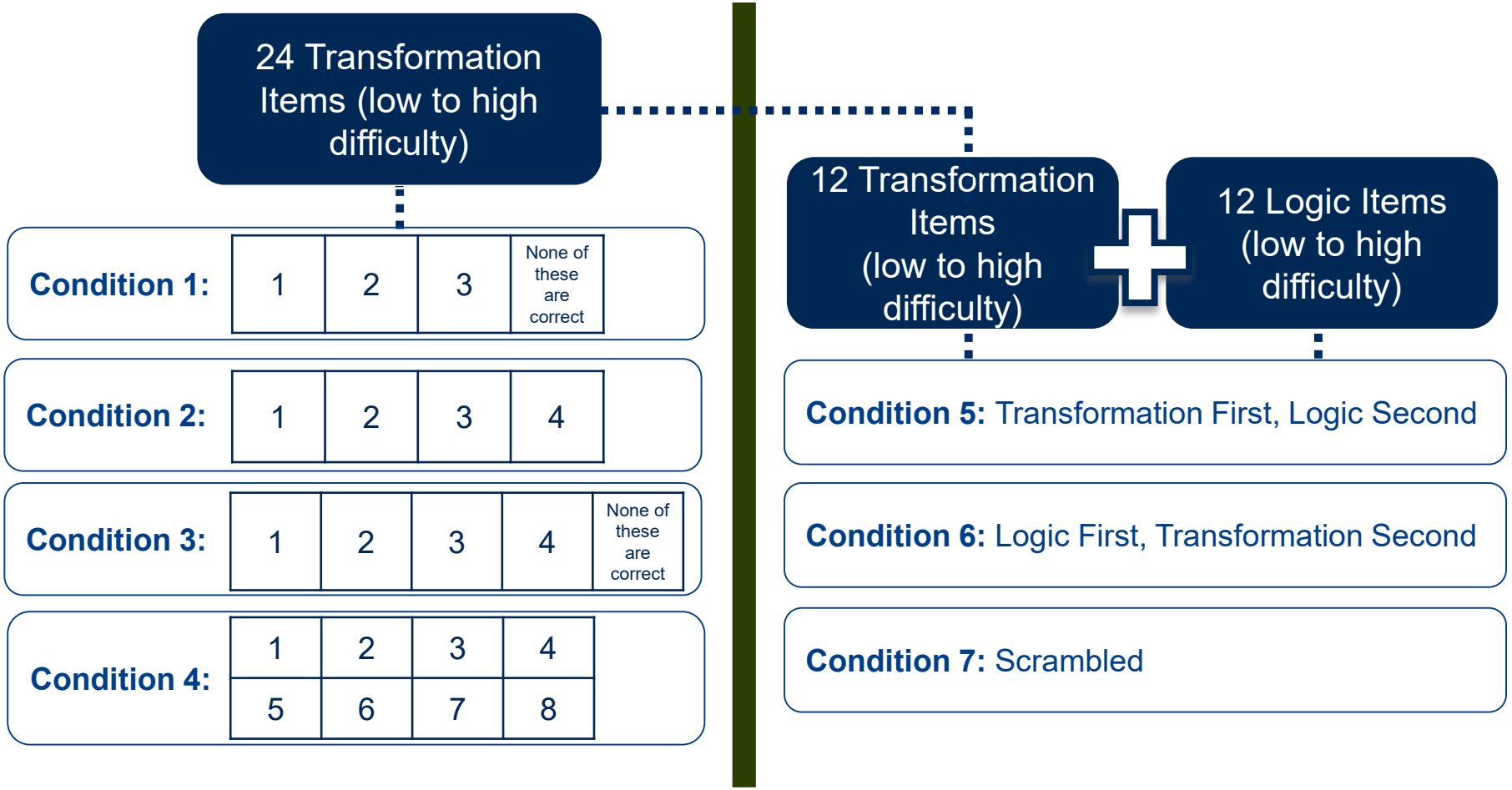
Logic item feature:

- **Nature of logic rule**
  - AND
  - OR
  - XOR (Exclusive OR)



Sample Logic Item ("AND")

PHASE 1 PILOT: DESIGN (CONTINUED)



PHASE 1 PILOT: DATA COLLECTION

Sample

- Non-military sample, ages 18–35, U.S. citizen
- Targeted *N* = 3,500 participants

Measures

- One complex reasoning form (24 items)
- Post-test questionnaire (demographics, perceived difficulty of items, test-taking experience)
- Two attention check items

Method

- Administered on Qualtrics platform
- No fixed time limit; recorded time to completion
- Desktop or laptop only

Dates of Data Collection:
19 July – 4 August (~2 weeks)

	N	*% of Total Invited*
Invited to participate in pilot	7,039	--
Accessed pilot	4,459	63.3%
Completed pilot	3,778	53.7%
Completed pilot w/ valid data	**3,491**	**49.6%**

Final Analysis Sample for Pilot

PHASE 1 PILOT: DEMOGRAPHICS AND EDUCATIONAL BACKGROUND OF FINAL ANALYSIS SAMPLE

All Participants

Condition Description	*N*	Mean Age	% Female	% Asian	% Black	% Hispanic	% HS Degree/GED/ < 1 yr College
3 + NOTAC	501	27.2	59.3	6.6	10.8	12.0	36.3
4	503	26.7	62.2	6.2	11.3	9.5	37.9
4 + NOTAC	490	26.7	59.8	5.9	15.5	10.2	35.1
8	493	27.0	61.1	5.5	10.8	11.6	36.5
Transform 1st, Logic 2nd	507	26.4	59.0	6.5	11.4	10.1	32.3
Scrambled	492	26.6	57.3	7.7	11.6	10.0	33.3
Logic 1st, Transform 2nd	505	26.9	61.8	6.3	13.1	8.9	30.0
Overall	***3,491***	***26.8***	***60.1***	***6.4***	***12.1***	***10.3***	***34.5***

Participants Completed in 30 Min or Less with HS Degree, GED, or < 1 yr of College

Condition Description	*N*	Mean Age	% Female	% Asian	% Black	% Hispanic	% HS Degree	% GED	% < 1 yr College
3 + NOTAC	182	25.2	52.7	5.5	16.5	23.1	65.9	9.3	24.7
4	188	25.7	59.2	5.3	14.9	22.3	62.8	14.7	22.5
4 + NOTAC	172	25.2	60.4	4.6	22.6	16.2	70.3	11.6	18.0
8	179	24.9	61.6	3.4	13.4	25.7	64.4	11.7	23.9
Transform 1st, Logic 2nd	164	24.7	53.7	4.9	19.5	19.5	65.2	9.8	25.0
Scrambled	164	25.4	52.6	7.9	13.8	21.0	67.8	7.9	24.3
Logic 1st, Transform 2nd	151	25.0	64.0	3.6	12.2	18.3	65.2	9.7	25.0
Overall	***1,200***	***25.2***	***57.7***	***5.0***	***16.1***	***20.9***	***65.9***	***10.7***	***23.3***

PHASE 1 PILOT: SUMMING UP (ALL PARTICIPANTS, *N* = 3,491)

Metric	Transform Only 8	Transform Only 3 + NOTAC	Transform Only 4	Transform Only 4 + NOTAC	Transform + Logic Grouped	Transform + Logic Scrambled
Unidimensionality	Yes	No	Yes	No	No	No
Reliability	α = .87 SEM = 1.98 *avg* CITC = .46	α = .78 SEM = 1.97 *avg* CITC = .31	α = .85 SEM = 1.99 *avg* CITC = .42	α = .83 SEM = 1.96 *avg* CITC = .37	α = .75 SEM = 2.10 *avg* CITC = .32	α = .75 SEM = 2.04 *avg* CITC = .30
Observed Difficulty	*M* = 12.29 *SD* = 5.48 avg *p* = .51	*M* = 13.29 *SD* = 4.21 avg *p* = .54	*M* = 15.00 *SD* = 5.15 avg *p* = .63	*M* = 11.96 *SD* = 4.76 avg *p* = .50	*M* = 10.32 *SD* = 4.19 avg *p* = .43	*M* = 9.15 *SD* = 4.08 avg *p* = .38
Group Score Differences	F-M *d* = .21 B-W *d* = -.39 H-W *d* = -.17 A-W *d* = .36	F-M *d* = .31 B-W *d* = -.31 H-W *d* = -.15 A-W *d* = .17	F-M *d* = .22 B-W *d* = -.58 H-W *d* = -.21 A-W *d* = .00	F-M *d* = .16 B-W *d* = -.20 H-W *d* = -.19 A-W *d* = .12	F-M *d* = .10 B-W *d* = -.34 H-W *d* = -.22 A-W *d* = .59	F-M *d* = .22 B-W *d* = -.08 H-W *d* = .04 A-W *d* = .24
Completion Time (30 minutes or less)	*M* = 12.74 *SD* = 5.86	*M* = 10.88 *SD* = 4.94	*M* = 11.36 *SD* = 5.14	*M* = 12.39 *SD* = 5.79	*M* = 13.54 *SD* = 6.18	*M* = 13.24 *SD* = 6.36
Perceived Difficulty	*M* = 3.92 *SD* = .95	*M* = 3.89 *SD* = .92	*M* = 3.98 *SD* = .89	*M* = 3.90 *SD* = .95	*M* = 3.50 *SD* = .96	*M* = 3.37 *SD* = 1.00

Note. NOTAC = None of these are correct. SEM = Standard error of measurement. CITC = Corrected item-total correlation.

