AI Psychometrics Inc.

Theta AI™

Cohort Insight Report

5-option situational judgment · Best+Worst responses · Cohort

Sample cohort · 24 synthetic participants. This report shows how cohort results are presented. With fewer than 50 participants, use group patterns as descriptive evidence rather than population benchmarks.

Where this group's judgment is strong and where it wavers, where risky choices pile up, which items work, and who to ask what in the interview.

Prepared for SYNTHETIC-DEMO (not customer data)

Cohort SYNTH-N24-v3b (pseudonymous)

24 members · 45 pages · v6.1

Generated October 8, 2026

Pre-calibration · descriptive | key_version theta-key-2026-10-05 | scoring bw-s1-v1 | No composite rank · no verdict | Missing stays missing

02 Executive summary
03 Visual dashboard
04 Comparison evidence 1
05 Comparison evidence 2
07 Comparison evidence 3
09 Comparison evidence 4
11 Comparison evidence 5
13 Comparison evidence 6
14 Comparison evidence 7
15 Comparison evidence 8
17 Comparison evidence 9
18 Risk concentration
19 Competency profile
20 Development · onboarding
22 Interview questions
24 Interviewer guide in depth
26 Judgment landscape
27 Choice distributions 20×5
28 Distractor analysis
29 Item health · key review
32 Item discrimination story
34 Ambiguity map
35 Gap · inversions
36 Values priorities
37 Decision style
38 Consistency · type stability
39 Response integrity
40 Response time in depth
41 Roster
42 Method · limits
43 Competency interpretation
44 Appendix R

Verify this report
TH-ZNUN-8B3Y-P

© 2026 AI Psychometrics Inc. · ai-psychometrics.com

EXECUTIVE SUMMARY · DECISION PAGE

The one-minute conclusion for leaders. Every number comes from a chart or table later in the report.

The conclusion in one sentence

3 of 24 are above the cut on both axes and 12 are borderline. Harmful Best picks cluster in CB-05-D, CB-06-D, CB-09-A: start interviews there, and onboarding with 'People & collaboration'.

3 strengths · situations the group judged best

17 **CB-11-B** · mean 5.4/8 · model action = Best 38%
16 **CB-10-A** · mean 5.3/8 · model action = Best 29%
04 **CB-02-C** · mean 5.2/8 · model action = Best 38%

3 points to check · where harmful Best picks cluster

09 **CB-05-D** · harmful Best 13 (54%) · mean 4.0/8
11 **CB-06-D** · harmful Best 12 (50%) · mean 4.4/8
15 **CB-09-A** · harmful Best 12 (50%) · mean 4.4/8

3 questions to ask in interviews

09 "In a situation like CB-05-D, what did you do first, and why?" · strong answer leans toward "Politely decline, say it cannot be used, and tell compliance if needed"
11 "In a situation like CB-06-D, what did you do first, and why?" · strong answer leans toward "Ask one-on-one how they are settling in and set up ways to join, like a buddy"
15 "In a situation like CB-09-A, what did you do first, and why?" · strong answer leans toward "Stop the work, check the cause, act on it, and log it"

Who to ask and signs of a weak answer are on the 'Interview questions' page.

4 key findings

- Most solid in 'Execution under pressure', weakest in 'People & collaboration'.
- 24 people (100%) picked a harmful option as Best at least once; far-misses cluster in CB-05-D, CB-06-D, CB-09-A.
- 38.1% of responses are inverted, highest in CB-05-D at 50%.
- Response speed and judgment score are barely related (item r -0.40–0.39).

Is the evidence enough?

- **24** people · Small tier (20-49) · discrimination for reference only
- Split-half **0.88** (90% interval 0.78–0.94)
- Items working **20/20** (discrimination r ≥ 0.15 · reference only)
- **12** key-review items · review candidates
- Integrity signals: **1** people · for review, not proof
- Reversal review: 4 candidates · answers on most items scored Worst above Best. A review marker only; it does not say why.
- Not assessed: 0 candidates · the number of answered items did not fit the check, or an answer value was missing, so no judgment was made.
- Descriptive, pre-calibration (key theta-key-2026-10-05 · calibration unavailable)

What to do this week

- First onboarding topic: case reviews in 'People & collaboration'
- Where interview questions go first: CB-05-D, CB-06-D
- Review candidates: ambiguous or weak items 01, 02, 03, 05, 06, 07, 08, 09, 10, 13, 18, 19

Limits of the data

- 24-person live responses · descriptive before calibration
- No ranking, cut-off or composite score
- No demographic data, so no group fairness analysis
- Reliability figures describe this sample; they are not a validation result

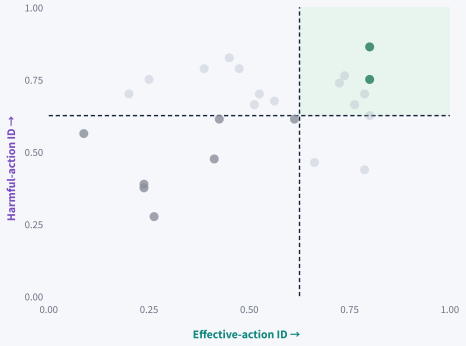
Pages to read before deciding

- **p18** Risk concentration
- **p20** Development · onboarding plan
- **p22** Interview question bank
- **p24** Interviewer guide in depth
- **p42** Method · tiers · limits

Descriptive · pre-calibration

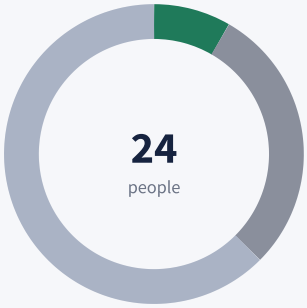
Of 24 people, 2 are above the cut on both axes and 15 are borderline. Risk concentrates most in the CB-05-D scenario.

Judgment landscape · 24 people



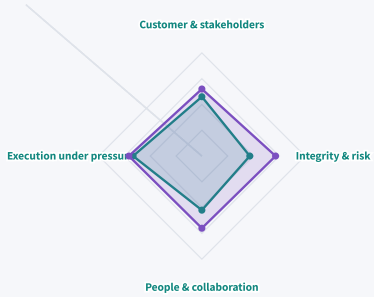
Dot color = type · faded = borderline (shown-type uncertainty rule) · green area = both axes at or above the 0.625 cut.

Type mix



- 2 Balanced
- 0 Strong on effective action, risk identification to develop
- 0 Strong on risk identification, effective action to develop
- 7 Insufficient evidence on both axes
- 15 Borderline

Competency profile · cohort median



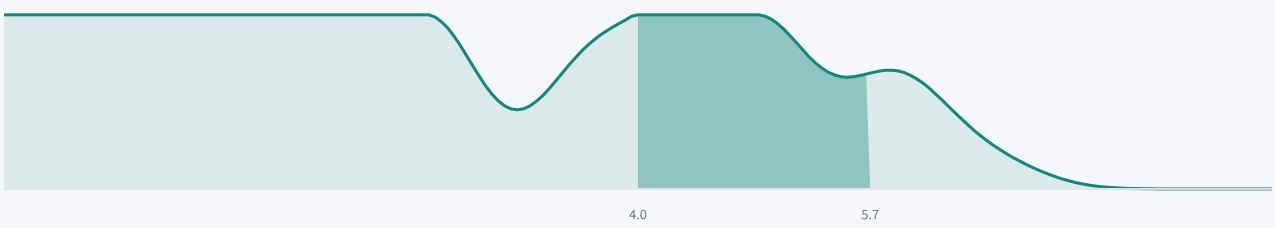
Teal = median Best credit, purple = median Worst credit (0–4) · competency tags are a sample mapping .

All 20 situations · score · risk · discrimination

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
Mean item score	4.5	4.2	4.9	5.2	4.6	4.3	4.5	4.4	4.0	4.3	4.4	5.0	4.3	4.8	4.4	5.3	5.4	4.1	4.7	5.0
Harmful as Best %	38	46	42	33	46	42	42	46	54	46	50	33	42	38	50	29	33	46	29	33
Discrimination r	0.70	0.44	0.71	0.42	0.35	0.45	0.70	0.38	0.42	0.46	0.61	0.37	0.36	0.45	0.58	0.59	0.46	0.26	0.50	0.55

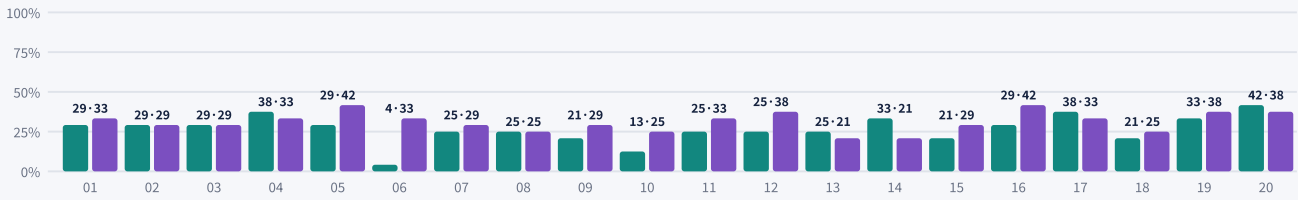
No low-discrimination items (all r ≥ 0.15).

Mean item score distribution



Dark part = middle 50%. No overall rank, no pass mark. (kernel density, h=0.35)

Agreement with the key by situation



Teal = share picking the model action as Best · purple = share picking the most harmful action as Worst.

RELATIVE-COMPARISON EVIDENCE · OVERVIEW DASHBOARD

24 members · 480 pairs · α 0.88. Type stability is 80%+ for 67%; 100% picked a harmful option at least once.

Headline indicators

Members	Response pairs	Cronbach α
24 members	480 items × members	0.88 CI 0.79–0.92

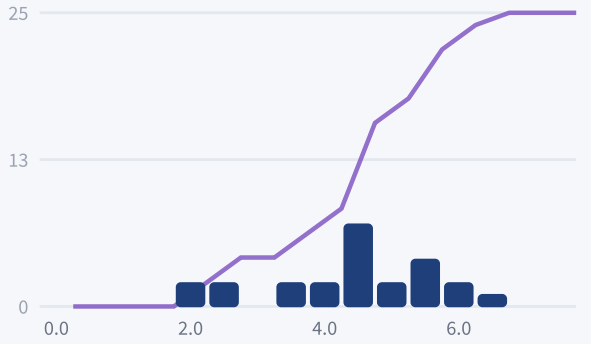
Split-half	Any harmful pick	Type stability ≥80%
0.88 CI 0.78–0.94	100.0% 24 members	16 members · 66.7%

Type and score distributions

Judgment-type distribution (N · share)

Balanced	2 · 8.3%
Insufficient evidence	7 · 29.2%
Execution strength	0 · 0.0%
Borderline	15 · 62.5%

Mean-score distribution · purple line = cumulative share



Median 4.9 · mean 4.61 · range 0-8

Stability bands

Type-stability bands (1,000 resamples) · 80% bar



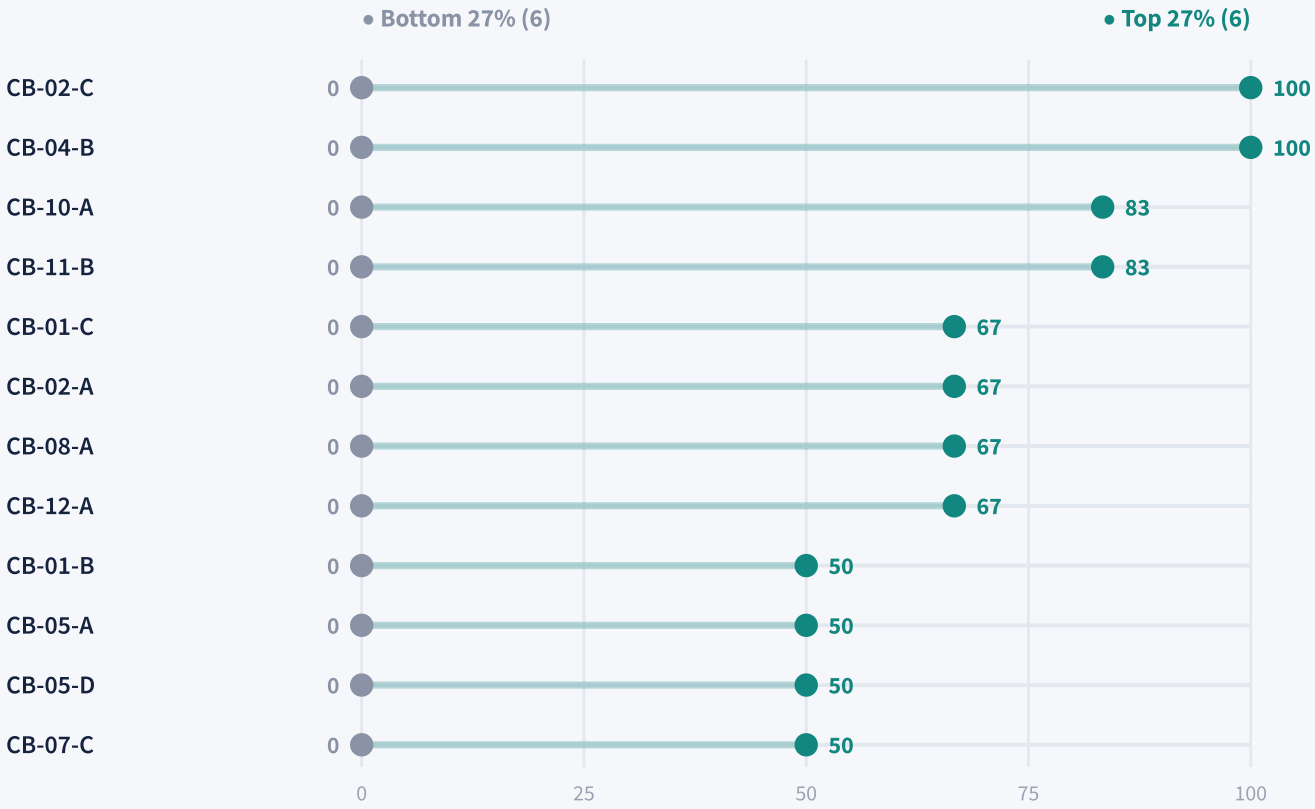
How to read this dashboard

• This page summarizes the eight deep-dive sections that follow (difficulty, means, harmful picks & inversions, discrimination, stability, axis alternation, bands, synthesis). Every figure is computed directly from the 24 × 20 response records.
• Figures describe the cohort; they cannot grade any person or produce a selection ranking.

RELATIVE-COMPARISON EVIDENCE · 20-SITUATION DIFFICULTY

The top 27% picked the model action at 90%+ in 2 of 20 situations. The widest between-group gap is CB-02-C (100pp).

Model-pick dumbbells · by situation

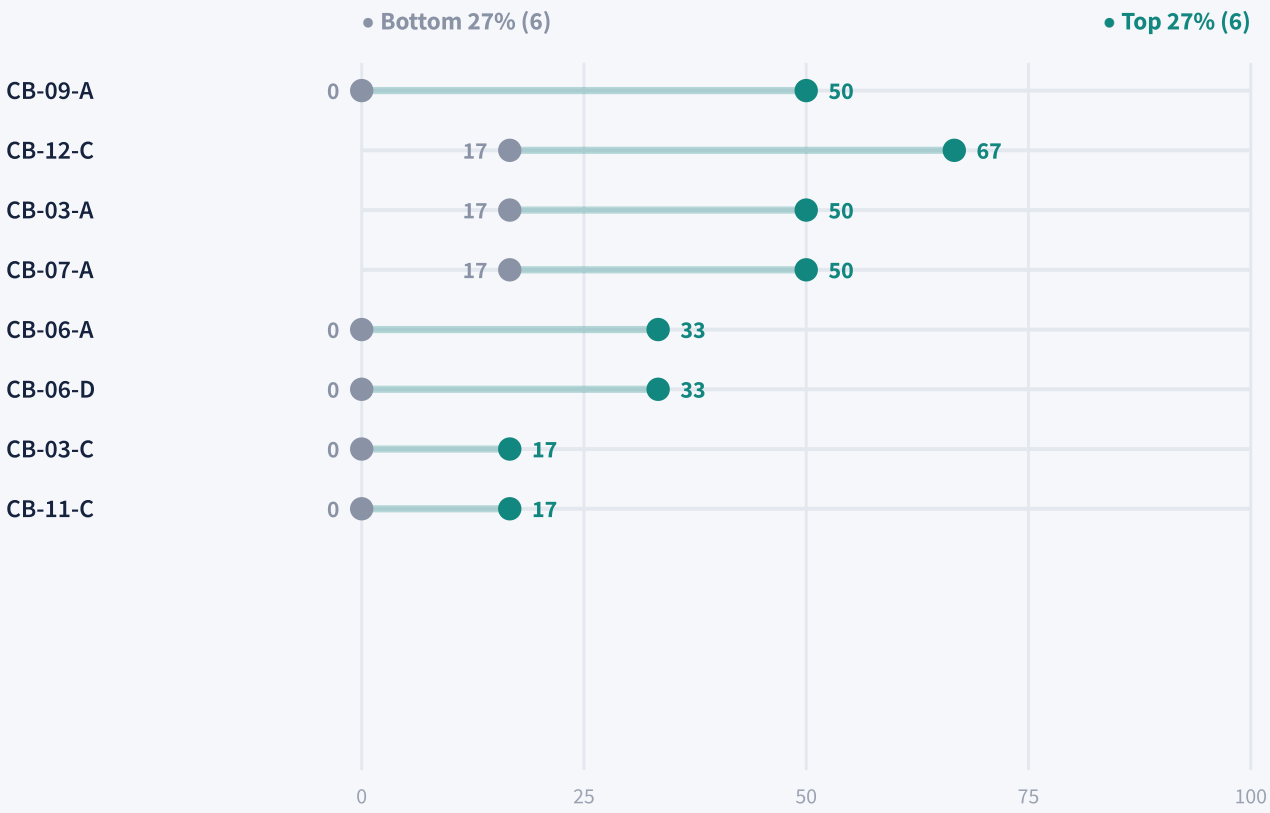


Figures

Situation (widest gap first)	Bottom 27%	Top 27%	Gap	Mean /8
CB-02-C	0%	100%	**100pp**	5.2
CB-04-B	0%	100%	**100pp**	4.5
CB-10-A	0%	83%	**83pp**	5.3
CB-11-B	0%	83%	**83pp**	5.4
CB-01-C	0%	67%	**67pp**	4.2
CB-02-A	0%	67%	**67pp**	4.9
CB-08-A	0%	67%	**67pp**	4.8
CB-12-A	0%	67%	**67pp**	4.7
CB-01-B	0%	50%	**50pp**	4.5
CB-05-A	0%	50%	**50pp**	4.4
CB-05-D	0%	50%	**50pp**	4.0
CB-07-C	0%	50%	**50pp**	4.3

Model-pick rates continued · gap rank 13-20.

Model-pick dumbbells · by situation



Figures

Situation (widest gap first)	Bottom 27%	Top 27%	Gap	Mean /8
CB-09-A	0%	50%	**50pp**	4.4
CB-12-C	17%	67%	**50pp**	5.0
CB-03-A	17%	50%	**33pp**	4.6
CB-07-A	17%	50%	**33pp**	5.0
CB-06-A	0%	33%	**33pp**	4.3
CB-06-D	0%	33%	**33pp**	4.4
CB-03-C	0%	17%	**17pp**	4.3
CB-11-C	0%	17%	**17pp**	4.1

How to read, and what shows

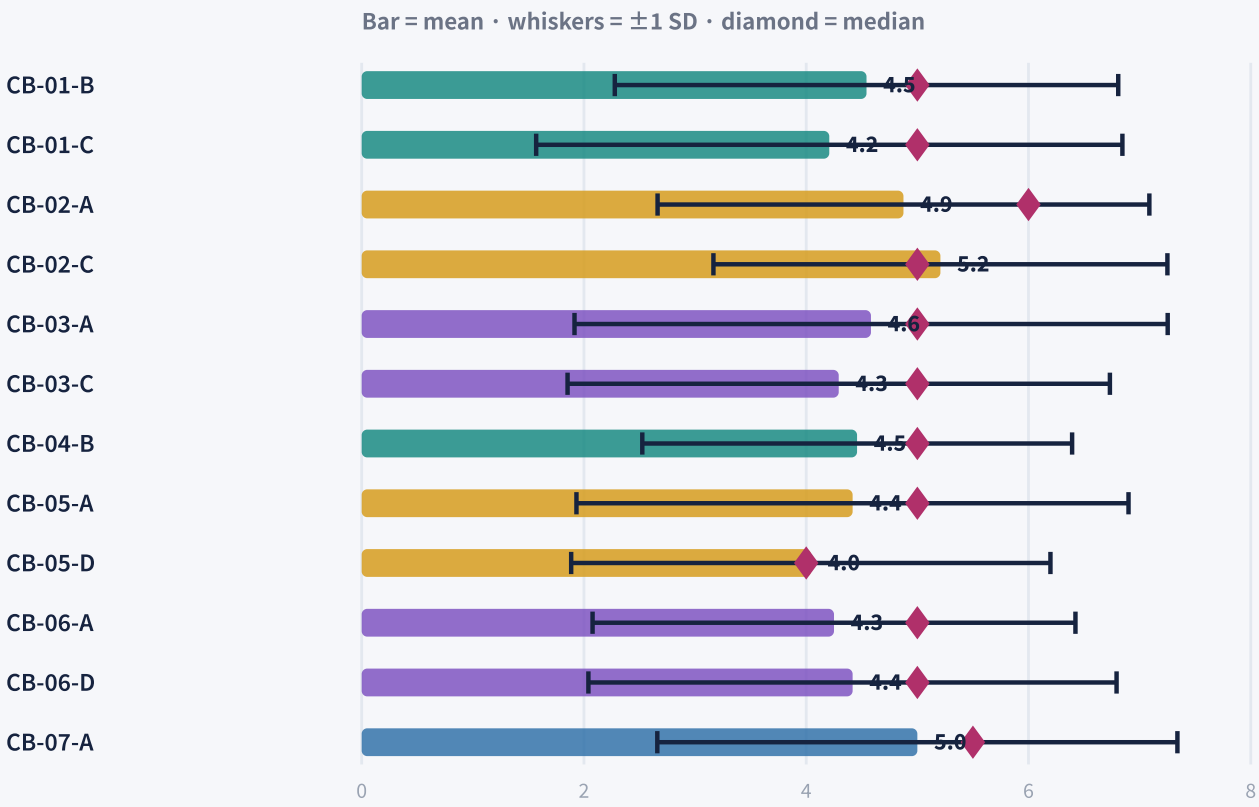
- Dot = share of the group picking the model action (key 5) as Best. Gray = bottom 27% (6), teal = top 27% (6) · groups split by total mean score.
- Wider gaps mark situations where judgment separates most - widest CB-02-C 100pp, narrowest CB-11-C 17pp.
- A low top-group rate marks a situation the whole cohort finds hard - evidence for onboarding priorities.

Calculation basis · the bottom/top rates on this page are the share picking the model action (key 5) as Best, within this cohort's 24 members split by total mean score into the bottom 27% and top 27% (6 each). Situations where the top group reaches 90%+: 2 of 20.

RELATIVE-COMPARISON EVIDENCE · PER-SITUATION MEANS

Situation means span 1.3 points, from 4.0 (CB-05-D) to 5.4 (CB-11-B). Judgment splits most on CB-12-C (SD 2.7).

Score distribution by situation · mean ±SD · median



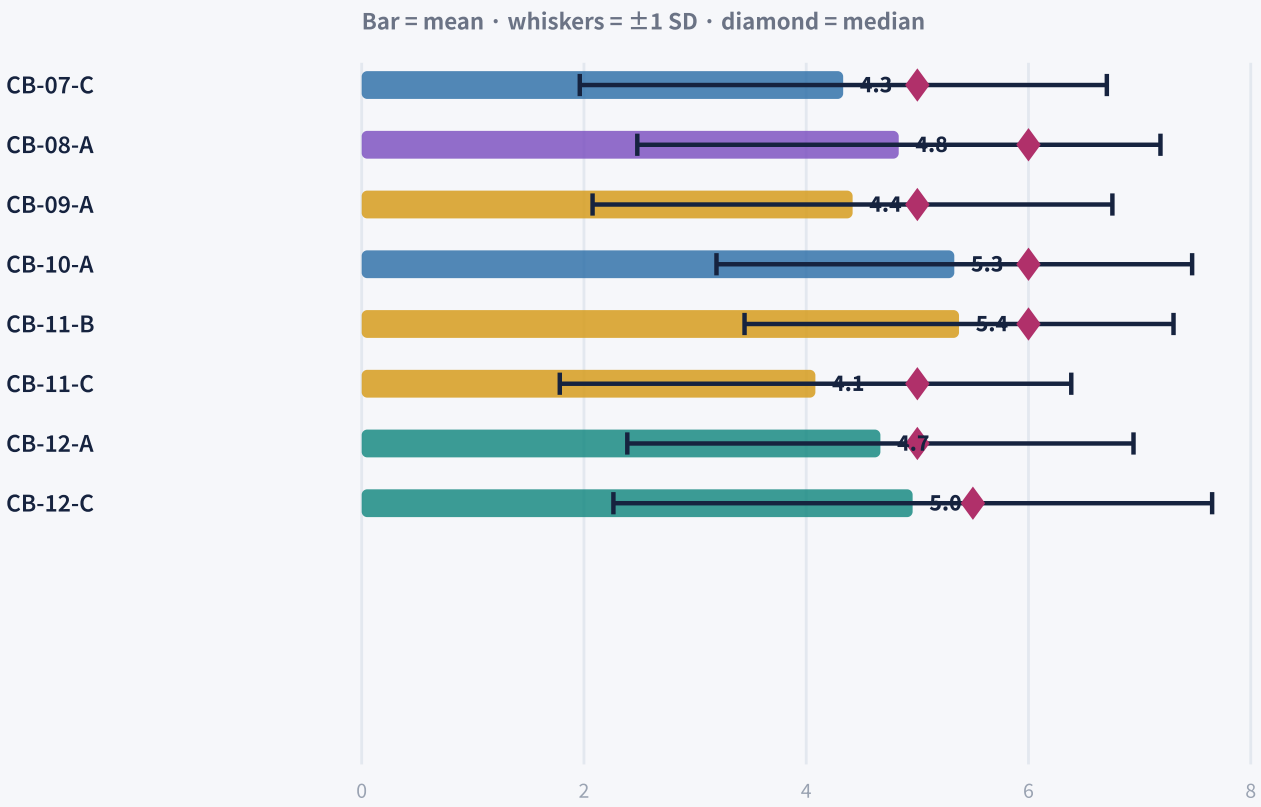
Figures

Situation	Mean	SD	Median	Q1–Q3	Range
CB-01-B	**4.5**	2.3	5.0	3.0–6.0	1–8
CB-01-C	**4.2**	2.6	5.0	2.8–6.3	0–8
CB-02-A	**4.9**	2.2	6.0	3.0–7.0	1–8
CB-02-C	**5.2**	2.0	5.0	3.0–7.0	2–8
CB-03-A	**4.6**	2.7	5.0	2.8–6.3	0–8
CB-03-C	**4.3**	2.4	5.0	2.0–6.0	0–7
CB-04-B	**4.5**	1.9	5.0	3.0–5.3	1–8
CB-05-A	**4.4**	2.5	5.0	2.8–7.0	1–8
CB-05-D	**4.0**	2.2	4.0	2.0–5.3	0–8
CB-06-A	**4.3**	2.2	5.0	2.8–6.0	1–8
CB-06-D	**4.4**	2.4	5.0	2.8–7.0	1–8
CB-07-A	**5.0**	2.3	5.5	5.0–6.3	0–8

Descriptive · pre-calibration

Score distribution continued · situations 13-20.

Score distribution by situation · mean ±SD · median



Figures

Situation	Mean	SD	Median	Q1–Q3	Range
CB-07-C	**4.3**	2.4	5.0	2.0–6.0	1–8
CB-08-A	**4.8**	2.4	6.0	3.0–7.0	1–8
CB-09-A	**4.4**	2.3	5.0	3.0–6.3	0–8
CB-10-A	**5.3**	2.1	6.0	4.5–7.0	1–8
CB-11-B	**5.4**	1.9	6.0	5.0–7.0	1–8
CB-11-C	**4.1**	2.3	5.0	2.5–6.0	0–7
CB-12-A	**4.7**	2.3	5.0	2.8–6.0	0–8
CB-12-C	**5.0**	2.7	5.5	3.0–7.3	0–8

How to read, and what shows

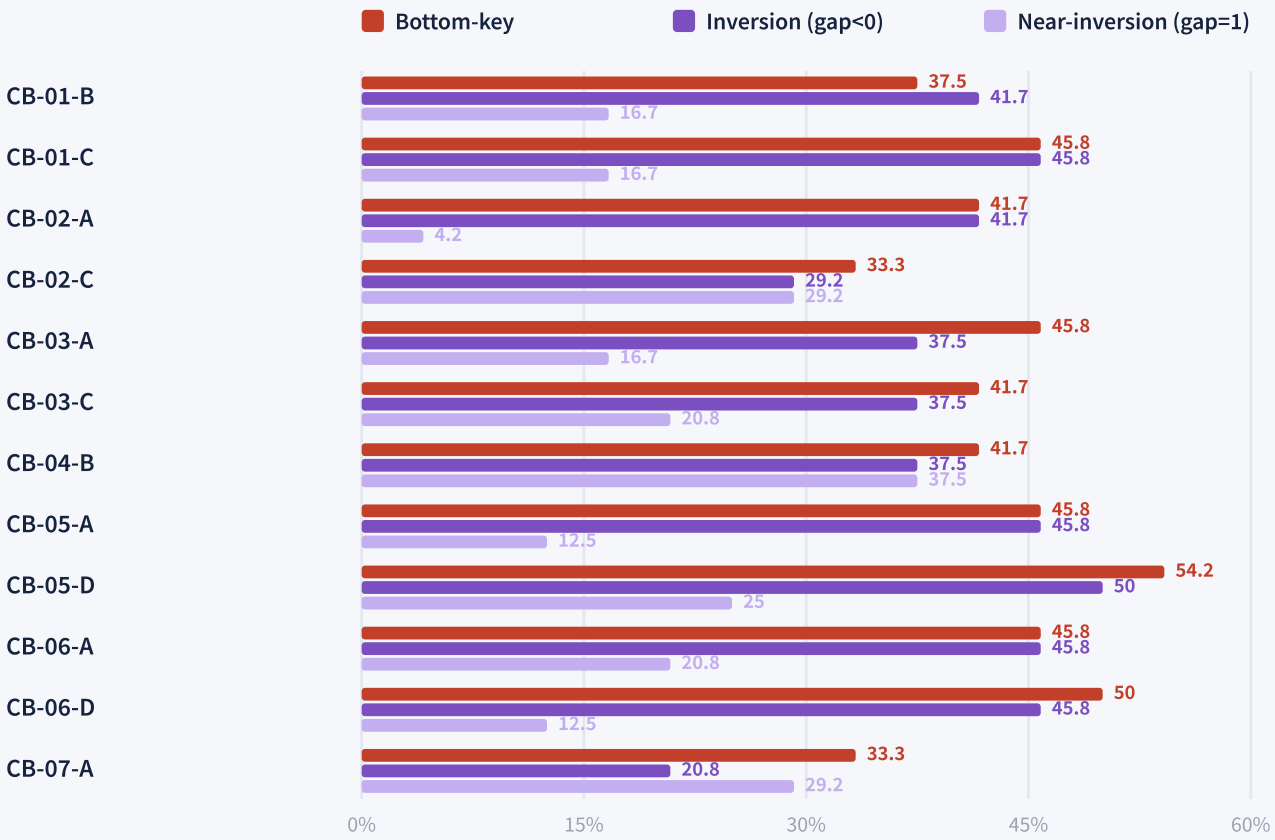
- Score = 0-8 per situation (Best rank + (5 - Worst rank)). Every figure is recomputed from the 24 members' raw item scores.
- Highest mean CB-11-B 5.4/8, lowest CB-05-D 4.0/8. The widest spread (SD) is CB-12-C (2.7) - the situation where this cohort disagrees most.
- Bar colour = competency axis. Low means with long whiskers are cohort-level development priorities.

Calculation basis · SD is the sample standard deviation (n−1) of situation scores (0-8) across this cohort's 24 members. Per-situation SD ranges 1.9-2.7; means range 4.0-5.4. Every figure is recomputed from member-level raw item scores.

RELATIVE-COMPARISON EVIDENCE · BOTTOM-KEY PICKS & INVERSIONS

Bottom-key Best picks exceed 10% in 20 situations, peaking at CB-05-D (54.2%). Inversions concentrate on CB-05-D (50%).