PHASE 1 PILOT: SUMMING UP (ALL PARTICIPANTS, *N* = 3,491)

Metric	Transform Only 8	Transform Only 3 + NOTAC	Transform Only 4	Transform Only 4 + NOTAC	Transform + Logic Grouped	Transform + Logic Scrambled
Unidimensionality	Yes	No	Yes	No	No	No
Reliability	α = .87 SEM = 1.98 *avg* CITC = .46	α = .78 SEM = 1.97 *avg* CITC = .31	α = .85 SEM = 1.99 *avg* CITC = .42	α = .83 SEM = 1.96 *avg* CITC = .37	α = .75 SEM = 2.10 *avg* CITC = .32	α = .75 SEM = 2.04 *avg* CITC = .30
Observed Difficulty	*M* = 12.29 *SD* = 5.48 avg *p* = .51	*M* = 13.29 *SD* = 4.21 avg *p* = .54	*M* = 15.00 *SD* = 5.15 avg *p* = .63	*M* = 11.96 *SD* = 4.76 avg *p* = .50	*M* = 10.32 *SD* = 4.19 avg *p* = .43	*M* = 9.15 *SD* = 4.08 avg *p* = .38
Group Score Differences	F-M *d* = .21 B-W *d* = -.39 H-W *d* = -.17 A-W *d* = .36	F-M *d* = .31 B-W *d* = -.31 H-W *d* = -.15 A-W *d* = .17	F-M *d* = .22 B-W *d* = -.58 H-W *d* = -.21 A-W *d* = .00	F-M *d* = .16 B-W *d* = -.20 H-W *d* = -.19 A-W *d* = .12	F-M *d* = .10 B-W *d* = -.34 H-W *d* = -.22 A-W *d* = .59	F-M *d* = .22 B-W *d* = -.08 H-W *d* = .04 A-W *d* = .24
Completion Time (30 minutes or less)	*M* = 12.74 *SD* = 5.86	*M* = 10.88 *SD* = 4.94	*M* = 11.36 *SD* = 5.14	*M* = 12.39 *SD* = 5.79	*M* = 13.54 *SD* = 6.18	*M* = 13.24 *SD* = 6.36
Perceived Difficulty	*M* = 3.92 *SD* = .95	*M* = 3.89 *SD* = .92	*M* = 3.98 *SD* = .89	*M* = 3.90 *SD* = .95	*M* = 3.50 *SD* = .96	*M* = 3.37 *SD* = 1.00

Note. NOTAC = None of these are correct. SEM = Standard error of measurement. CITC = Corrected item-total correlation.

PHASE 1 PILOT: SUMMING UP (COMPLETED ≤ 30 MINUTES WITH HS DEGREE/GED/< 1 YR OF COLLEGE, *N* = 1,200)

Metric	Transform Only 8	Transform Only 3 + NOTAC	Transform Only 4	Transform Only 4 + NOTAC	Transform + Logic Grouped	Transform + Logic Scrambled
Unidimensionality	--	--	--	--	--	--
Reliability	α = .88 SEM = 1.78 *avg* CITC = .39	α = .79 SEM = 1.78 *avg* CITC = .28	α = .86 SEM = 1.92 *avg* CITC = .40	α = .84 SEM = 1.76 *avg* CITC = .37	α = .67 SEM = 2.11 *avg* CITC = .27	α = .69 SEM = 1.98 *avg* CITC = .26
Observed Difficulty	*M* = 11.60 *SD* = 5.15 avg *p* = .48	*M* = 12.56 *SD* = 3.89 avg *p* = .52	*M* = 13.89 *SD* = 5.12 avg *p* = .58	*M* = 11.66 *SD* = 4.39 avg *p* = .49	*M* = 9.71 *SD* = 3.68 avg *p* = .40	*M* = 8.27 *SD* = 3.56 avg *p* = .35
Group Score Differences	F-M *d* = .20	F-M *d* = .26	F-M *d* = .26	F-M *d* = .18	F-M *d* = -.02	F-M *d* = .34
Completion Time (30 minutes or less)	*M* = 12.32 *SD* = 5.70	*M* = 10.59 *SD* = 4.91	*M* = 11.34 *SD* = 5.41	*M* = 12.04 *SD* = 5.51	*M* = 13.24 *SD* = 6.16	*M* = 12.79 *SD* = 6.21
Perceived Difficulty	*M* = 3.83 *SD* = .97	*M* = 3.79 *SD* = .93	*M* = 3.86 *SD* = .91	*M* = 3.83 *SD* = .96	*M* = 3.41 *SD* = .97	*M* = 3.26 *SD* = 1.00

Note. NOTAC = None of these are correct. SEM = Standard error of measurement. CITC = Corrected item-total correlation.

PHASE 2: RECOMMENDED ITEM SPECS

- Transformation items only
- No NOTAC response option
- Four response options
  - Refine item difficulty model and item selection to ensure appropriate level of difficulty and minimize group score differences by race-ethnicity, where feasible

PHASE 2: PROPOSAL FOR PILOT

Objective

- Collect data on refined pool of CR items representative of the population of CR items with a participant sample representative of military applicants
- Use results to refine CR item specs and select potential pool of CR items for follow-on research (Computational Thinking) and field testing

Design and Measures

- 125+ CR items, multiple forms (5 or more forms, 25 items each, with a subset of items common to all forms)
- General mental ability (GMA) test (e.g., retired form or items from APT/PiCAT)
- Post-test questionnaire (demographics, perceived difficulty of items, test-taking experience)
- Two attention check items

Sample

- Non-military sample, ages 18–35, U.S. citizen, HS Degree/GED/<1 year of college
- Targeted *N* = 3,000 participants (~ 600 participants per form)