Bottom-key picks · inversions · near-inversions by situation



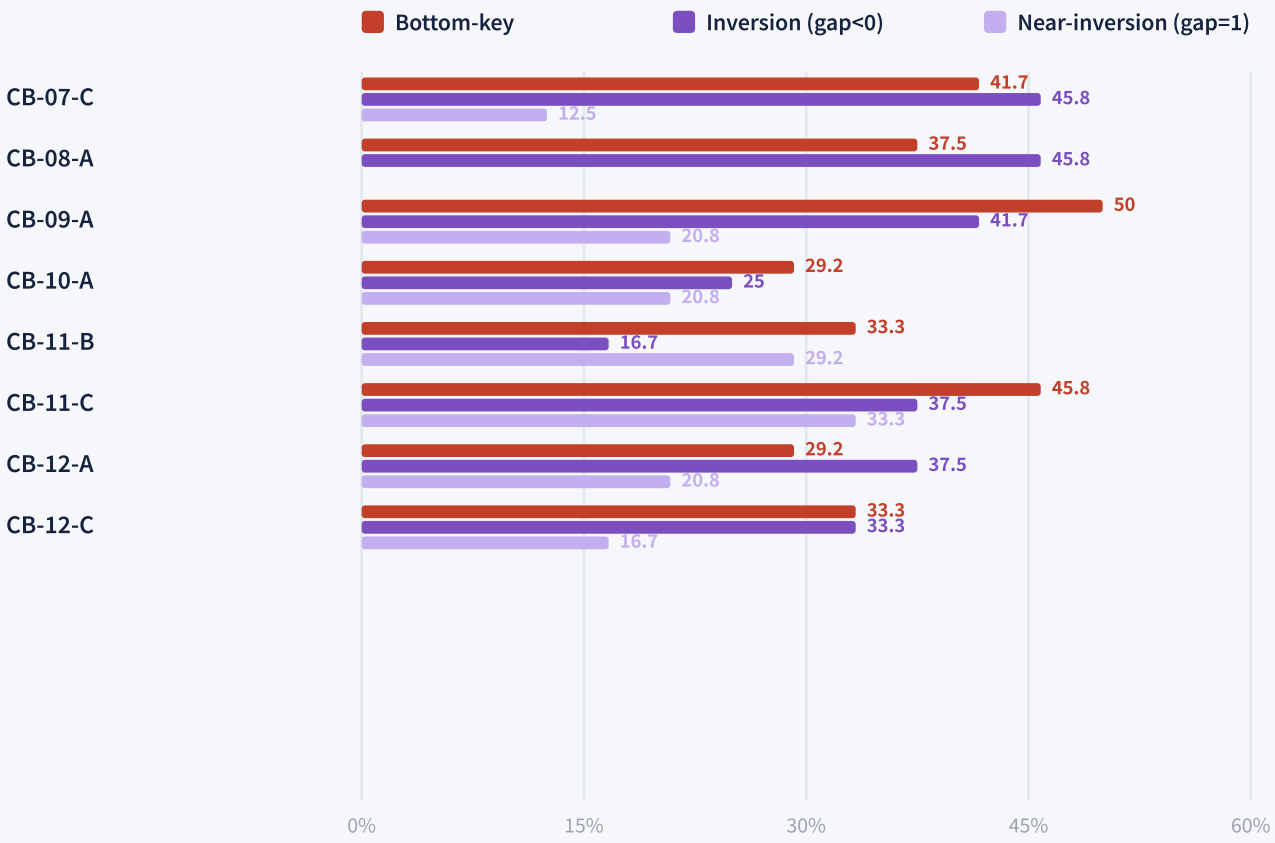
Figures

Situation	Bottom-key	Inversion	Near-inversion	Model Best
CB-01-B	37.5%	41.7%	16.7%	29.2%
CB-01-C	45.8%	45.8%	16.7%	29.2%
CB-02-A	41.7%	41.7%	4.2%	29.2%
CB-02-C	33.3%	29.2%	29.2%	37.5%
CB-03-A	45.8%	37.5%	16.7%	29.2%
CB-03-C	41.7%	37.5%	20.8%	4.2%
CB-04-B	41.7%	37.5%	37.5%	25%
CB-05-A	45.8%	45.8%	12.5%	25%
CB-05-D	54.2%	50%	25%	20.8%
CB-06-A	45.8%	45.8%	20.8%	12.5%
CB-06-D	50%	45.8%	12.5%	25%
CB-07-A	33.3%	20.8%	29.2%	25%

Descriptive · pre-calibration

Bottom-key picks continued · situations 13-20.

Bottom-key picks · inversions · near-inversions by situation



Figures

Situation	Bottom-key	Inversion	Near-inversion	Model Best
CB-07-C	41.7%	45.8%	12.5%	25%
CB-08-A	37.5%	45.8%	0%	33.3%
CB-09-A	50%	41.7%	20.8%	20.8%
CB-10-A	29.2%	25%	20.8%	29.2%
CB-11-B	33.3%	16.7%	29.2%	37.5%
CB-11-C	45.8%	37.5%	33.3%	20.8%
CB-12-A	29.2%	37.5%	20.8%	33.3%
CB-12-C	33.3%	33.3%	16.7%	41.7%

How to read, and what shows

• Bottom-key Best = a Best pick on one of the two lowest-ranked key options (key ranks 1–2). Inversion = the Best pick's key rank is below the Worst pick's (gap<0). Near-inversion = ranks one step apart (gap=1) - a near-flipped judgment.
• Of 480 response pairs: 183 inversions (38.1%), 95 near-inversions (19.8%). 24 members (100.0%) picked at least one bottom-key Best.
• Situations high on both inversion and near-inversion are key/option review candidates; bottom-key-only situations are cohort coaching targets.
• Severity-tagged risky picks (fatal/high) are counted on the dashboard, item health and interview bank pages.

RELATIVE-COMPARISON EVIDENCE · GROUP DISCRIMINATION

12 of 12 items sit above the working bar (r ≥ 0.15). The strongest separation is CB-06-D (D 0.56).

Discrimination by situation · D-index & item-rest r



Figures

Situation	D-index	Item-rest r	α if removed	Flag
CB-01-B	**0.54**	0.70	-0.013	working
CB-01-C	**0.44**	0.44	-0.004	working
CB-02-A	**0.52**	0.71	-0.013	working
CB-02-C	**0.38**	0.42	-0.004	working
CB-03-A	**0.42**	0.35	-0.000	working
CB-03-C	**0.42**	0.45	-0.004	working
CB-04-B	**0.52**	0.70	-0.012	working
CB-05-A	**0.33**	0.38	-0.002	working
CB-05-D	**0.31**	0.42	-0.004	working
CB-06-A	**0.33**	0.46	-0.005	working
CB-06-D	**0.56**	0.61	-0.010	working
CB-07-A	**0.25**	0.37	-0.002	working

How to read, and what shows

- D-index = (top 27% mean - bottom 27% mean) / 8 · item-rest r = correlation of the item with the rest total · working bar r ≥ 0.15 (dashed).
- Overall reliability α = 0.878. An alpha-if-removed above the overall α marks an item whose removal would raise reliability.
- All 12 items sit above the working bar.

Descriptive · pre-calibration

Situations 13–20 · 8 of 8 items sit above the working bar (r ≥ 0.15). The strongest separation is CB-12-C (D 0.60).

Discrimination by situation · D-index & item-rest r



Figures

Situation	D-index	Item-rest r	α if removed	Flag
CB-07-C	**0.25**	0.36	-0.001	working
CB-08-A	**0.35**	0.45	-0.005	working
CB-09-A	**0.44**	0.58	-0.009	working
CB-10-A	**0.46**	0.59	-0.009	working
CB-11-B	**0.33**	0.46	-0.005	working
CB-11-C	**0.25**	0.26	0.002	working
CB-12-A	**0.42**	0.50	-0.006	working
CB-12-C	**0.60**	0.55	-0.008	working

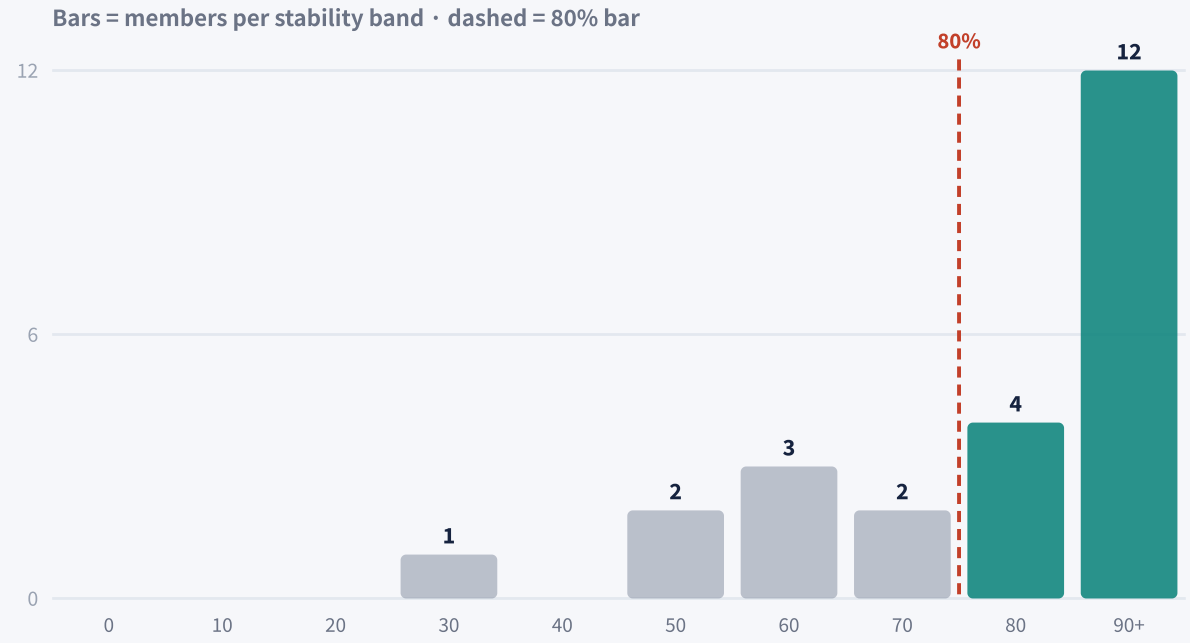
How to read, and what shows

- D-index = (top 27% mean - bottom 27% mean) / 8 · item-rest r = correlation of the item with the rest total · working bar r ≥ 0.15 (dashed).
- Overall reliability α = 0.878. An alpha-if-removed above the overall α marks an item whose removal would raise reliability.
- All 8 items sit above the working bar.

RELATIVE-COMPARISON EVIDENCE · TYPE STABILITY RESAMPLING

16 of 24 members hold their type in 80%+ of 1,000 resamples. Undifferentiated is the least stable type (mean 78.8%).

Type stability distribution · 1,000 resamples



By type

Judgment type	N	Mean stability	≥80%
Borderline	15	**84.0%**	10
Undifferentiated	7	**78.8%**	4
Balanced	2	**98.9%**	2
Effective-leaning	0	**0.0%**	0
Harm-leaning	0	**0.0%**	0

How to read, and what shows

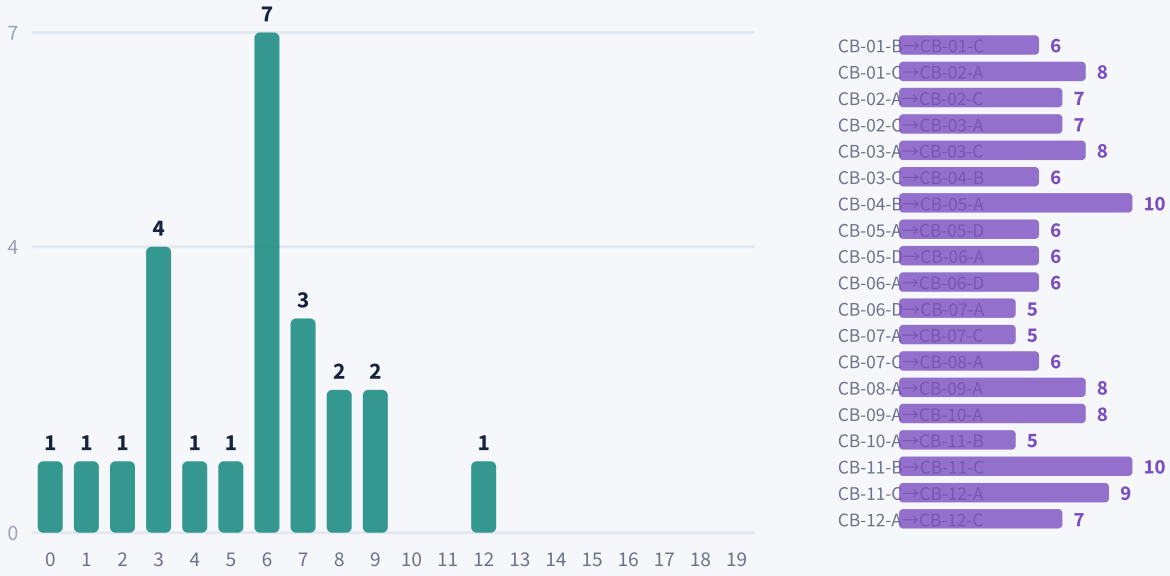
- Stability = share of 1,000 deterministic resamples (seed bw-v6|<member>) reproducing the member's type. Identical data always reproduces the same result.
- At 80%+ the type is firm; below it, a different item mix can move the type. Cohort mean 83.7%, borderline (shown-type uncertainty rule) 15 members.
- A type is not a label on a person but a response pattern on this item set; stability shows how sensitive that pattern is to the item mix.

RELATIVE-COMPARISON EVIDENCE · AXIS ALTERNATION

Members alternate axes 5.5 times on average; 1 never alternate. The busiest boundary is CB-04-B→CB-05-A (10 members).

Axis alternation · by member and boundary

Left = alternations per member · right = boundaries where flips happen



Figures

Alternations	N	Share	Cumulative
0	1	4.2%	4.2%
1	1	4.2%	8.3%
2	1	4.2%	12.5%
3	4	16.7%	29.2%
4	1	4.2%	33.3%
5	1	4.2%	37.5%
6	7	29.2%	66.7%
7	3	12.5%	79.2%
8	2	8.3%	87.5%
9	2	8.3%	95.8%
10	0	0.0%	95.8%
11	0	0.0%	95.8%
12	1	4.2%	100.0%
13	0	0.0%	100.0%
14	0	0.0%	100.0%
15	0	0.0%	100.0%
16	0	0.0%	100.0%
17	0	0.0%	100.0%
18	0	0.0%	100.0%
19	0	0.0%	100.0%

How to read, and what shows

• A situation's axis state = E when effective-identification (Best rank) beats harm-rejection (6 - Worst rank), H when weaker, skipped when tied. One E↔H flip between consecutive situations counts as one alternation.
• Zero alternations: 1 members answered consistently on one axis. More alternations = responses led by different axes across situations - a pattern description, not an integrity verdict.
• Busiest boundary: CB-04-B→CB-05-A, 10 members.

Descriptive · pre-calibration

42.1% of responses sit in the 6-8 band. Only 12.7% land in the lowest band (0-2).

Situation × band matrix

Cell = members in the band (N) · bottom row = band totals

	0-2	2-4	4-6	6-8
CB-01-B	3	7	4	10
CB-01-C	4	7	4	9
CB-02-A	2	8	1	13
CB-02-C	0	7	7	10
CB-03-A	5	4	4	11
CB-03-C	5	4	5	10
CB-04-B	1	8	9	6
CB-05-A	5	6	3	10
CB-05-D	2	10	6	6
CB-06-A	3	8	5	8
CB-06-D	4	7	3	10
CB-07-A	4	1	7	12

Competency bands

Competency bands · bar = 25-75% interval, tick = median (0-8)



Figures

Situation	0-2	2-4	4-6	6-8	Top band	Share
CB-01-B	3	7	4	10	**6-8**	41.7%
CB-01-C	4	7	4	9	**6-8**	37.5%
CB-02-A	2	8	1	13	**6-8**	54.2%
CB-02-C	0	7	7	10	**6-8**	41.7%
CB-03-A	5	4	4	11	**6-8**	45.8%
CB-03-C	5	4	5	10	**6-8**	41.7%
CB-04-B	1	8	9	6	**4-6**	37.5%
CB-05-A	5	6	3	10	**6-8**	41.7%
CB-05-D	2	10	6	6	**2-4**	41.7%
CB-06-A	3	8	5	8	**2-4·6-8**	33.3%
CB-06-D	4	7	3	10	**6-8**	41.7%
CB-07-A	4	1	7	12	**6-8**	50.0%

Descriptive · pre-calibration

Situation × band matrix continued · situations 13-20 and band totals.

Situation × band matrix

Cell = members in the band (N) · bottom row = band totals

	0-2	2-4	4-6	6-8
CB-07-C	3	8	3	10
CB-08-A	1	10	0	13
CB-09-A	4	6	5	9
CB-10-A	1	5	5	13
CB-11-B	2	2	7	13
CB-11-C	6	3	8	7
CB-12-A	1	8	5	10
CB-12-C	5	3	4	12
Band totals	**61**	**122**	**95**	**202**

Figures

Situation	0-2	2-4	4-6	6-8	Top band	Share
CB-07-C	3	8	3	10	**6-8**	41.7%
CB-08-A	1	10	0	13	**6-8**	54.2%
CB-09-A	4	6	5	9	**6-8**	37.5%
CB-10-A	1	5	5	13	**6-8**	54.2%
CB-11-B	2	2	7	13	**6-8**	54.2%
CB-11-C	6	3	8	7	**4-6**	33.3%
CB-12-A	1	8	5	10	**6-8**	41.7%
CB-12-C	5	3	4	12	**6-8**	50.0%

How to read, and what shows

- A band is a 2-point slice of the 0-8 item score. A cell counts the members landing in that band on that situation.
- Of all 288 pairs, 12.7% sit in the 0-2 band and 42.1% in 6-8. A situation whose mass sits in one band separates members less.
- Competency bands are the 25-75% intervals of per-member competency means. Wider interval = more spread in the cohort.

RELATIVE-COMPARISON EVIDENCE · NINE-GRAPH SYNTHESIS

Across 20 situations, mean D is 0.41; separation spans from CB-02-C at 100pp down to CB-03-C at 17pp.

Nine pieces of evidence at a glance

1. Difficulty gap (top - bottom 27%)
100pp / 17pp
CB-02-C / CB-03-C



2. Per-situation mean range
4.0 – 5.4
CB-05-D – CB-11-B



3. Inverted picks (of all pairs)
38.1%
gap < 0 of 480 pairs



4. Discrimination D
mean 0.41
lowest CB-11-C 0.25



5. Type stability ≥80%
16 · 67%
borderline 15 least stable



6. Axis alternations (mean)
5.5
CB-04-B→CB-05-A 10



7. 6-8 band share
42.1%
0-2 band 12.7%



8. Reliability (α · split-half)
0.88 · 0.88
20/20 items at or above r = 0.15



9. Members with a harmful pick
100.0%
24 of 24



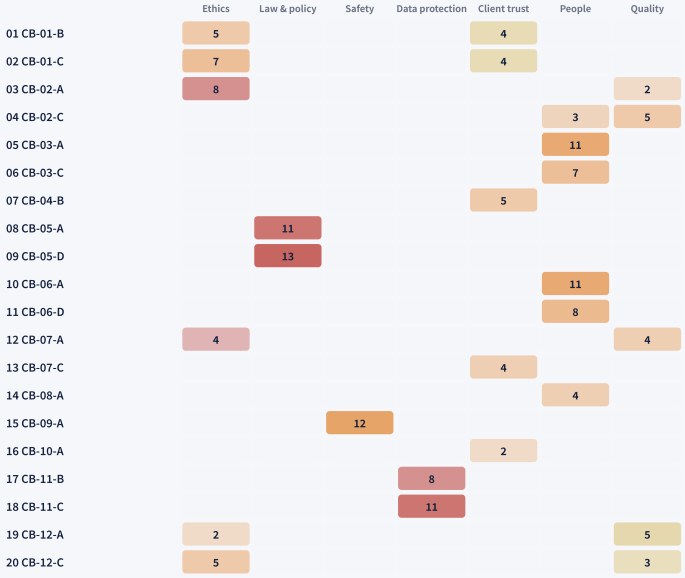
Synthesis in three sentences

① Across 20 situations, mean D is 0.41; 42% of responses sit in the 6-8 band. ② Separation varies widely by situation: CB-02-C shows a 100pp top-bottom gap while CB-03-C manages only 17pp, leaving it nearly flat. ③ Types are stable for 67%, but the 15 borderline members can shift with the item mix, and the 100% harmful-pick rate with 38.1% inversions marks interview territory.

RISK CONCENTRATION

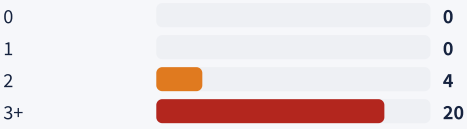
Harmful picks are not spread evenly. They pile up most in CB-05-D and CB-06-D.

People picking a harmful option as Best · scenario × risk area



Number = people · color = highest severity in the cell (red Critical · orange High · yellow Caution, severity tags) · darker = more · blank = no option in that area · 6 key-rank-2 options without a panel tag are unclassified - row totals: 06 CB-03-C 3 · 07 CB-04-B 5 · 11 CB-06-D 4 · 13 CB-07-C 6 · 14 CB-08-A 5 · 16 CB-10-A 5 · these people are not counted in any cell above

Risky picks per person



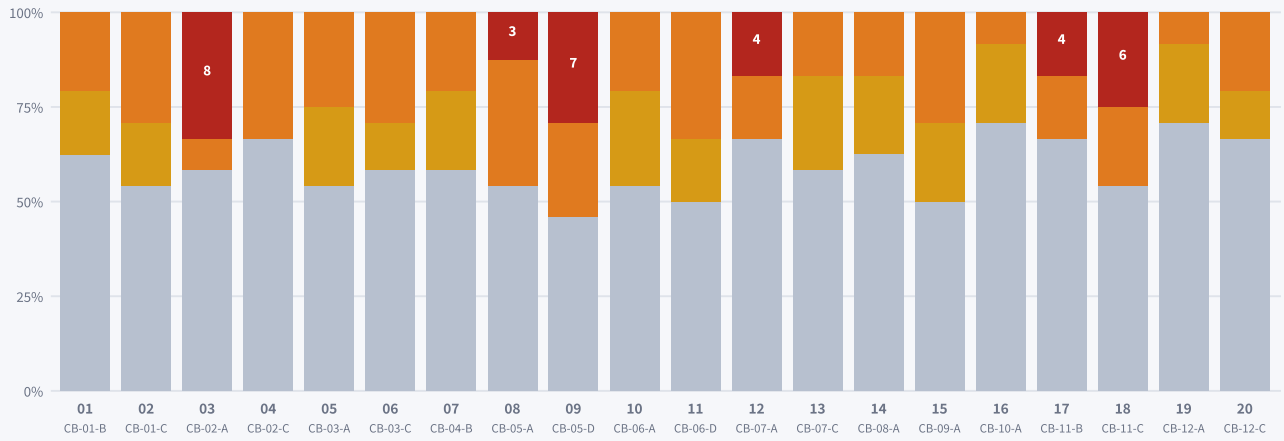
Repeated pattern (same area 2+) 13 · Most-harmful option (key 1) at least once 20

Most attractive harmful options · top 6

③ CB-02-A	Leave the original numbers in place…	key 1	Critical	33%
⑧ CB-05-A	Assume approval will come and send the…	key 2	High	33%
⑪ CB-06-D	Leave them out of important work because…	key 1	High	33%
② CB-01-C	Have staff post rebuttal comments from…	key 1	High	29%
⑥ CB-03-C	Take one side and propose removing the…	key 1	High	29%
⑨ CB-05-D	Take the competitor's price list and use…	key 1	Critical	29%

Options shown as summaries only · usable as priorities for training and onboarding content.

Harm level of Best picks by situation

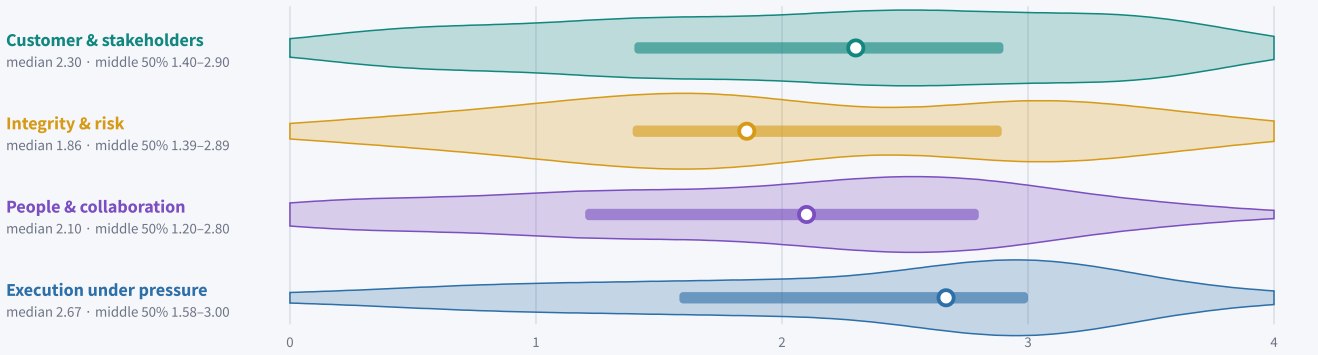


Bar = mix of harm severity among options picked as Best in that situation (%, severity tags). Situations with large orange/red parts come first for interviews and training.

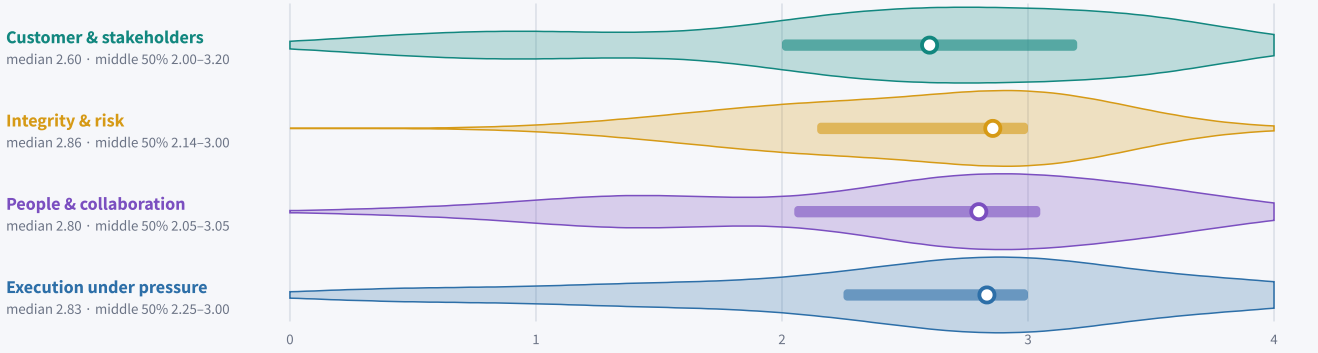
COMPETENCY PROFILE · COHORT

This cohort is most solid in ‘Execution under pressure’ and most spread out in ‘Customer & stakeholders’. Use it to set the onboarding focus.

Best credit by competency



Worst credit by competency



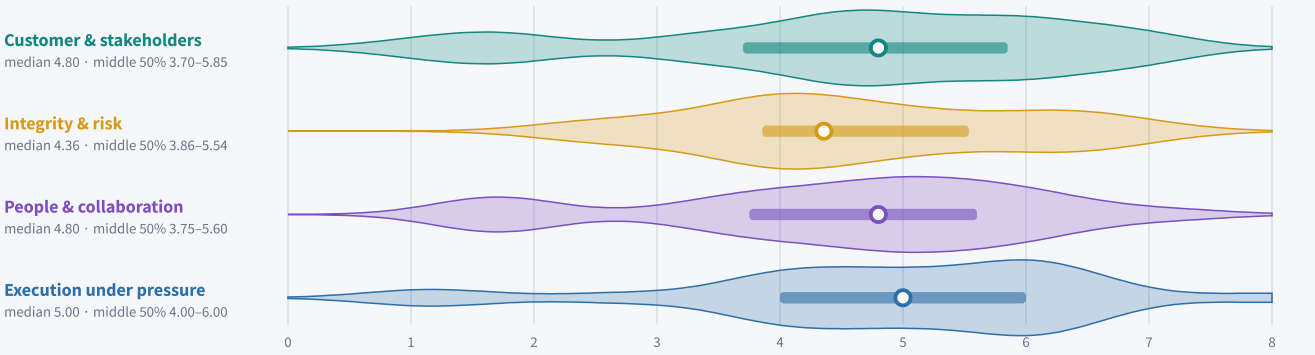
How to read

Shape = spread, dark bar = middle 50%, white dot = median. With 2–4 situations per competency, use this for group trends rather than individual reads.

DEVELOPMENT · ONBOARDING PLAN

Starting onboarding with ‘Customer & stakeholders’ helps the most people. Each suggestion is tied to a situation this cohort actually missed.

Cohort spread by competency · Best + Worst credit



Integrity & risk · median 4.4/8

Evidence · most-missed situation CB-05-D (mean 4.0/8 · far-miss 38% · inverted 50%)
Who · 16 people (67%) at 5 or below in this competency
Week 1 · CB-05-D case workshop: model action “Politely decline, say it cannot be used, and tell…” vs the most common wrong pick
Month 1 · review one similar real decision with a mentor
Month 3 · short re-check on situations in the same competency (new item version, same scoring)

Customer & stakeholders · median 4.8/8

Evidence · most-missed situation CB-01-C (mean 4.2/8 · far-miss 25% · inverted 46%)
Who · 14 people (58%) at 5 or below in this competency
Week 1 · CB-01-C case workshop: model action “Contact the customer to check the issue, then post a…” vs the most common wrong pick
Month 1 · review one similar real decision with a mentor
Month 3 · short re-check on situations in the same competency (new item version, same scoring)

DEVELOPMENT · ONBOARDING PLAN

Starting onboarding with ‘Customer & stakeholders’ helps the most people. Each suggestion is tied to a situation this cohort actually missed. · continued

People & collaboration · median 4.8/8

Evidence · most-missed situation CB-06-A (mean 4.3/8 · far-miss 42% · inverted 46%)
Who · 15 people (63%) at 5 or below in this competency
Week 1 · CB-06-A case workshop: model action “Fix today's file together, then find the cause and…” vs the most common wrong pick
Month 1 · review one similar real decision with a mentor
Month 3 · short re-check on situations in the same competency (new item version, same scoring)

Execution under pressure · median 5.0/8

Evidence · most-missed situation CB-07-C (mean 4.3/8 · far-miss 33% · inverted 46%)
Who · 13 people (54%) at 5 or below in this competency
Week 1 · CB-07-C case workshop: model action “Line up backup venues fast and, once set, tell…” vs the most common wrong pick
Month 1 · review one similar real decision with a mentor
Month 3 · short re-check on situations in the same competency (new item version, same scoring)

Basis and limits

- Each card comes from the situation this cohort missed most, not a generic training list.
- These are onboarding suggestions, not a validated program.

Onboarding evidence record

Checkpoint	Work evidence to record
Week 1	A case discussion, options considered and the reason for the chosen action.
Month 1	One similar real decision and a mentor’s observation.
Month 3	A same-competency check with the form version and scoring basis recorded.

This is a suggested record, not a measured training effect. Track an observed change separately from the original assessment result.

INTERVIEW QUESTION BANK · BUILT FROM THE COHORT'S RISKY PICKS

Structured questions for the 5 scenarios with the most risky picks, and who each question is for.

⑨ CB-05-D · 13 people · Critical 7

“In ‘CB-05-D’ — A colleague who joined from a competitor offers to show you the… What would you do?”

Follow-up: What if someone proposes to ‘Take the competitor's price list and use it…’?

Strong answer	Concern	IDs (ID order)
Reaches the principle of the model action (‘Politely decline, say it cannot be used, and tell…’)	Defends ‘Take the competitor's price list and use it…’	Sample 002, Sample 005, Sample 006, Sample 010 +9 more

⑪ CB-06-D · 12 people · Critical 8

“In ‘CB-06-D’ — A remote new hire has barely interacted with the team for a… What would you do?”

Follow-up: What if someone proposes to ‘Leave them out of important work because they…’?

Strong answer	Concern	IDs (ID order)
Reaches the principle of the model action (‘Ask one-on-one how they are settling in and set up…’)	Defends ‘Leave them out of important work because they…’	Sample 002, Sample 004, Sample 006, Sample 007 +8 more

⑮ CB-09-A · 12 people · Critical 7

“In ‘CB-09-A’ — During a site check you see a minor but repeated safety rule… What would you do?”