Method

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

ACKNOWLEDGMENTS

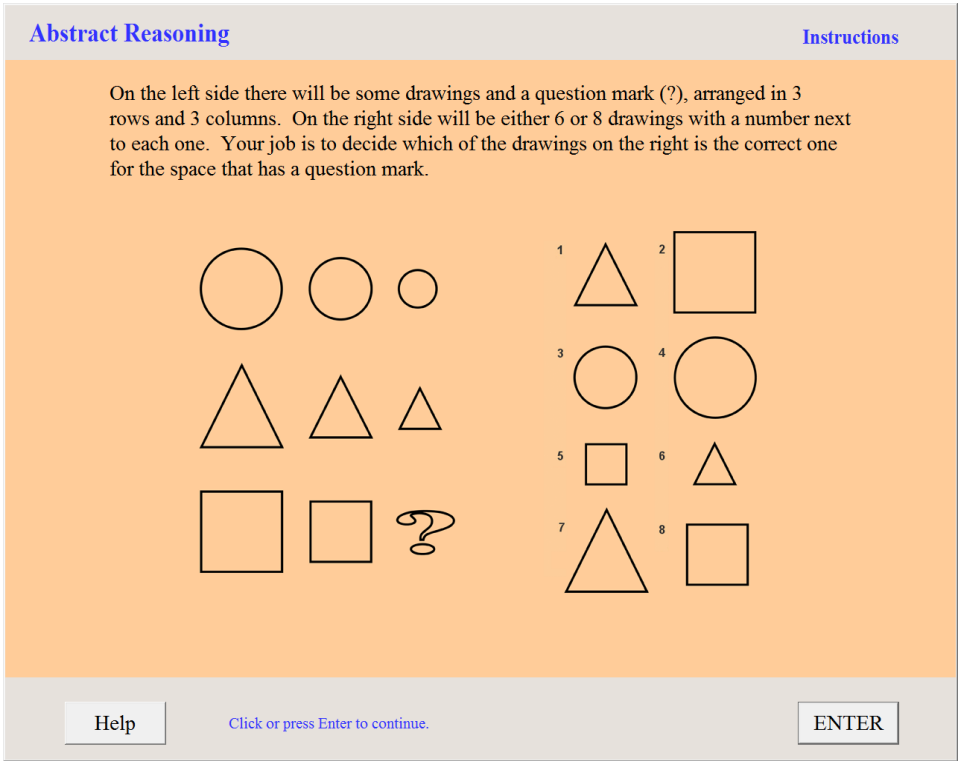
Scott Oppler, *HumRRO*

Matthew Brown, *HumRRO*

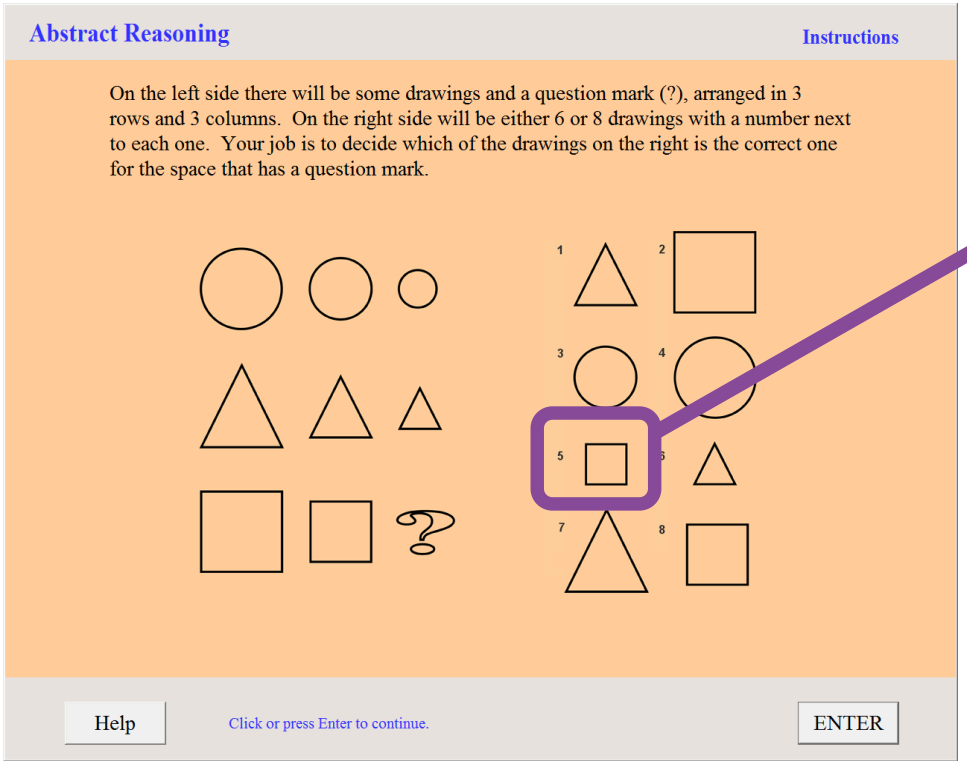
Katherine Klein, *HumRRO*

Backup Slides

INITIAL DEVELOPMENT EFFORT: SAMPLE ABSTRACT REASONING TEST (ART) ITEM #1

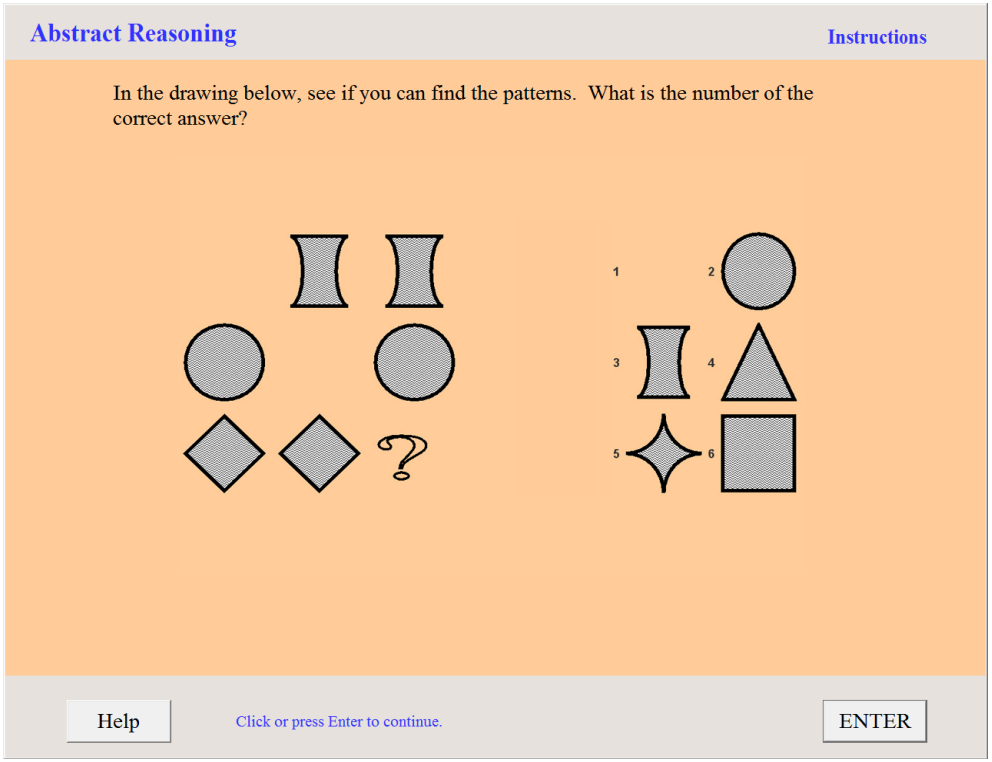


INITIAL DEVELOPMENT EFFORT: SAMPLE ABSTRACT REASONING TEST (ART) ITEM #1

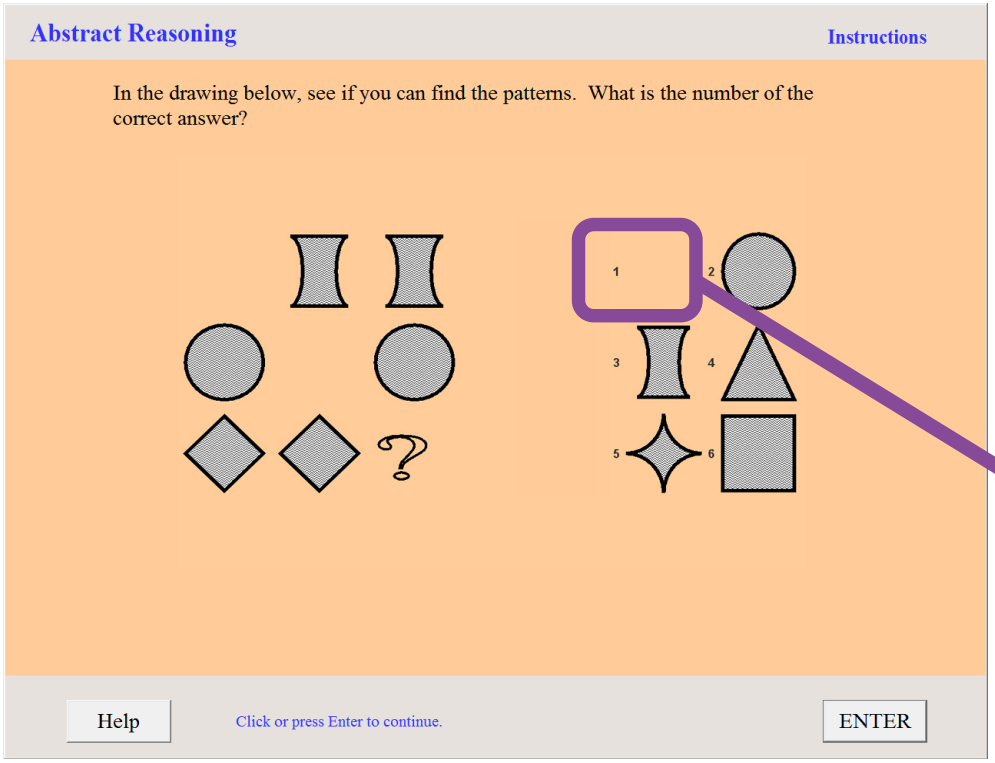


Option #5 is correct answer

INITIAL DEVELOPMENT EFFORT: SAMPLE ABSTRACT REASONING TEST (ART) ITEM #2



INITIAL DEVELOPMENT EFFORT: SAMPLE ABSTRACT REASONING TEST (ART) ITEM #2

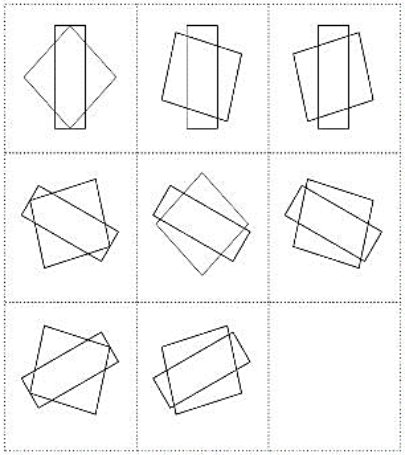


Option #1 is correct answer

PHASE 1 PILOT: ITEMS

Sample Transformation Item

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

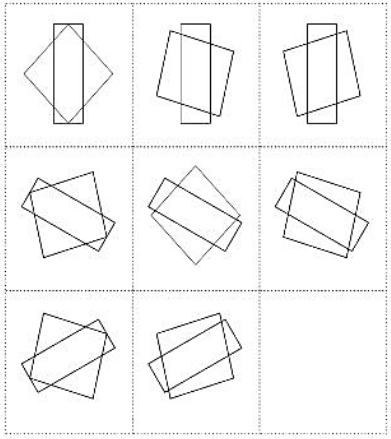


Which of the following images best completes the pattern(s) in the grid?

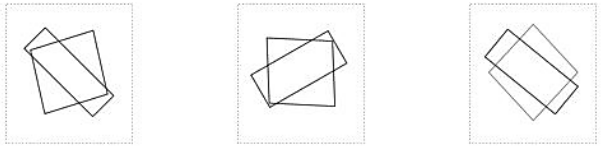
PHASE 1 PILOT: ITEMS

Sample Transformation Item

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



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

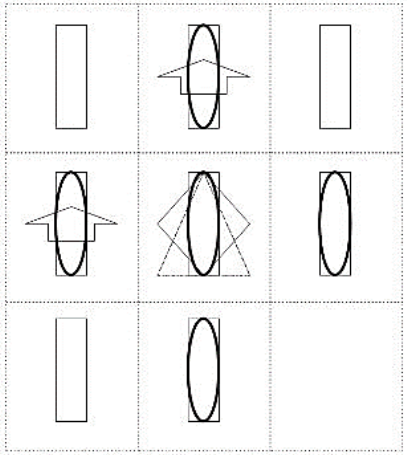


Option #4 is correct answer

PHASE 1 PILOT: ITEMS

Sample Logic Item

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

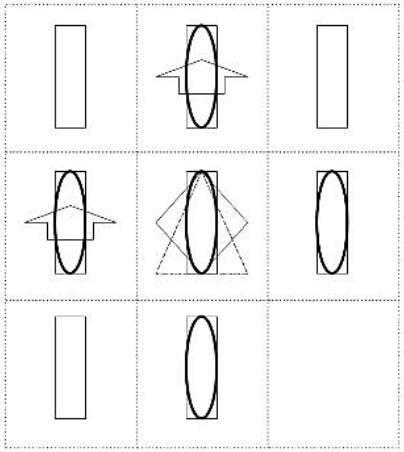


Which of the following images best completes the pattern(s) in the grid?