Follow-up: What if someone proposes to ‘The schedule is tight, so bring it up at the…’?

Strong answer	Concern	IDs (ID order)
Reaches the principle of the model action (‘Stop the work, check the cause, act on it, and log…’)	Defends ‘The schedule is tight, so bring it up at the…’	Sample 002, Sample 004, Sample 005, Sample 006 +8 more

② CB-01-C · 11 people · Critical 7

“In ‘CB-01-C’ — A customer post claiming a product defect is spreading on social… What would you do?”

Follow-up: What if someone proposes to ‘Have staff post rebuttal comments from…’?

Strong answer	Concern	IDs (ID order)
Reaches the principle of the model action (‘Contact the customer to check the issue, then post…’)	Defends ‘Have staff post rebuttal comments from…’	Sample 006, Sample 007, Sample 009, Sample 010 +7 more

⑤ CB-03-A · 11 people · Critical 6

“In ‘CB-03-A’ — In a meeting, a colleague sarcastically put down your proposal in… What would you do?”

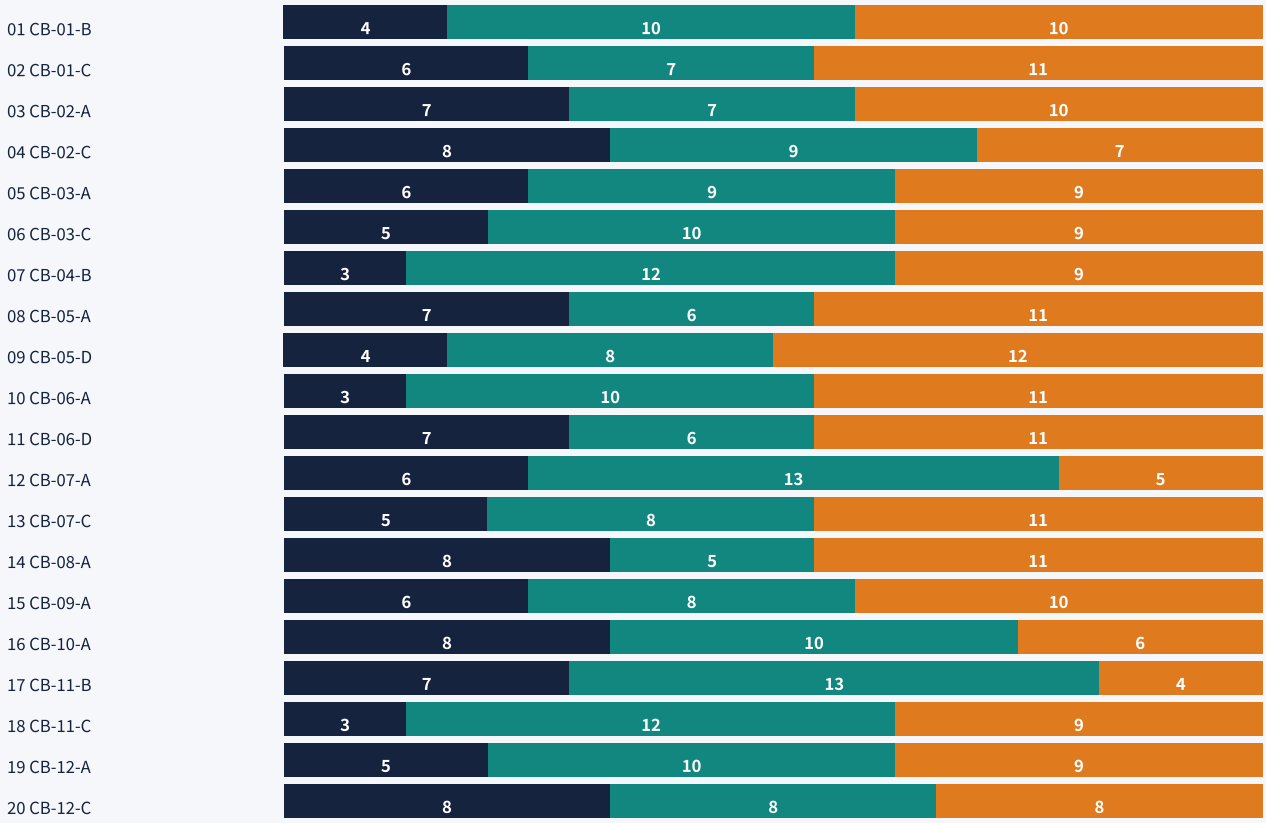
Follow-up: What if someone proposes to ‘Quietly spread complaints about that…’?

Strong answer	Concern	IDs (ID order)
Reaches the principle of the model action (‘Raise the moment one-on-one after the meeting and…’)	Defends ‘Quietly spread complaints about that…’	Sample 001, Sample 002, Sample 004, Sample 006 +7 more

Top 5 scenarios by risky picks (fatal/high) · generated the same way as page 13 (Interview · development guide) of each individual report · questions for all 20 scenarios are in the HTML version and audit appendix · the interviewer makes the final choice.

Item score bands by situation · wide orange bands are the interview priorities.

Item score bands by situation · 24 people



Navy 7–8 · teal 5–6 · orange 0–3 (item score 0–8; 4 cannot occur under S1). Situations with a wide orange band are where the interview questions go first.

Comparison basis

Each row includes all 24 people once. The three segments are mutually exclusive score bands. Segment labels are counts, not percentages. Use the band pattern to choose a follow-up situation, then check the individual's actual chosen action.

INTERVIEWER GUIDE IN DEPTH

The 4 situations with the most risky picks, and the interview focus for each type: who to ask what, across this section.

Interview focus by type

Type · people	Interview focus	Sign of a strong answer
2 **Balanced**	Confirm strengths · reasoning behind the hardest decision	Can they explain with a real case why their judgment holds
7 **Insufficient evidence on both axes**	Both axes · start by checking how they read the situation	Do they identify the key stakeholders
15 **Borderline**	Interval touches the cut · focus on the risky-pick situations	Does the judgment change when they re-explain the situation

Members by shown type

Type counts use the same scorer output as the dashboard, roster and report headline. All 24 members are counted once. Borderline is an uncertainty label, not a low score.

Balanced · 2 people

Sample 001, Sample 017

Undifferentiated · 7 people

Sample 002, Sample 005, Sample 008, Sample 010, Sample 012, Sample 018, Sample 024

Borderline · 15 people

Sample 003, Sample 004, Sample 006, Sample 007, Sample 009, Sample 011, Sample 013, Sample 014, Sample 015, Sample 016, Sample 019, Sample 020, Sample 021, Sample 022, Sample 023

INTERVIEWER GUIDE IN DEPTH

The 4 situations with the most risky picks, and the interview focus for each type: who to ask what, across this section. · continued

09 CB-05-D · 13 people

Ask · “In a situation like CB-05-D, what did you do first, and why?”
Strong · in the direction of the model action: “Politely decline, say it cannot be…”
Weak · defends the most common risky pick: “Take the competitor's price list and…”
Who · Sample 002, Sample 005, Sample 006, Sample 010, Sample 012, Sample 013 …

11 CB-06-D · 12 people

Ask · “In a situation like CB-06-D, what did you do first, and why?”
Strong · in the direction of the model action: “Ask one-on-one how they are settling in…”
Weak · defends the most common risky pick: “Leave them out of important work…”
Who · Sample 002, Sample 004, Sample 006, Sample 007, Sample 010, Sample 012 …

15 CB-09-A · 12 people

Ask · “In a situation like CB-09-A, what did you do first, and why?”
Strong · in the direction of the model action: “Stop the work, check the cause, act on…”
Weak · defends the most common risky pick: “The schedule is tight, so bring it up…”
Who · Sample 002, Sample 004, Sample 005, Sample 006, Sample 007, Sample 008 …

02 CB-01-C · 11 people

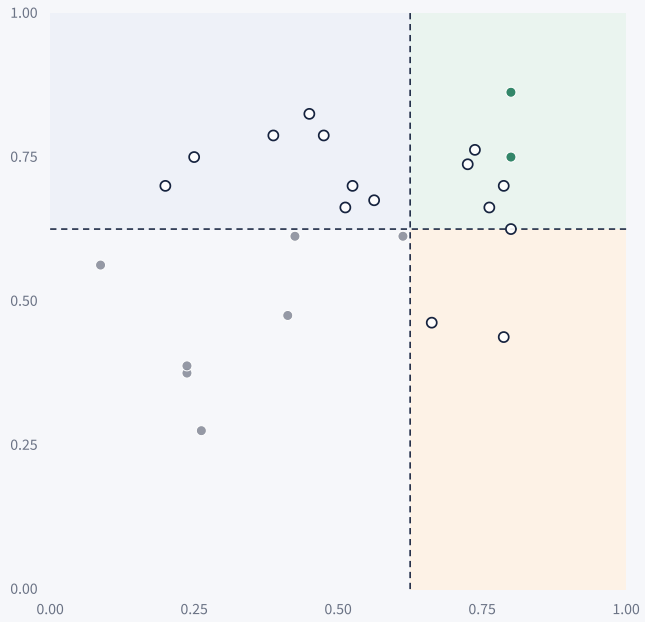
Ask · “In a situation like CB-01-C, what did you do first, and why?”
Strong · in the direction of the model action: “Contact the customer to check the…”
Weak · defends the most common risky pick: “Have staff post rebuttal comments from…”
Who · Sample 006, Sample 007, Sample 009, Sample 010, Sample 012, Sample 016 …

How to use this

• Use with the deep-question page in each individual report. This page organises who to ask what, across the cohort.
• Record answers as ticks and one sentence; do not create a new score.

JUDGMENT LANDSCAPE · INTERVAL STRUCTURE

15 people are shown as Borderline under the scoring rule; their type should be treated as uncertain. About 24 items are needed for 0.90 reliability (Spearman-Brown).



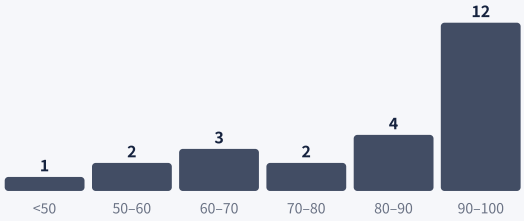
Color = type · hollow = borderline (shown-type uncertainty rule) · one pseudonymous person per dot · no ranking

Type distribution



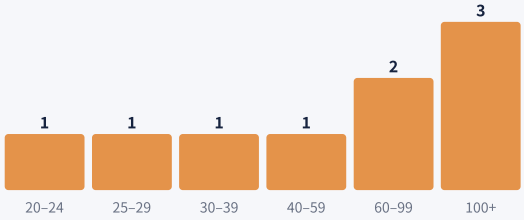
- Balanced 2
- Strong on effective action, risk identification to develop 0
- Strong on risk identification, effective action to develop 0
- Insufficient evidence on both axes 7
- Borderline · no type 15

Type stability (1,000 resamples)



People by % of resamples staying in the same zone · 80%+: 16

Borderline · items needed to set a type (estimate)



Items needed ≈ 20 × (interval half-width ÷ distance to cut)² · median 62 items · finite estimates: 9/15; remaining members have no finite estimate · use additional evidence, not a guaranteed type resolution

Balanced

2

mean effective-action 0.80 · mean harm-ID 0.81

Insufficient evidence on both axes

7

mean effective-action 0.33 · mean harm-ID 0.47

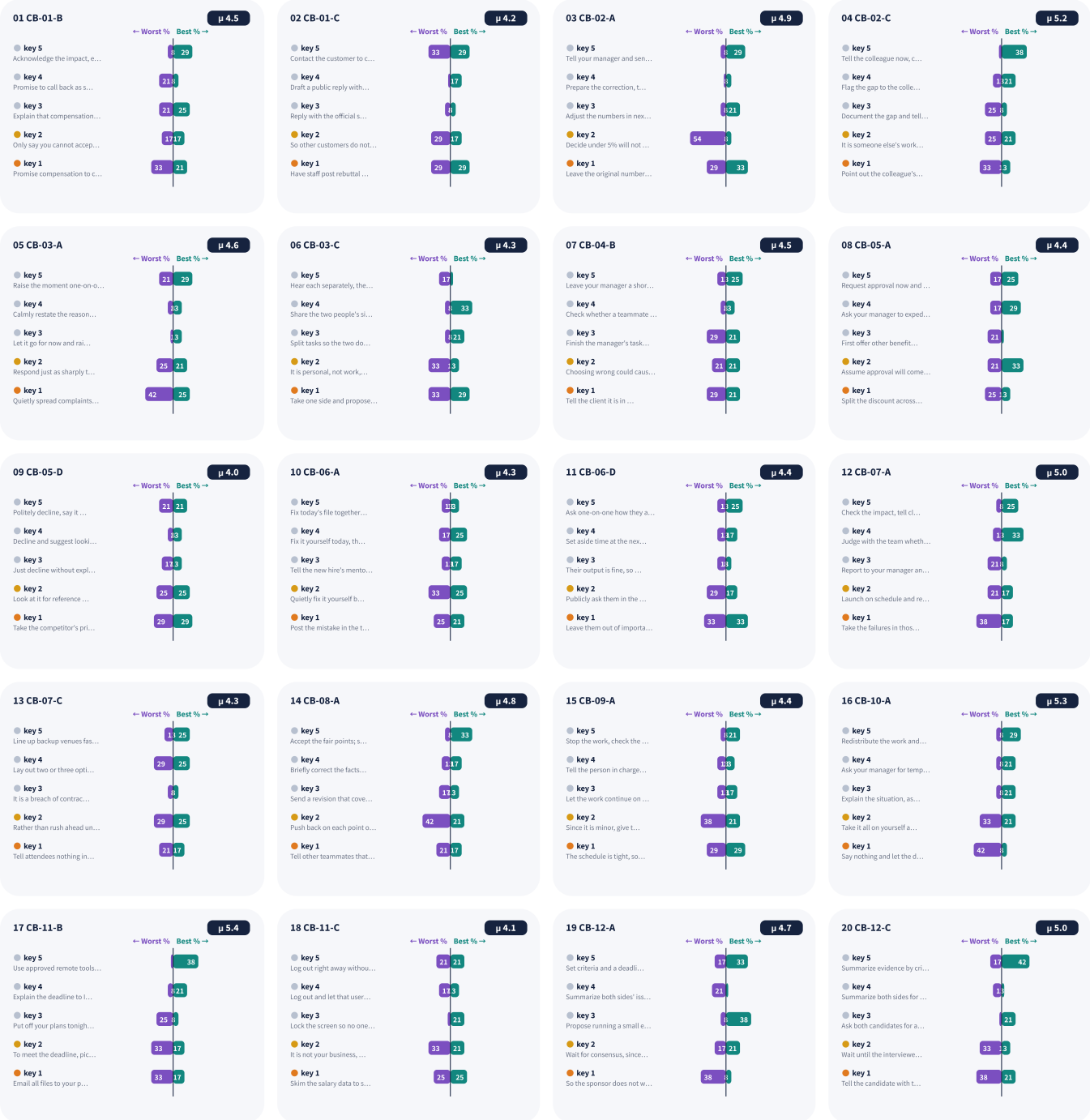
Borderline

15

mean effective-action 0.57 · mean harm-ID 0.69

CHOICE DISTRIBUTIONS · 20 SITUATIONS × 5 OPTIONS

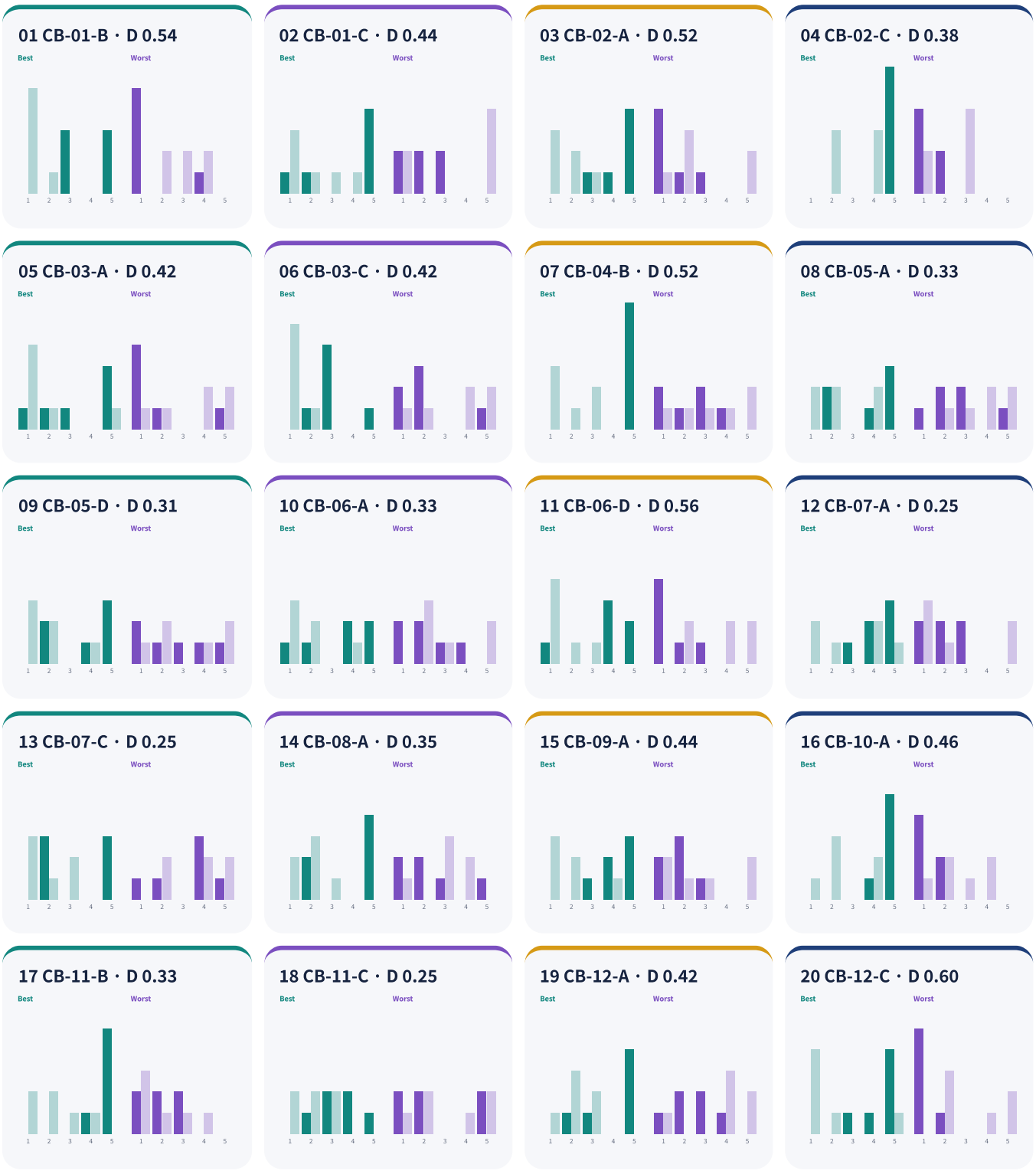
In most situations the model action is the top Best pick and the most harmful action the top Worst pick. The exceptions are covered on the item-health page.



Options in key order (key 5 = model … key 1 = most harmful), with the share of 24 people picking each as Best (teal) and Worst (purple). μ = mean item score. When another option's Best bar is longer than the model action's, the key needs review.

DISTRACTOR ANALYSIS · TOP 27% VS BOTTOM 27%

In 15 of 20 situations the top group picked the model action as Best and the most harmful action as Worst more often. The rest show up in the distractor analysis.



How to read, and what shows

- Bars = pick share by key 1…5 · solid = top 27% (6 people) · light = bottom 27% (6).
- A working item: for Best, the top group is higher at key 5; for Worst, the top group is higher at key 1.
- Middle options picked more by the bottom group (attractive distractors): 01 · key 2, 02 · key 3, 02 · key 4, 03 · key 2, 04 · key 2, 04 · key 4, 07 · key 2, 07 · key 3, 08 · key 4, 10 · key 2, 11 · key 2, 11 · key 3, 12 · key 2, 13 · key 3, 14 · key 2, 14 · key 3 +9

ITEM HEALTH · KEY REVIEW

One line per item: is it working, and what to do next. This cohort makes the next round of the test better.

Scenario	Best=key 5	Worst=key 1	Discrimination r [95% CI]	Top non-model Best	Harmful Best	Flags	Recommended action
01 CB-01-B	29%	33%	0.70 [0.41, 0.86]	key 3 · 25%	38%	Competing option, Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
02 CB-01-C	29%	29%	0.44 [0.04, 0.72]	key 1 · 29%	46%	Competing option, Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
03 CB-02-A	29%	29%	0.71 [0.43, 0.87]	key 1 · 33%	42%	Competing option, Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
04 CB-02-C	38%	33%	0.42 [0.02, 0.70]	key 4 · 21%	33%	Risk hotspot	Keep · auto-suggest in interview bank
05 CB-03-A	29%	42%	0.35 [-0.06, 0.66]	key 1 · 25%	46%	Competing option, Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
06 CB-03-C	4%	33%	0.45 [0.05, 0.72]	key 4 · 33%	42%	Competing option, Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
07 CB-04-B	25%	29%	0.70 [0.41, 0.86]	key 2 · 21%	42%	Risk hotspot	Keep · auto-suggest in interview bank
08 CB-05-A	25%	25%	0.38 [-0.03, 0.68]	key 2 · 33%	46%	Competing option, Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
09 CB-05-D	21%	29%	0.42 [0.02, 0.70]	key 1 · 29%	54%	Competing option, Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
10 CB-06-A	13%	25%	0.46 [0.07, 0.73]	key 2 · 25%	46%	Competing option, Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording

Rules: low discrimination r < 0.15 · competing option = a non-model option taken as Best by 25%+ · risk hotspot = harmful options taken as Best by 20%+ Key changes apply only as a new key_version.

Item-health table continued · scenarios 11-20.

Scenario	Best=key 5	Worst=key 1	Discrimination r [95% CI]	Top non-model Best	Harmful Best	Flags	Recommended action
11 CB-06-D	25%	33%	0.61 [0.27, 0.81]	key 1 · 33%	50%	Competing option Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
12 CB-07-A	25%	38%	0.37 [-0.04, 0.67]	key 4 · 33%	33%	Competing option Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
13 CB-07-C	25%	21%	0.36 [-0.05, 0.66]	key 2 · 25%	42%	Competing option Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
14 CB-08-A	33%	21%	0.45 [0.05, 0.72]	key 2 · 21%	38%	Risk hotspot	Keep · auto-suggest in interview bank
15 CB-09-A	21%	29%	0.58 [0.23, 0.79]	key 1 · 29%	50%	Competing option Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
16 CB-10-A	29%	42%	0.59 [0.24, 0.80]	key 2 · 21%	29%	Risk hotspot	Keep · auto-suggest in interview bank
17 CB-11-B	38%	33%	0.46 [0.07, 0.73]	key 4 · 21%	33%	Risk hotspot	Keep · auto-suggest in interview bank
18 CB-11-C	21%	25%	0.26 [-0.16, 0.60]	key 1 · 25%	46%	Competing option Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
19 CB-12-A	33%	38%	0.50 [0.12, 0.75]	key 3 · 38%	29%	Competing option Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
20 CB-12-C	42%	38%	0.55 [0.18, 0.78]	key 1 · 21%	33%	Risk hotspot	Keep · auto-suggest in interview bank

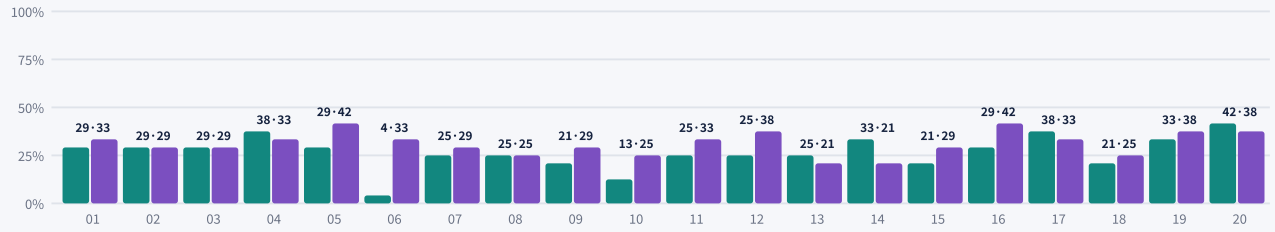
Continued from the previous page (scenarios 11-20).

Descriptive · pre-calibration

Summary metrics and the per-situation key agreement chart.

Mean discrimination	Agreement with key (Best)	Next-round actions
0.49	**27%**	**0**
mean corrected item-total r, 20 items	share of all responses with key 5 as Best	items to review

Agreement with the key by situation · model action as Best / most harmful as Worst



Teal = share picking key 5 (model action) as Best, purple = share picking key 1 (most harmful) as Worst. A situation low on both is a key-review candidate.

Reading key agreement

Each bar uses the same 24-person denominator. Best key 5 is agreement with the model action; Worst key 1 is agreement with the lowest-ranked action. Low agreement can invite a wording or scoring-key review. It is not proof of a defective item or of an individual skill gap. Item-rest correlation and its interval belong beside this chart.

ITEM DISCRIMINATION STORY

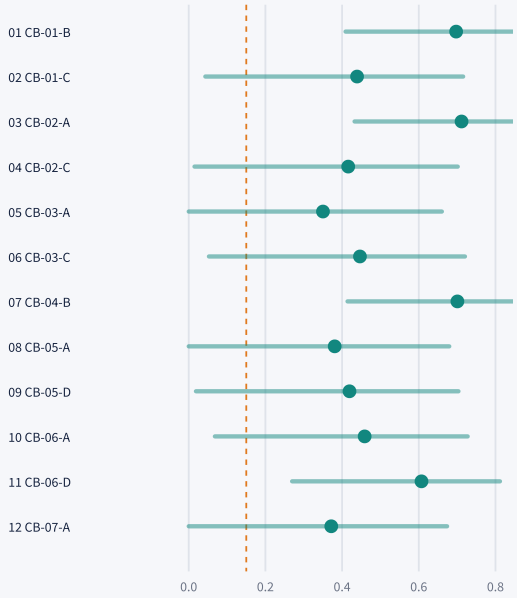
12 of 12 items clearly separate the top and bottom groups.

Ease × discrimination · circle size = item-rest correlation



Orange = D < 0.25. Vertical = (top 27% mean − bottom 27% mean) / 8.

Item-rest correlation · 95% interval



Dashed = 0.15 review line.

Item stories · why each item works, or does not

Situation	p	D	Item-rest r [95%]	Entropy B / W	Reading
01 CB-01-B	0.57	0.54	0.70 [0.41, 0.86]	0.96 / 0.95	works, but picks are spread → review key wording
02 CB-01-C	0.53	0.44	0.44 [0.04, 0.72]	0.95 / 0.80	works, but picks are spread → review key wording
03 CB-02-A	0.61	0.52	0.71 [0.43, 0.87]	0.91 / 0.69	works, but picks are spread → review key wording
04 CB-02-C	0.65	0.38	0.42 [0.02, 0.70]	0.92 / 0.90	works, but picks are spread → review key wording
05 CB-03-A	0.57	0.42	0.35 [-0.06, 0.66]	0.96 / 0.86	works, but picks are spread → review key wording
06 CB-03-C	0.54	0.42	0.45 [0.05, 0.72]	0.90 / 0.90	works, but picks are spread → review key wording
07 CB-04-B	0.56	0.52	0.70 [0.41, 0.86]	0.99 / 0.94	works, but picks are spread → review key wording
08 CB-05-A	0.55	0.33	0.38 [-0.03, 0.68]	0.83 / 0.99	works, but picks are spread → review key wording
09 CB-05-D	0.51	0.31	0.42 [0.02, 0.70]	0.96 / 0.96	works, but picks are spread → review key wording
10 CB-06-A	0.53	0.33	0.46 [0.07, 0.73]	0.98 / 0.95	works, but picks are spread → review key wording
11 CB-06-D	0.55	0.56	0.61 [0.27, 0.81]	0.94 / 0.94	works, but picks are spread → review key wording
12 CB-07-A	0.63	0.25	0.37 [-0.04, 0.67]	0.94 / 0.92	works, but picks are spread → review key wording

Readings are rule-based (D < 0.25 → weak, entropy ≥ 0.85 → spread, p < 0.76 → hard).

ITEM DISCRIMINATION STORY

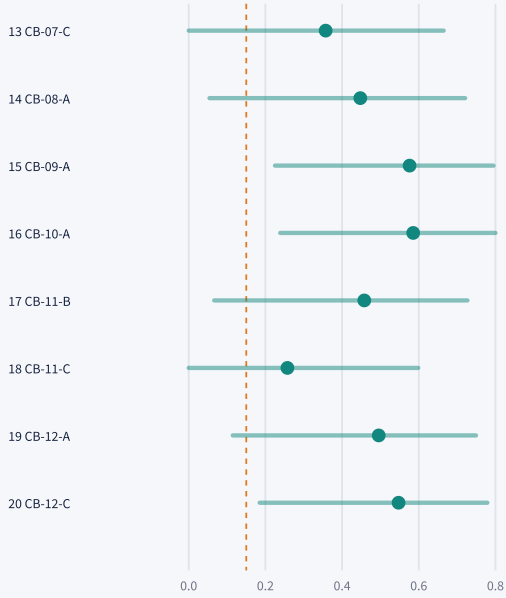
Situations 13–20 · 8 of 8 items clearly separate the top and bottom groups.

Ease × discrimination · circle size = item-rest correlation



Orange = D < 0.25. Vertical = (top 27% mean − bottom 27% mean) / 8.

Item-rest correlation · 95% interval



Dashed = 0.15 review line.

Item stories · why each item works, or does not

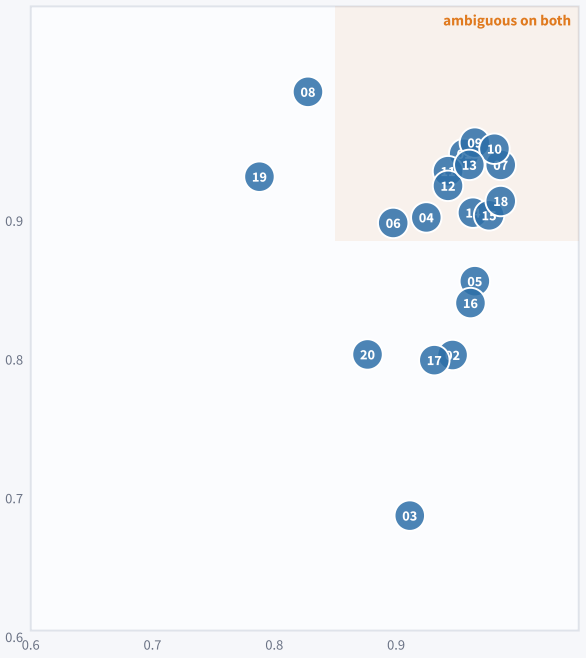
Situation	p	D	Item-rest r [95%]	Entropy B / W	Reading
13 CB-07-C	0.54	0.25	0.36 [-0.05, 0.66]	0.96 / 0.94	works, but picks are spread → review key wording
14 CB-08-A	0.60	0.35	0.45 [0.05, 0.72]	0.96 / 0.91	works, but picks are spread → review key wording
15 CB-09-A	0.55	0.44	0.58 [0.23, 0.79]	0.98 / 0.90	works, but picks are spread → review key wording
16 CB-10-A	0.67	0.46	0.59 [0.24, 0.80]	0.96 / 0.84	works, but picks are spread → review key wording
17 CB-11-B	0.67	0.33	0.46 [0.07, 0.73]	0.93 / 0.80	works, but picks are spread → review key wording
18 CB-11-C	0.51	0.25	0.26 [-0.16, 0.60]	0.99 / 0.91	weak · small gap between top and bottom
19 CB-12-A	0.58	0.42	0.50 [0.12, 0.75]	0.79 / 0.93	works, but picks are spread → review key wording
20 CB-12-C	0.62	0.60	0.55 [0.18, 0.78]	0.88 / 0.80	works, but picks are spread → review key wording

Readings are rule-based (D < 0.25 → weak, entropy ≥ 0.85 → spread, p < 0.76 → hard).