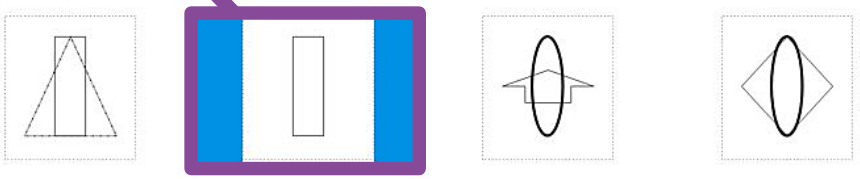
PHASE 1 PILOT: ITEMS

Sample Logic Item

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



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



Option #2 is correct answer

PHASE 1 PILOT: PRACTICE-ORDER EFFECTS (TRANSFORMATION ONLY CONDITIONS)

Condition Label	Condition Description	*n*	*M*	*SD*	Min	Max
C1	3 + NOTAC	260	13.29	4.21	4	22
C2	3 + NOTAC	241	12.87	4.12	1	22
C3	4	249	15.00	5.15	0	24
C4	4	254	15.28	5.21	2	24
C5	4 + NOTAC	245	12.12	4.64	2	24
C6	4 + NOTAC	245	11.96	4.76	1	24
C7	8	239	12.50	5.69	0	23
C8	8	254	12.09	5.27	0	24
C9 & C10	Transform 1st, Logic 2nd	507	10.55	4.16	1	24
C11 & C14	Scrambled	492	9.15	4.08	1	23
C12 & C13	Logic 1st, Transform 2nd	505	10.09	4.23	0	24

- Observed *no* significant effects for:
  - Block order, $F(1, 1979) = 0.68$, $p = .41$
  - Item block, $F(1, 1979) = 1.73$, $p = .19$
  - Block order * Item block, $F(3, 1979) = 1.14$, $p = .65$
  - Response option condition * Block order * Item block, $F(3, 1979) = 0.89$, $p = .44$
- Observed significant effects for:
  - Response option condition, $F(3, 1979) = 40.96$, $p < .001$
  - Response option condition * Item block, $F(1, 1979) = 38.55$, $p < .001$
    - Item block effects were only observed in two of four response option conditions
      - 3 + NOTAC ($d = .31$ favoring Block 1)
      - 8 option conditions ($d = .22$ favoring Block 2)

PHASE 1 PILOT: PRACTICE-ORDER EFFECTS (TRANSFORMATION + LOGIC CONDITIONS)

Condition Label	Condition Description	*n*	*M*	*SD*	Min	Max
C1	3 + NOTAC	260	13.29	4.21	4	22
C2	3 + NOTAC	241	12.87	4.12	1	22
C3	4	249	15.00	5.15	0	24
C4	4	254	15.28	5.21	2	24
C5	4 + NOTAC	245	12.12	4.64	2	24
C6	4 + NOTAC	245	11.96	4.76	1	24
C7	8	239	12.50	5.69	0	23
C8	8	254	13.09	5.27	0	24
C9 & C10	Transform 1st, Logic 2nd	507	10.55	4.16	1	24
C11 & C14	Scrambled	492	9.15	4.08	1	23
C12 & C13	Logic 1st, Transform 2nd	505	10.09	4.23	0	24

- Observed *no* significant effects for:
  - Logic block order, $F(1, 1498) = 0.30$, $p = .86$
  - Response option conditions, $F(3, 1498) = 2.14$, $p = .12$
  - Logic block order * Logic item block, $F(1, 1498) = 0.08$, $p = .77$
  - Logic block order * response option condition, $F(2, 1498) = 0.66$, $p = .52$
  - Logic item block * response option condition, $F(2, 1498) = 0.09$, $p = .91$
  - Response option condition * Logic block order * Logic item block, $F(2, 1498) = 0.40$, $p = .67$
- Observed significant effect for:
  - Logic item block, $F(1, 1498) = 7.08$, $p = .01$

PHASE 1 PILOT: DIMENSIONALITY

Modified Parallel Analyses

- Compared observed, second-factor eigenvalues to simulated second-factor eigenvalues (from 100 randomly generated samples) to determine whether the items were unidimensional
- Transformation items met the criteria for unidimensionality in the Transformation only conditions, except for the NOTAC conditions (3 + NOTAC, 4 + NOTAC)
- Items in all three Transformation + Logic conditions were not unidimensional, presentation order made no difference

Condition Label	Condition Description	*n*	Observed Eigenvalue	Simulated Eigenvalue	*p*
C1 + C2	3 + NOTAC	501	3.32	1.78	.01
C3 + C4	4	503	1.80	1.37	.13
C5 + C6	4 + NOTAC	490	3.65	1.82	.01
C7 + C8	8	493	1.81	1.47	.06
C9 + C10	Transform 1st, Logic 2nd	507	2.31	1.29	.01
C11 + C14	Scrambled	492	2.07	1.20	.02
C12 + C13	Logic 1st, Transform 2nd	505	2.58	1.23	.01

PHASE 1 PILOT: DIMENSIONALITY (CONTINUED)

Confirmatory Factor Analysis

- **Transformation & Logic Items (responses from Transformation + Logic conditions, combined)**
  - Two-factor model provided significantly superior model fit compared to a single-factor model ($\Delta\chi^2$ (1) = 406.555, ΔCFI = .08, ΔRMSEA = .01)

Model	Model χ^2	df	CFI	RMSEA	SRMR	$\Delta\chi^2$	ΔCFI	ΔRMSEA
One-factor	1297.65	252	.79	.05	.05			
Two-factor	891.09	251	.87	.04	.05	406.55	.08	.01

- **NOTAC Items (responses from 3 + NOTAC and 4+ NOTAC, combined)**
  - Two-factor model provided better overall model fit compared to the one-factor model ($\Delta\chi^2$ (1) = 255.28, ΔCFI = .08, ΔRMSEA = .01)
    - Second factor contained all six items where the NOTAC options was the keyed (correct) answer
  - Model fit was further improved from the two-factor model by cross-loading all six items to both latent factors but constraining the correlation between factors ($\Delta\chi^2$ (5) = 95.16, ΔCFI = .03, ΔRMSEA = .00)

Model	Model χ^2	df	CFI	RMSEA	SRMR	$\Delta\chi^2$	ΔCFI	ΔRMSEA
One-factor	1065.98	252	.74	.06	.08			
Two-factor	810.70	251	.82	.05	.08	255.28	.08	.01
Alt Two-factor *	742.32	246	.85	.05	.07	68.38	.03	.00

PHASE 1 PILOT: INTERNAL CONSISTENCY (ALL PARTICIPANTS)

The NOTAC and Transformation + Logic items were less internally consistent.

- The four-response option and eight-response option conditions showed slightly higher internal consistency (α = .85 - .87) than the NOTAC conditions (α = .78 - .83) [CITCs (avg *r* = .42 - .46) compared to the NOTAC items (avg *r* = .32 - .37)]
- Internal consistency estimates for the Transformation + Logic items were < .80, regardless of presentation order (α's = .74 - .76)
- Among the NOTAC conditions, performance on the six items where NOTAC was the keyed (correct) response were less correlated with performance on the other NOTAC items
  - CITCs (avg *r* = .47, ranging from .29 to .60) compared to the NOTAC items (avg *r* = .20, ranging between -.19 and .48)