AMBIGUITY MAP · PICK ENTROPY

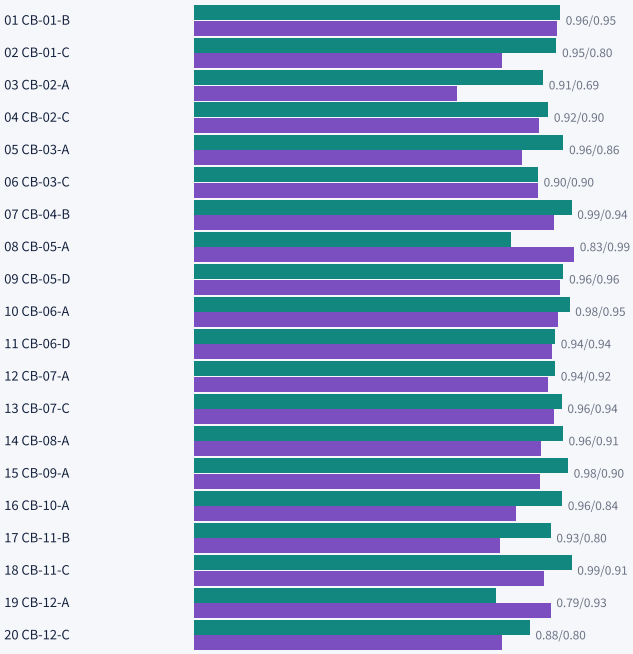
Picks spread widely in 20 situations. The item review should tell whether this is people's judgment or item ambiguity.

Best × Worst entropy



Entropy = how evenly picks spread over 5 options (0 = everyone the same, 1 = perfectly even).

Entropy by situation · Best / Worst



Key-review candidates · entropy 0.85+

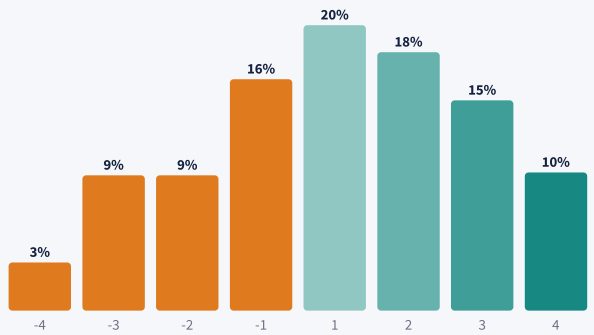
01 **CB-01-B** · Best 0.96 · Worst 0.95 · top Best key 5 (29%) · top Worst key 1 (33%)
02 **CB-01-C** · Best 0.95 · Worst 0.80 · top Best key 5 (29%) · top Worst key 5 (33%)
03 **CB-02-A** · Best 0.91 · Worst 0.69 · top Best key 1 (33%) · top Worst key 2 (54%)
04 **CB-02-C** · Best 0.92 · Worst 0.90 · top Best key 5 (38%) · top Worst key 1 (33%)
05 **CB-03-A** · Best 0.96 · Worst 0.86 · top Best key 5 (29%) · top Worst key 1 (42%)
06 **CB-03-C** · Best 0.90 · Worst 0.90 · top Best key 4 (33%) · top Worst key 2 (33%)
07 **CB-04-B** · Best 0.99 · Worst 0.94 · top Best key 5 (25%) · top Worst key 3 (29%)
08 **CB-05-A** · Best 0.83 · Worst 0.99 · top Best key 2 (33%) · top Worst key 1 (25%)
09 **CB-05-D** · Best 0.96 · Worst 0.96 · top Best key 1 (29%) · top Worst key 1 (29%)
10 **CB-06-A** · Best 0.98 · Worst 0.95 · top Best key 2 (25%) · top Worst key 2 (33%)
11 **CB-06-D** · Best 0.94 · Worst 0.94 · top Best key 1 (33%) · top Worst key 1 (33%)
12 **CB-07-A** · Best 0.94 · Worst 0.92 · top Best key 4 (33%) · top Worst key 1 (38%)
13 **CB-07-C** · Best 0.96 · Worst 0.94 · top Best key 5 (25%) · top Worst key 2 (29%)
14 **CB-08-A** · Best 0.96 · Worst 0.91 · top Best key 5 (33%) · top Worst key 2 (42%)
15 **CB-09-A** · Best 0.98 · Worst 0.90 · top Best key 1 (29%) · top Worst key 2 (38%)
16 **CB-10-A** · Best 0.96 · Worst 0.84 · top Best key 5 (29%) · top Worst key 1 (42%)
17 **CB-11-B** · Best 0.93 · Worst 0.80 · top Best key 5 (38%) · top Worst key 1 (33%)
18 **CB-11-C** · Best 0.99 · Worst 0.91 · top Best key 1 (25%) · top Worst key 2 (33%)
19 **CB-12-A** · Best 0.79 · Worst 0.93 · top Best key 3 (38%) · top Worst key 1 (38%)
20 **CB-12-C** · Best 0.88 · Worst 0.80 · top Best key 5 (42%) · top Worst key 1 (38%)

High entropy can mean the item is ambiguous or several actions all look reasonable. Review candidates. Keys are not changed from this data.

CHOICE QUALITY · GAP AND INVERSIONS

24% of responses have a gap of +3 or more. Inversions are most frequent in CB-05-D (50%).

Judgment gap distribution · key(Best) − key(Worst)

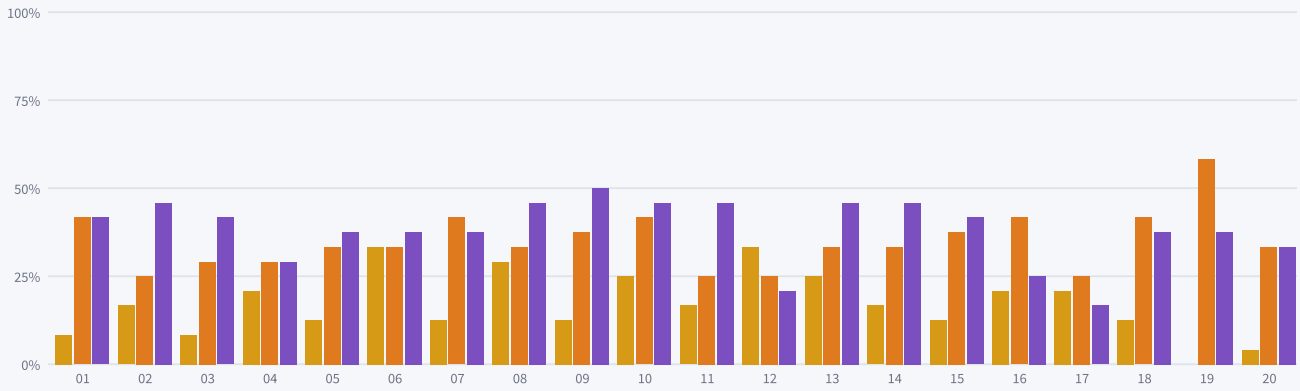


24×20 = 480 responses. Orange = inverted (negative).

Gap summary

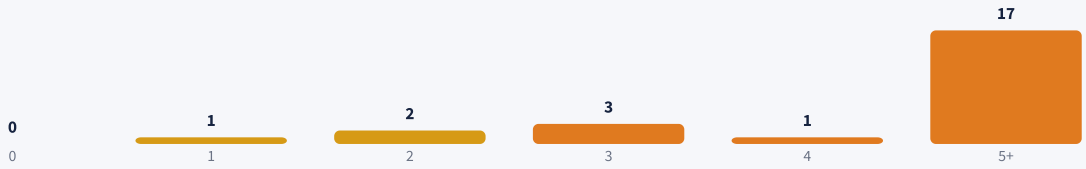
- +4 (perfect order): 10%
- +3 or more: 24%
- Inverted: 38% · 7.6 per person
- People with no inversion: 0
- 3+ inversions: 21 people → interview check

Near-miss · far-miss · inversion rate by situation



Near-misses are a light signal; far-misses and inversions set the training and interview priorities. (near = key 4 · far = key 2-3 · inversion = gap < 0)

Inversions per person



Calculation basis · computed from 480 responses by this cohort's 24 members. Total inversions 183 (38.1%) · gap +4 (perfect order) 10% · +3 or more 24% · people with no inversion 0 · 3+ inversions 21. Inversion = gap < 0; gap = key(Best) − key(Worst).

VALUE TENSION MAP

Integrity draws the most 25%+ rejection: 6 of 24 people.

Judgment tension vs company priority values · people distribution (in-house value framework draft, G4 default)

	reject 0%	reject 1–24%	reject 25–49%	reject 50%+	median support
Integrity	3	15	6	0	65%
Customer	6	15	3	0	58%
Team	5	16	3	0	67%
Results · speed	19	0	3	2	100%
Process	3	18	3	0	60%
Autonomy	19	0	0	5	100%

Cell = people by the share of defensible (non-harmful) options linked to this value that they picked as Worst · harmful options excluded · no total fit score · no ◆ marks until the culture profile is confirmed

Integrity
6
people reject 25%+

Results · speed
5
people reject 25%+

Autonomy
5
people reject 25%+

Customer
3
people reject 25%+

Team
3
people reject 25%+

Process
3
people reject 25%+

How to read this, and limits

- The priority value with the most tension is 'Integrity' (6 people). This is a point to explain in onboarding why the company values it, not a hiring criterion.
- Value priorities are the sample company's draft framework (G4 default).
- These are priorities shown in judgments of 'what works', not personal values.

Onboarding conversation guide · by priority value

Integrity
In week one, share one real decision that shows why integrity comes first here and how it plays out.

Customer
In week one, share one real decision that shows why customer comes first here and how it plays out.

Team
In week one, share one real decision that shows why team comes first here and how it plays out.

Results · speed
In week one, share one real decision that shows why results · speed comes first here and how it plays out.

Process
In week one, share one real decision that shows why process comes first here and how it plays out.

Autonomy
In week one, share one real decision that shows why autonomy comes first here and how it plays out.

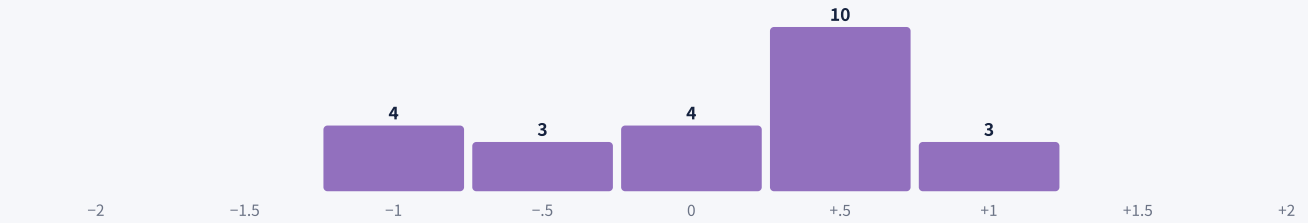
Start with the value that shows the most tension · run it with the team, not person by person · not a hiring criterion.

Descriptive · pre-calibration

DECISION STYLE DISTRIBUTION

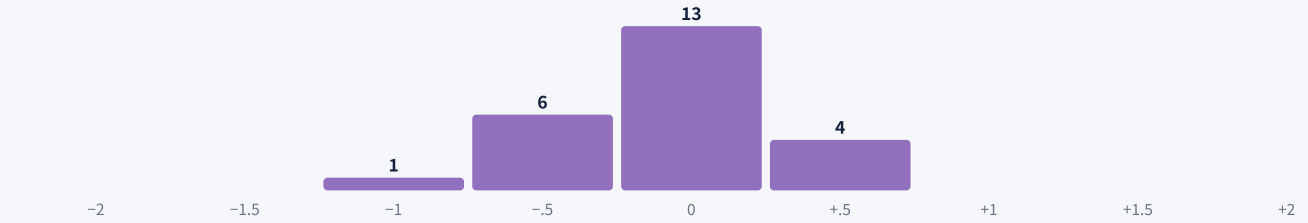
This cohort leans Collaborative. Use it to check balance when forming teams.

Directive ↔ Collaborative · median +0.60



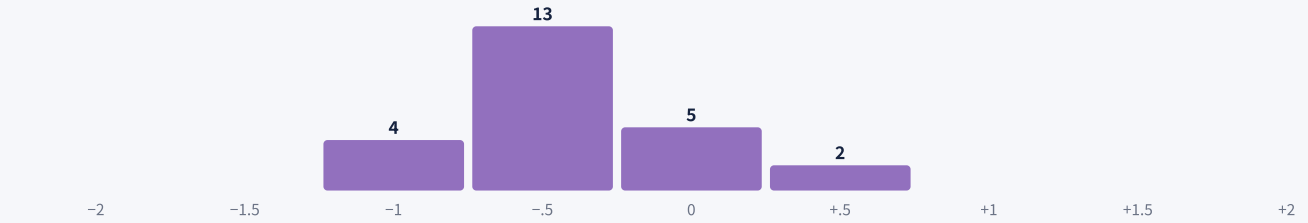
← Directive · Collaborative →

Speed ↔ Consensus · median +0.15



← Speed · Consensus →

Principle ↔ Pragmatic · median -0.15



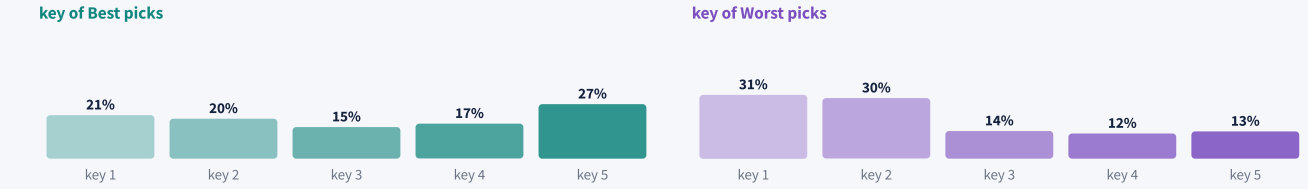
← Principle · Pragmatic →

Style means by type

Type	Directive↔Collaborative	Speed↔Consensus	Principle↔Pragmatic	N
Balanced	+1.17	+0.17	−0.63	2
Insufficient evidence on both axes	−0.31	−0.13	+0.29	7

Score −2 to +2 = the approach in options picked as Best vs Worst · style is a difference in approach, not better or worse · use it to see team-mix diversity · faking resistance not yet validated.

Where Best and Worst picks land on the key · all 20 situations



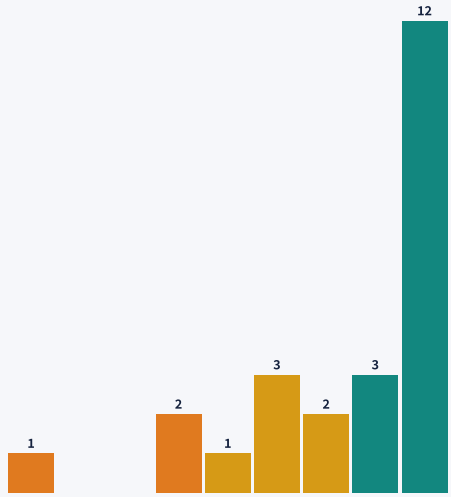
24 people × 20 situations = 480 responses. The more Best picks sit near key 5 and Worst picks near key 1, the more the cohort judges in the same direction.

MEASUREMENT CONSISTENCY · TYPE STABILITY

Split-half 0.88, α 0.88. 67% of people keep the same type under resampling.

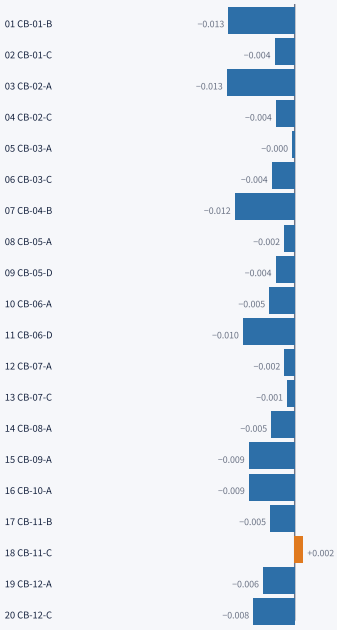
Split-half (odd/even, S-B)	Cronbach α	Type stable 80%+	Shown as borderline
0.88	**0.88**	**67%**	**15**
90% 0.78–0.94	90% 0.79–0.92	16 people	shown-type uncertainty rule

Type stability per person



Share of 1,000 item resamples giving the same type. Teal 80%+, gold 60–80%, orange below 60%.

How α changes if an item is removed



Point-estimate type → type shown in the report

Point-estimate type	Balanced	Strong on effe	Strong on risk	Insufficient e	Borderline
Balanced	2				5
Strong on effective action, risk identification to develop					2
Strong on risk identification, effective action to develop					8
Insufficient evidence on both axes				7	

The shown type follows the same scorer rule as the roster. A raw undifferentiated type stays undifferentiated; other raw types are shown as Borderline when either bootstrap lower bound is at or below the cut.