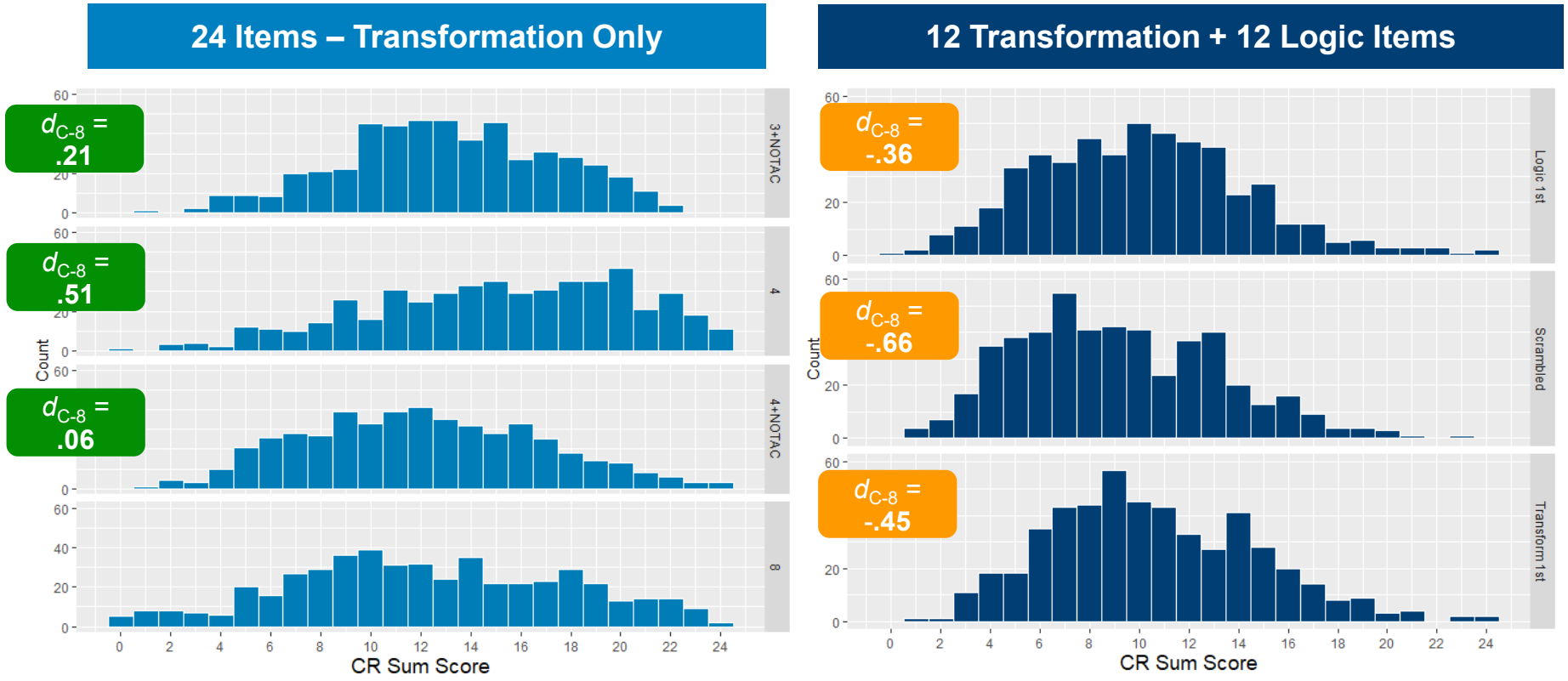
Statistic	3 + NOTAC	4	4 + NOTAC	8	Transform 1st, Logic 2nd	Scrambled	Logic 1st, Transform 2nd
α	**.78**	**.85**	**.83**	**.87**	**.74**	**.75**	**.76**
Avg CITC	.32	.42	.37	.44	.31	.30	.33
Min CITC	-.19	.23	.01	.23	.10	.07	.14
5th Pct	-.12	.25	.05	.27	.11	.07	.15
25th Pct	.22	.34	.33	.41	.29	.22	.22
50th Pct	.33	.42	.38	.45	.34	.30	.35
75th Pct	.44	.50	.47	.48	.40	.41	.39
95th Pct	.51	.60	.60	.60	.53	.50	.56
Max CITC	.51	.60	.60	.60	.54	.51	.56

(COMPLETED IN <30 MINUTES AND HS DEGREE/GED/< 1 YR OF COLLEGE)

Same pattern of results as for All Participants, internal consistency estimates were lower for the Transformation + Logic items (α's ≤ .70).

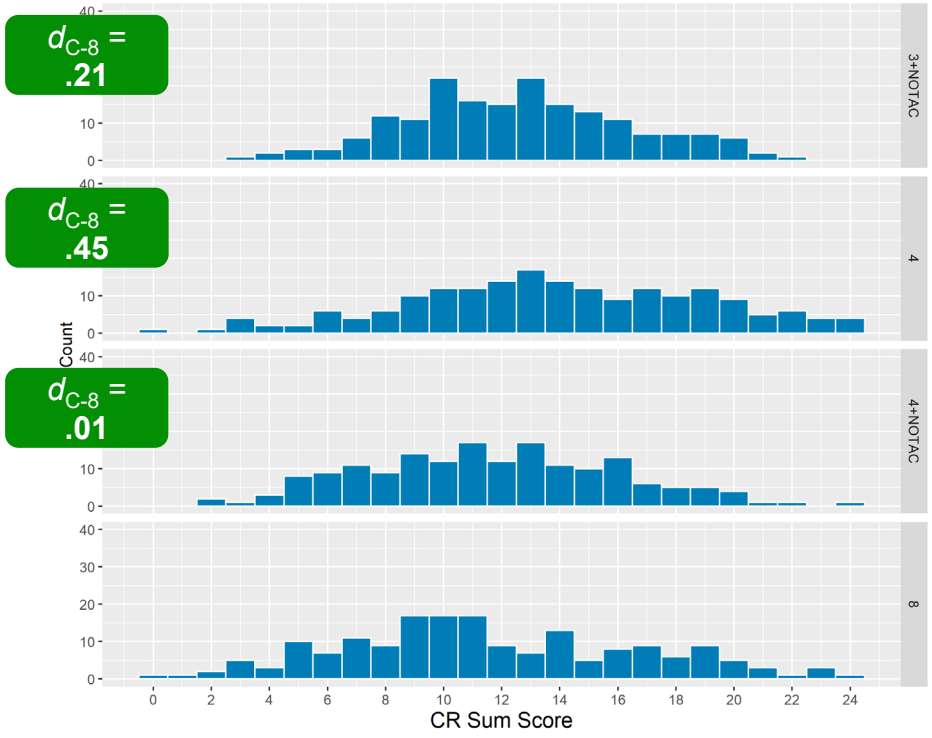
Statistic	3 + NOTAC	4	4 + NOTAC	8	Transform 1st, Logic 2nd	Scrambled	Logic 1st, Transform 2nd
α	**.79**	**.86**	**.84**	**.88**	**.64**	**.69**	**.70**
Avg CITC	.28	.41	.37	.40	. 25	.26	.28
Min CITC	-.28	.21	-.05	.19	-.03	.06	.01
5th Pct	-.28	.21	-.04	.19	-.02	.07	.02
25th Pct	.13	.33	.30	.31	.13	.15	.18
50th Pct	.33	.40	.38	.39	.25	.25	.26
75th Pct	.40	.50	.46	.46	.35	.36	.40
95th Pct	.54	.56	.65	.61	.55	.39	.58
Max CITC	.54	.56	.65	.61	.58	.47	.58

PHASE 1 PILOT: TEST SCORE DISTRIBUTIONS BY CONDITION (ALL PARTICIPANTS)

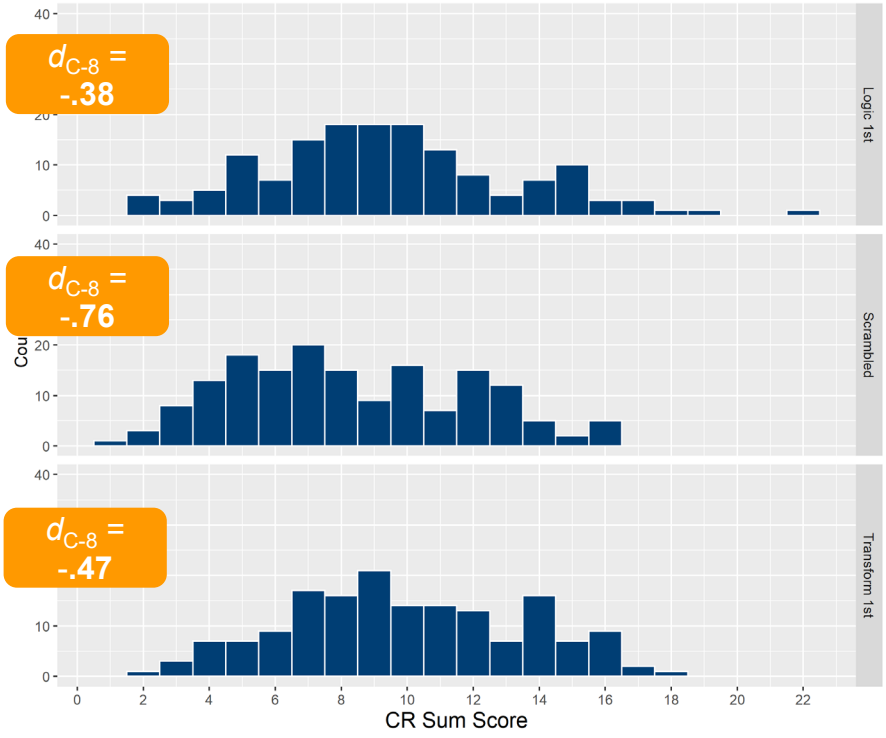


PHASE 1 PILOT: TEST SCORE DISTRIBUTIONS BY CONDITION (COMPLETED IN ≤ 30 MINUTES AND HS DEGREE/ GED/< 1 YR OF COLLEGE)

24 Items – Transformation Only

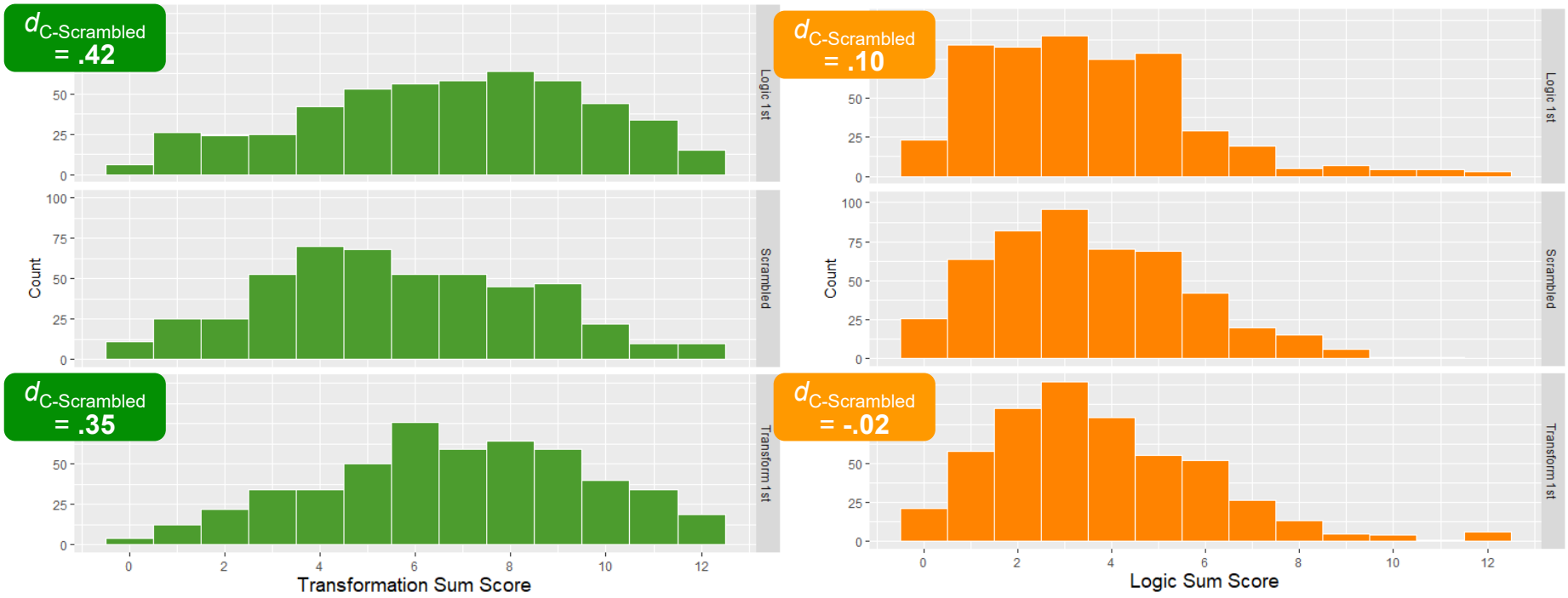


12 Transformation + 12 Logic Items



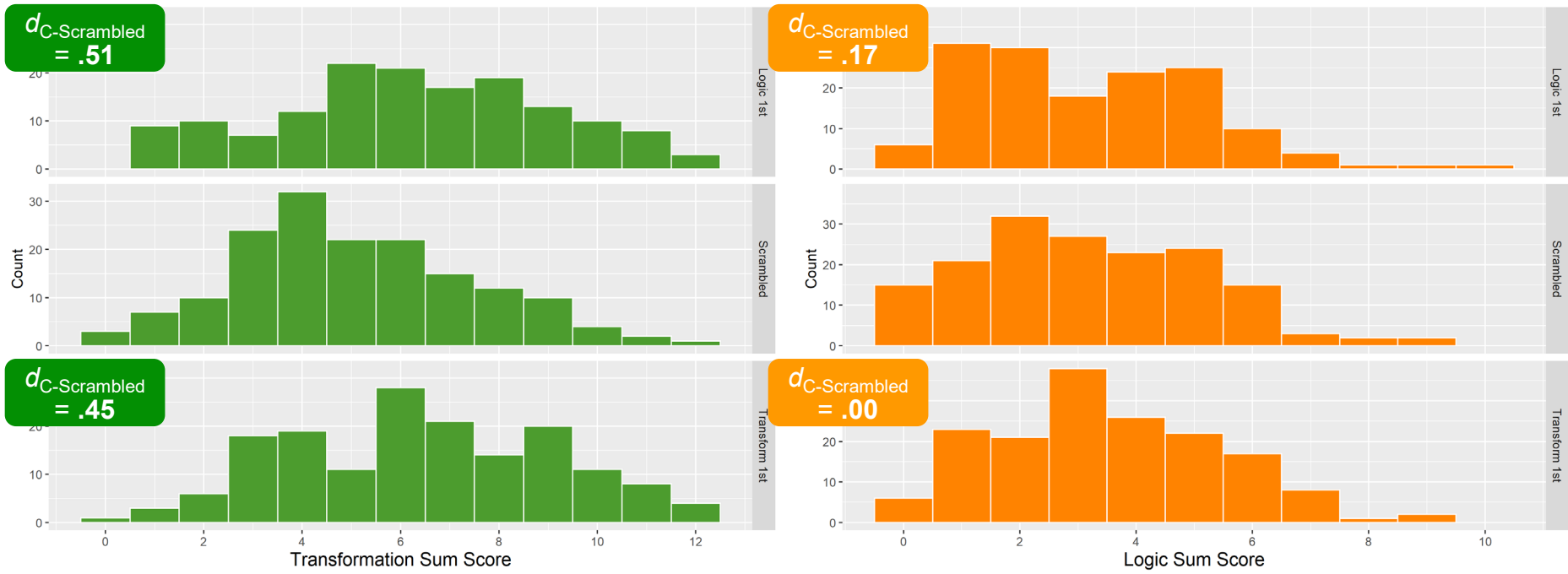
PHASE 1 PILOT: TRANSFORMATION AND LOGIC SUBSCORE DISTRIBUTIONS BY CONDITION

All Participants



PHASE 1 PILOT: TRANSFORMATION AND LOGIC SUBSCORE DISTRIBUTIONS BY CONDITION (CONTINUED)

Completed in ≤ 30 minutes and HS Degree/ GED/< 1 yr of College



PHASE 1 PILOT: ITEM DIFFICULTY BY TRANSFORMATION ONLY CONDITION

All Participants

- Items in the four-response option items were less difficult (avg *p* = .63) compared to:
  - 3 + NOTAC conditions (avg *p* = .54)
  - 8-response option conditions (avg *p* = .51)
  - 4 + NOTAC conditions (avg *p* = .50)
- Among NOTAC conditions, the six items where NOTAC was keyed (correct) response were more difficult
  - Item difficulties (avg *p* = .53) compared to items where NOTAC was keyed response (avg *p* = .23)

Statistic	3 + NOTAC	4	4 + NOTAC	8
Average *p*	**.54**	**.63**	**.50**	**.51**
Min *p*	.07	.29	.08	.15
5th Pct	.10	.32	.10	.17
25th Pct	.37	.51	.33	.36
50th Pct	.52	.60	.46	.46
75th Pct	.75	.71	.67	.59
95th Pct	.94	.97	.96	.94
Max *p*	.94	.97	.96	.94

Completed in 30 Mins or Less with HS Degree/GED/< 1 yr of College

The four-response option items were less difficult (avg *p* = .58) compared to:

- 3 + NOTAC conditions (avg *p* = .52)
- 4 + NOTAC conditions (avg *p* = .49)
- 8-response option conditions (avg *p* = .48)

Statistic	3 + NOTAC	4	4 + NOTAC	8
Average *p*	**.52**	**.58**	**.49**	**.48**
Min *p*	.05	.20	.05	.14
5th Pct	.08	.23	.08	.15
25th Pct	.30	.46	.29	.31
50th Pct	.46	.52	.38	.40
75th Pct	.62	.59	.57	.49
95th Pct	.95	.92	.93	.92
Max *p*	.95	.96	.98	.97

Note. Results calculated after screening participants for completion times (equal to or less than 30 minutes) and highest educational attainment (HS Degree/GED/< 1 yr of College)

PHASE 1 PILOT: ITEM DIFFICULTY BY TRANSFORMATION + LOGIC CONDITION (ALL PARTICIPANTS)

Transformation Items

- Items were more difficult in the Scrambled condition (avg *p* = .47) than in the grouped conditions:
  - Transformation 1st condition (avg *p* = .57)
  - Logic 1st condition (avg *p* = .55)

Transformation Items	Transform 1st, Log 2nd	Logic 1st, Transform 2nd	Scrambled
Average *p*	**.57**	**.55**	**.47**
Min *p*	.25	.24	.19
5th Pct	.31	.29	.23
25th Pct	.44	.42	.31
50th Pct	.69	.68	.60
75th Pct	.54	.50	.43
90th Pct	.91	.90	.89
Max *p*	.96	.96	.94

Logic Items

- Items were more difficult than the Transformation items, regardless of presentation order:
  - Transformation 1st condition (avg *p* = .31)
  - Logic 1st condition (avg *p* = .29)
  - Scrambled condition (avg *p* = .29)

Logic Items	Transform 1st, Log 2nd	Logic 1st, Transform 2nd	Scrambled
Average *p*	**.31**	**0.29**	**0.29**
Min *p*	.21	.19	.14
5th Pct	.22	.20	.14
25th Pct	.23	.22	.24
50th Pct	.35	.33	.37
75th Pct	.35	.33	.37
90th Pct	.47	.43	.47
Max *p*	.54	.46	.51

PHASE 1 PILOT: ITEM DIFFICULTY BY TRANSFORMATION + LOGIC CONDITION (COMPLETED ≤ 30 MINS WITH HS DEGREE/GED/< 1 YR OF COLLEGE)

Transformation Items

- Items were more difficult in the Scrambled condition (avg *p* = .42) than in the grouped conditions:
  - Transformation 1st condition (avg *p* = .54)
  - Logic 1st condition (avg *p* = .52)

Transformation Items	Transform 1st, Log 2nd	Logic 1st , Transform 2nd	Scrambled
Average *p*	**.54**	**.52**	**.42**
Min *p*	.23	.16	.16
5th Pct	.27	.23	.16
25th Pct	.39	.38	.20
50th Pct	.52	.51	.29
75th Pct	.64	.63	.41
95th Pct	.89	.88	.88
Max *p*	.96	.97	.97

Logic Items

- Items were more difficult than the Transformation items, regardless of presentation order:
  - Transformation 1st condition (avg *p* = .29)
  - Logic 1st condition (avg *p* = .26)
  - Scrambled condition (avg *p* = .26)

Logic Items	Transform 1st, Log 2nd	Logic 1st , Transform 2nd	Scrambled
Average *p*	**.29**	**.26**	**.26**
Min *p*	.18	.15	.11
5th Pct	.19	.16	.11
25th Pct	.23	.18	.15
50th Pct	.27	.25	.20
75th Pct	.32	.32	.32
95th Pct	.44	.40	.43
Max *p*	.54	.42	.47

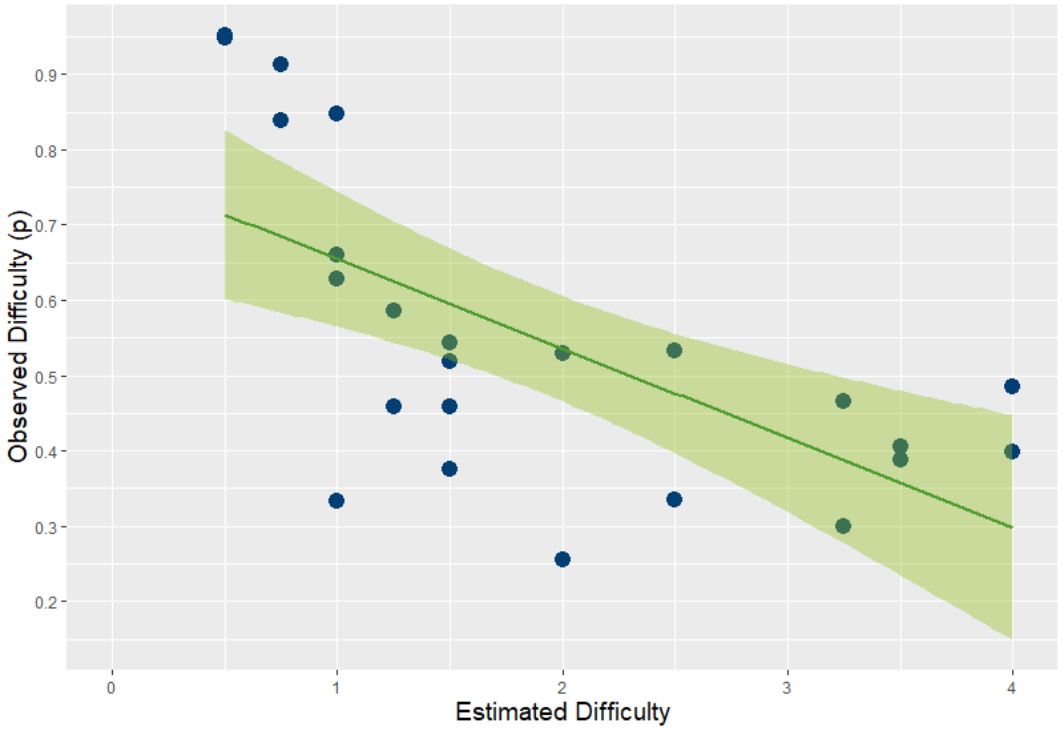
PHASE 1 PILOT: OBSERVED ITEM DIFFICULTY BY ESTIMATED DIFFICULTY (TRANSFORMATION ONLY)

Estimated item difficulty based on item features:

- Number of transformations
  - Vertical or horizontal
  - Diagonal (top-down or bottom-up)
- Rotation

Estimated item difficulty correlated with observed item difficulty at r = -0.64

- 95% CI = -0.84, -0.32



PHASE 1 PILOT: GROUP SCORE DIFFERENCES BY CONDITION

All Participants

Gender

- Female participants consistently scored higher than male participants across all conditions, on average (observed *d's* = .06 to .31).

Race and Ethnicity

- Asian participants scored equal to or higher than White, non-Hispanic participants, on average (observed *d's* = .00 to .80).
- Black participants scored lower than White, non-Hispanic participants, on average (observed *d's* = -.08 to -.58).
- White, Hispanic participants also scored lower than White, non-Hispanic participants in all but the Scrambled condition (observed *d's* = .04 to -.21).

Condition Description	Female	Asian, non-Hispanic	Black, non-Hispanic	White, Hispanic
3 + NOTAC	.31 (.35)	.17 (.22)	-.31 (-.40)	-.15 (-.19)
4	.22 (.27)	.00 (.00)	-.58 (-.69)	-.21 (-.25)
4 + NOTAC	.16 (.18)	.12 (.14)	-.20 (.24)	-.19 (-.23)
8	.21 (.24)	.36 (.42)	-.39 (-.45)	-.17 (-.19)
Transform 1st, Logic 2nd	.14 (.18)	.38 (.52)	-.40 (-.54)	-.15 (-.20)
Scrambled	.22 (.29)	.24 (.32)	-.08 (-.11)	.04 (.05)
Logic 1st, Transform 2nd	.06 (.08)	.80 (1.12)	-.28 (-.37)	-.29 (-.38)

Note. *d*'s corrected for measurement error in the CR items are reported in the parentheses.

Completed in 30 Mins or Less with HS Degree/GED/< 1 yr of College

- Female participants consistently scored higher than male participants across all Transformation Only conditions, on average same as for all participants (observed *d's* = .18 to .26).
- Results were mixed on the Transformation + Logic conditions (observed *d*'s = -.03 to .34).

Condition Description	Female
3 + NOTAC	.26 (.33)
4	.26 (.30)
4 + NOTAC	.18 (.21)
8	.20 (.23)
Transform 1st, Logic 2nd	-.03 (-.05)
Scrambled	.34 (.50)
Logic 1st, Transform 2nd	-.01 (-.01)

Note. Results calculated after screening participants for completion times (equal to or less than 30 minutes) and highest educational attainment (HS Degree/GED/< 1 yr of College). *d*'s corrected for measurement error in the CR items are reported in the parentheses.

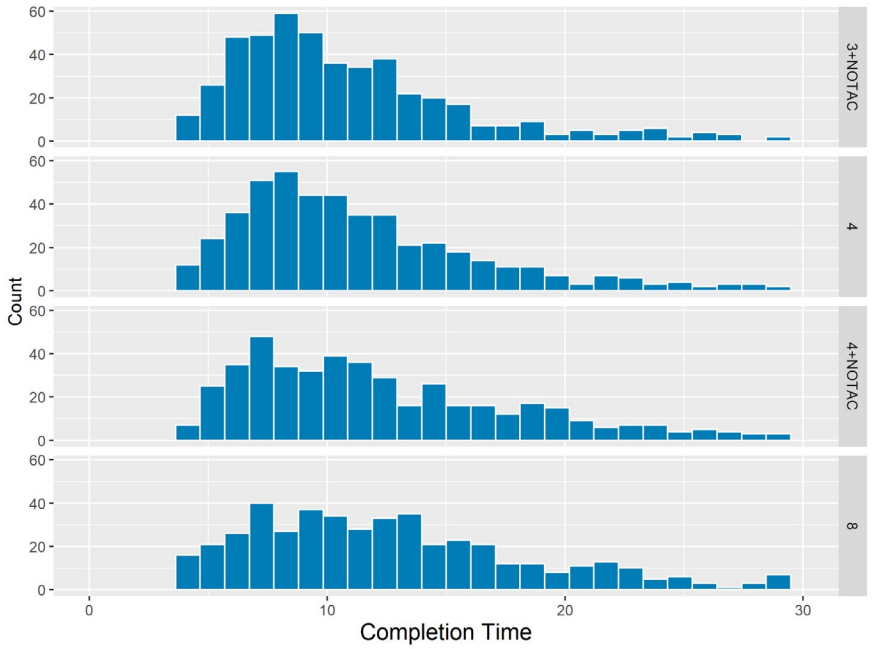
PHASE 1 PILOT: COMPLETION TIME BY CONDITION (ALL PARTICIPANTS, COMPLETED IN ≤ 30 MINUTES)

- Among the Transformation Only conditions, four-response options (with or without NOTAC) resulted in shorter average completion times than five- or eight-response options
- Transformation + Logic conditions required more time to complete than the Transformation Only conditions, on average, magnitude of the differences varied

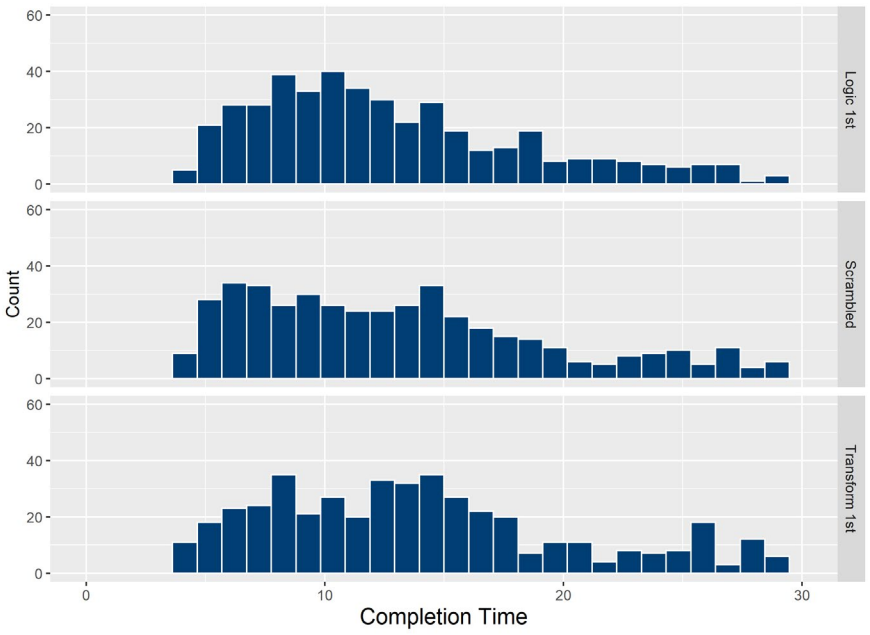
Condition Description	*n*	*M*	*SD*	$d_{C\text{-}8}$
3 options + NOTAC	468	10.88	4.94	-.34
4 options	474	11.36	5.14	-.25
4 options + NOTAC	453	12.39	5.79	-.06
8 options	455	12.74	5.86	--
Transform 1st , Logic 2nd	447	14.10	6.49	.22
Scrambled	438	13.24	6.36	.08
Logic 1st , Transform 2nd	440	12.98	5.87	.04

PHASE 1 PILOT: COMPLETION TIME (IN MINUTES) BY CONDITION (COMPLETED ≤ 30 MINUTES)

24 Items – Transformation Only

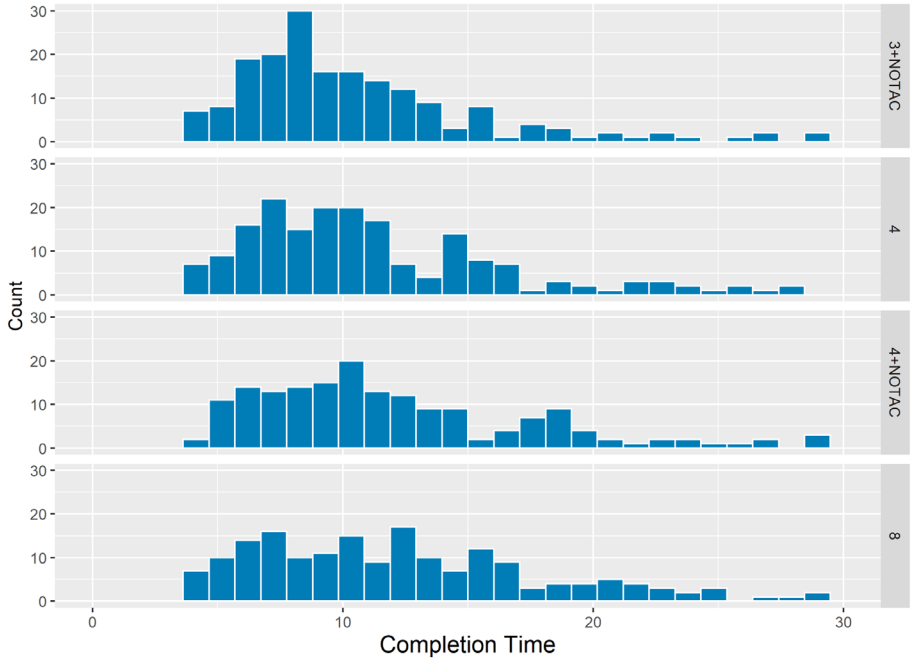


12 Transformation + 12 Logic Items

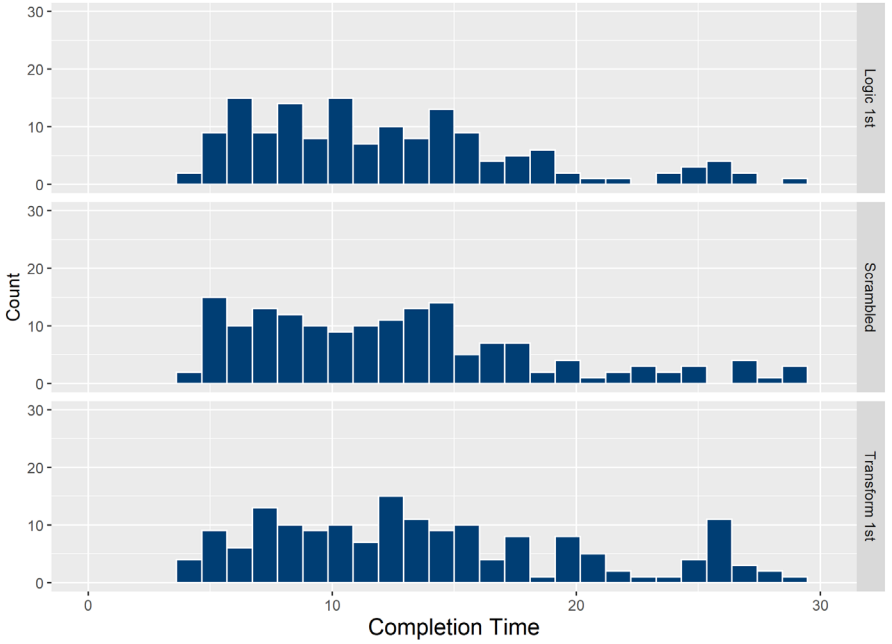


PHASE 1 PILOT: COMPLETION TIME (IN MINUTES) BY CONDITION (HS DEGREE/GED/< 1 YR OF COLLEGE)

24 Items – Transformation Only



12 Transformation + 12 Logic Items



PHASE 1 PILOT: PERCEIVED DIFFICULTY BY CONDITION

All Participants

- All participants reported their perceived performance on the CR items using a 5-point scale (1 = “Less than 20% of items correct” to 5 = “More than 80% of items correct”).
- Significant effect of experimental conditions on perceived CR performance, F (6, 3484) = 36.65, $p < .001$ (all participants, top table).
- Participants in Transformation Only conditions (3+NOTAC, 4, 4+NOTAC, and 8) reported more correct responses compared to participants in all three Transformation + Logic conditions (Transform 1st, Logic 1st, and Scrambled), on average.

Condition Description	*n*	*M*	*SD*	$d_{C\text{-}8}$
3 options + NOTAC	501	3.89	.92	-.03
4 options	503	3.98	.89	.07
4 options + NOTAC	490	3.90	.95	-.02
8 options	493	3.92	.95	N/A
Transform 1st , Logic 2nd	507	3.53	.96	-.41
Scrambled	492	3.37	1.00	-.56
Logic 1st , Transform 2nd	505	3.46	.95	-.48

Completed in 30 Mins or Less with HS Degree/GED/< 1 yr of College

- Observed similar effects among the subset of participants who completed in 30 minutes or less with HS Degree/GED/< 1 yr of College (bottom table).

Condition Description	*n*	*M*	*SD*	$d_{C\text{-}8}$
3 options + NOTAC	182	3.79	.93	-.04
4 options	188	3.86	.91	.03
4 options + NOTAC	172	3.83	.96	.00
8 options	179	3.83	.97	N/A
Transform 1st , Logic 2nd	164	3.59	.94	-.25
Scrambled	164	3.26	1.00	-.58
Logic 1st , Transform 2nd	151	3.21	1.00	-.63

Note. Results calculated after screening participants based on completion times (equal to or less than 30 minutes) and highest educational attainment (HS Degree/GED/< 1 yr of college)

PHASE 1 PILOT: PERCEIVED DIFFICULTY BY CONDITION AND GENDER, RACE, AND ETHNICITY

All Participants

Gender

- Significant difference in perceived performance by gender, F (2, 3488) = 21.78, p < .001.
- Male participants consistently reported scoring higher compared to female participants, avg d = .22.

Race and Ethnicity

- Differences in perceived performance among racial or ethnic groups were mixed and varied by condition.

Condition Description	Female	Asian, non-Hispanic	Black, non-Hispanic	White, Hispanic
3 + NOTAC	3.85 (.94)	3.97 (.77)	3.69 (.82)	3.92 (.93)
4	3.91 (.93)	4.03 (.95)	3.89 (.90)	3.77 (.88)
4 + NOTAC	3.84 (.94)	4.00 (.80)	3.87 (1.06)	3.86 (1.07)
8	3.88 (.91)	4.15 (.53)	3.72 (1.03)	3.88 (.96)
Transform 1st, Logic 2nd	3.39 (.96)	3.55 (.97)	3.78 (.90)	3.61 (.85)
Scrambled	3.24 (.96)	3.24 (1.00)	3.12 (1.12)	3.47 (.92)
Logic 1st, Transform 2nd	3.34 (.96)	3.75 (.76)	3.64 (.92)	3.51 (.87)

Completed in 30 Mins or Less with HS Degree/GED/< 1 yr of College

Gender

- Males reported consistently higher performance than females (avg d = .25).

Condition Description	Female
3 + NOTAC	3.72 (.96)
4	3.82 (.93)
4 + NOTAC	3.75 (.93)
8	3.78 (.99)
Transform 1st, Logic 2nd	3.40 (.95)
Scrambled	3.17 (.98)
Logic 1st, Transform 2nd	3.07 (.99)

Note. Results calculated after screening participants based on completion times (equal to or less than 30 minutes) and highest educational attainment (HS Degree/GED/< 1 yr of college)