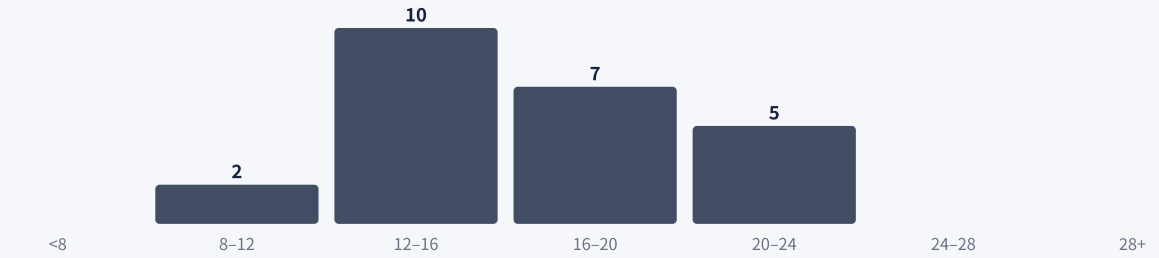
What these figures mean, and do not

- Descriptive values from this 24-person cohort, not a validity result.
- With real responses, the same method is rerun and published with a version.
- Based on 20 items; the values change if the item count changes.

RESPONSE INTEGRITY OVERVIEW

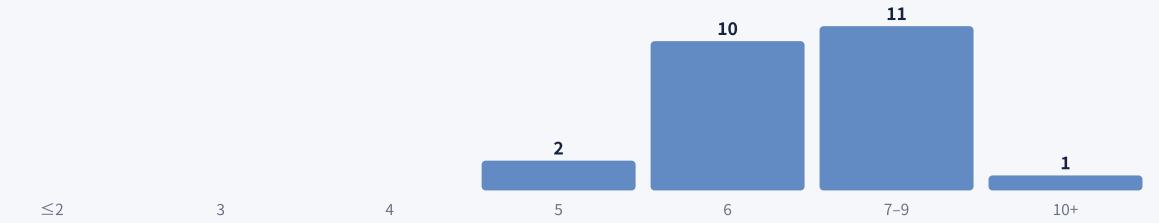
1 of 24 members carry a review signal. The rest are within normal range on both speed and position pattern.

Median response time per person (s)



Under 8 s = speed signal · short for reading a 20-scenario stem

Max times Best was picked at the same screen position



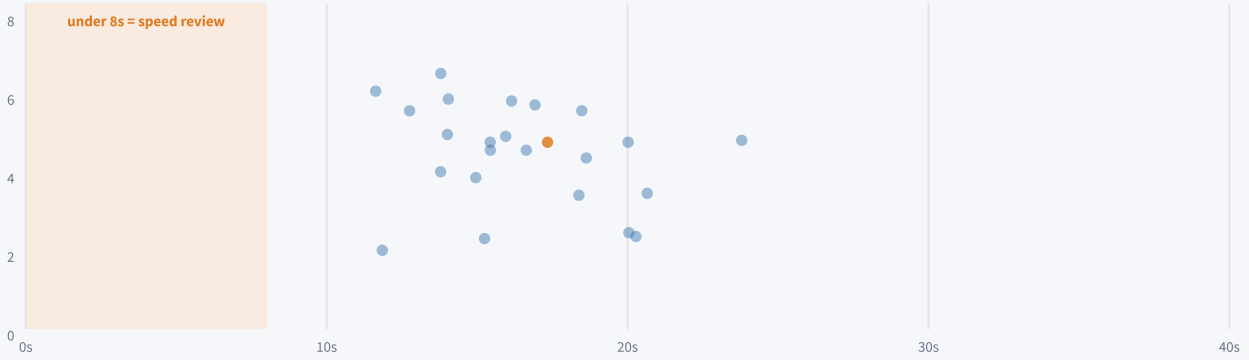
Options are randomized per person · same position 15+ of 20 = position-pattern signal (chance probability < 0.1%)

Responses with signals

ID	Signal	Median time	Position max	Type	Suggested
Sample 006	Position pattern	17.3s	19/20	Borderline	Human review first · retest if needed

Integrity signals are shown separately and never combined with judgment scores. A signal is for review, not proof of misconduct. Types and risky picks of flagged responses need careful reading.

Response time × mean item score · 24 people

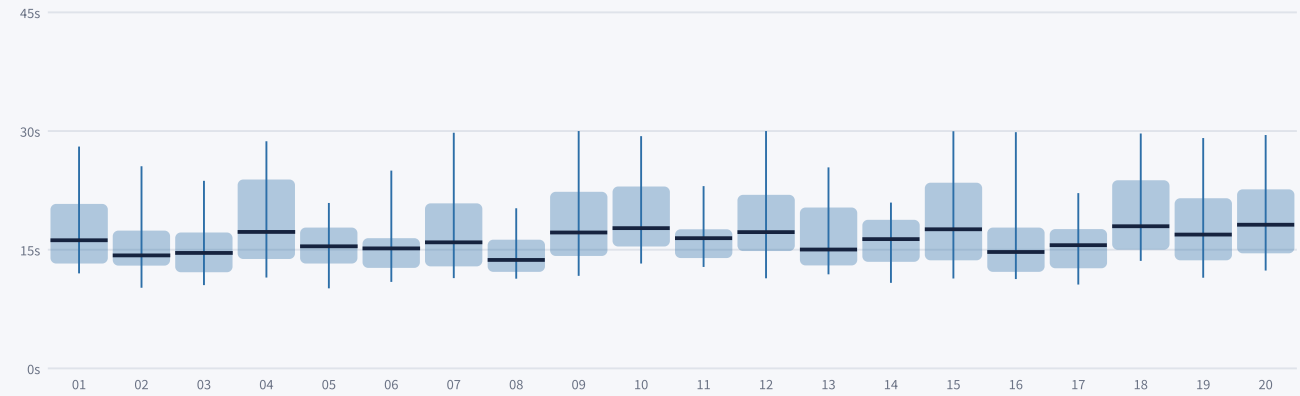


Orange = responses with a review signal. Fast and low-scoring dots come first for review. A signal is not proof.

RESPONSE TIME IN DEPTH

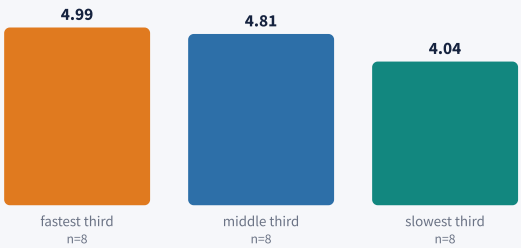
The judgment gap between the fastest and slowest responders is only 0.96. Speed is not a screening criterion.

Response time by item · 24 people



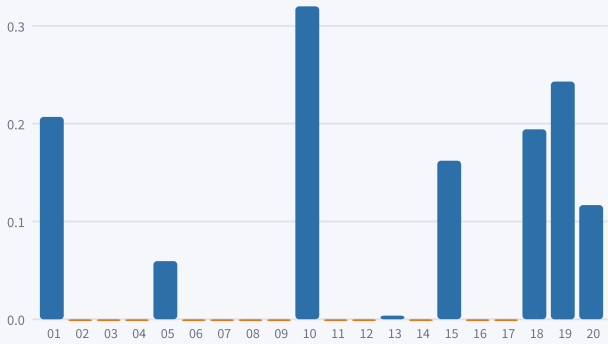
Box = middle 50% · line = median · whiskers = 5–95%.

Mean item score by speed group



Split into thirds by median response time.

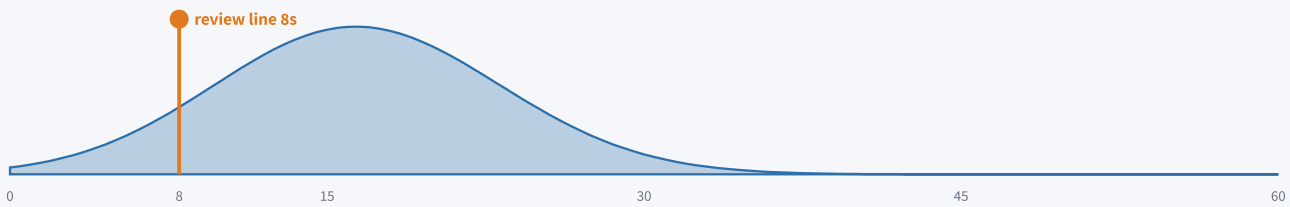
Time–score correlation by item



What can be said about speed

- Fastest third mean 4.99, slowest third 4.04. The difference is small, -0.96.
- Median below 8s (review signal): 0 people.
- Time explains little about judgment quality, so it is not used to screen people. Signals are for review, not proof.

Median response time per person



Orange line = the 8s review threshold. People to its left get a speed review signal.

ROSTER · ID ORDER (NO RANKING)

Two-axis scores, type and interview focus for all 24 people across this section.

ID	Eff.	Harm	Type	Stab.	Risky	Sig.	Interview focus
Sample 001	0.80	0.75	Balanced	99%	2●		⑤ ⑭
Sample 002	0.26	0.28	Low evid.	100%	15●		① ③ ⑤ ⑥ ⑧ ⑨ ⑩ ⑪ ⑫ ⑬ ⑭ ⑮ ⑯ ⑱ ⑳
Sample 003	0.80	0.63	Borderline	55%	2		⑧ ⑬
Sample 004	0.20	0.70	Borderline	87%	17●		① ③ ④ ⑤ ⑥ ⑦ ⑧ ⑩ ⑪ ⑫ ⑬ ⑭ ⑮ ⑯ ⑰ ⑱ ⑲
Sample 005	0.61	0.61	Low evid.	31%	5●		③ ④ ⑨ ⑮ ⑳
Sample 006	0.53	0.70	Borderline	82%	7●	⚑	② ⑤ ⑨ ⑪ ⑬ ⑮ ⑰
Sample 007	0.47	0.79	Borderline	96%	8●		② ③ ⑪ ⑬ ⑮ ⑯ ⑰ ⑱
Sample 008	0.43	0.61	Low evid.	52%	10●		① ③ ④ ⑤ ⑥ ⑧ ⑫ ⑮ ⑰ ⑱
Sample 009	0.72	0.74	Borderline	85%	5●		② ⑥ ⑧ ⑬ ⑭
Sample 010	0.24	0.38	Low evid.	100%	14●		① ② ③ ⑤ ⑥ ⑨ ⑩ ⑪ ⑭ ⑮ ⑯ ⑰ ⑲ ⑳
Sample 011	0.79	0.70	Borderline	88%	2●		⑤ ⑩
Sample 012	0.09	0.56	Low evid.	76%	20●		① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪ ⑫ ⑬ ⑭ ⑮ ⑯ ⑰ ⑱ ⑲ ⑳
Sample 013	0.74	0.76	Borderline	94%	5		⑨ ⑩ ⑬ ⑱ ⑲
Sample 014	0.51	0.66	Borderline	68%	8●		④ ⑦ ⑧ ⑫ ⑭ ⑮ ⑯ ⑱
Sample 015	0.45	0.82	Borderline	99%	10●		① ⑤ ⑥ ⑦ ⑨ ⑪ ⑫ ⑰ ⑲ ⑳
Sample 016	0.66	0.46	Borderline	70%	5●		② ⑦ ⑪ ⑬ ⑱
Sample 017	0.80	0.86	Balanced	99%	3●		② ⑨ ⑪
Sample 018	0.24	0.39	Low evid.	100%	16●		① ② ③ ④ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪ ⑬ ⑭ ⑮ ⑰ ⑲ ⑳
Sample 019	0.79	0.44	Borderline	99%	2		④ ⑤
Sample 020	0.25	0.75	Borderline	99%	14●		② ③ ④ ⑦ ⑧ ⑨ ⑩ ⑪ ⑫ ⑭ ⑮ ⑯ ⑱ ⑲
Sample 021	0.56	0.68	Borderline	62%	5●		③ ⑧ ⑩ ⑮ ⑳
Sample 022	0.76	0.66	Borderline	75%	3		⑦ ⑨ ⑩
Sample 023	0.39	0.79	Borderline	100%	10●		① ② ⑥ ⑦ ⑧ ⑨ ⑩ ⑪ ⑫ ⑱
Sample 024	0.41	0.47	Low evid.	94%	8●		① ② ⑤ ⑥ ⑦ ⑨ ⑱ ⑳

Eff./Harm = 0–1 axes (cut 0.625) · Stab. = same-type share over 1,000 resamples · Risky = harmful Best picks, ● = Critical risky pick · ⚑ = integrity review signal · interview focus = scenarios with risky picks · type shows Borderline when the shown-type uncertainty rule.

Descriptive · pre-calibration

How everything was computed, and what this report does not do.

What each N tier shows

N	Tier	What is shown
< 20	Blocked	Individual reports only · no cohort report
20–49	Small	Distributions, risk concentration, interview bank, integrity, roster · discrimination for reference only (wide CI) · no item actions · value/style distributions hidden
≥ 50	Standard	All pages · item-health actions · provisional norms possible (after key + calibration approval)

This cohort is N=24, Small tier (green row).

Definitions

Two-axis scores	x = mean (key(B)−1)/4, y = mean (5−key(W))/4. Descriptive before calibration, not θ.
Intervals · borderline	90% interval over 1,000 item resamples. If it crosses the cut (0.625), no type is set.
Risky picks	E1 = harmful option (severity 1–3) as Best · E2 = model action (key 5) as Worst.
Discrimination	Corrected item-total correlation, Fisher z 95% CI.
Integrity	Median time < 8 s, Best at the same screen position ≥ 15/20. Never combined with scores.
Versions	key_version theta-key-2026-10-05 · scoring bw-s1-v1. Published results are never overwritten; re-scoring is a new version.

What we do not do

- Rankings, composite scores, pass/fail verdicts
- θ or percentiles before calibration
- Folding integrity signals into scores
- Showing values/style to clients

What changes after calibration

- The internal scoring engine computes the judgment score from calibrated item parameters, shown with SE and 95% interval
- Percentiles only under a named, frozen reference group
- Borderline calls sharpened by item parameters
- The descriptive scores in this report stay as they are

Pre-calibration · descriptive | key_version theta-key-2026-10-05 | Reference cut 0.625 | No ranking · no composite | Missing stays missing

S1 score table · Best key × Worst key

key of the Worst pick

	1	2	3	4	5
B key 5	8	7	6	5	B=W ×
B key 4	7	6	5	B=W ×	3
B key 3	6	5	B=W ×	3	2
B key 2	5	B=W ×	3	2	1
B key 1	B=W ×	3	2	1	0

How to read the table

- Item score = (Best key − 1) + (5 − Worst key), 0–8
- Model action as Best and most harmful as Worst = 8
- The same option cannot be both Best and Worst (rejected by the server)
- This table is the only source of every score in this report. No weights, no hidden adjustment.

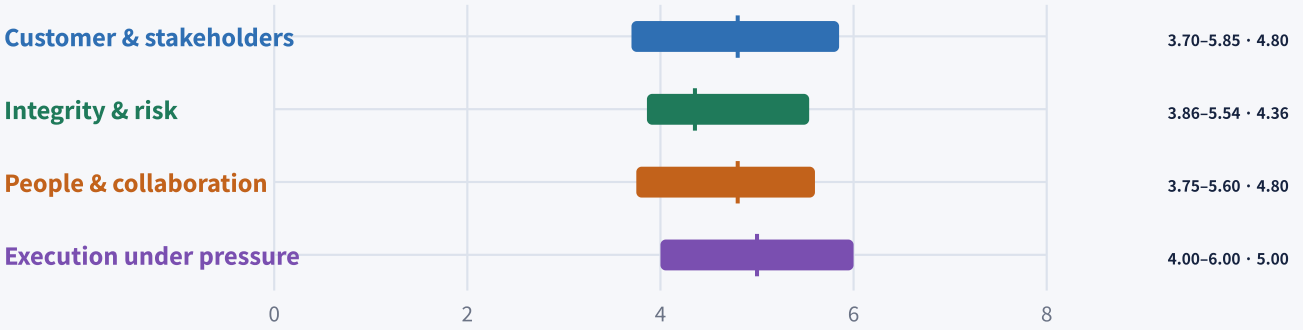
COMPETENCY INTERPRETATION · PRE-CALIBRATION DESCRIPTIVE

Four lenses, one assessment

Best + Worst credits describe four sample competencies. They are not an IRT θ score or an ability percentile.

Cohort	Competency lenses	Provisional floor	Situations
24	4	20	20
same assessment	Best + Worst credits	below 20: report blocked	5 options, Best + Worst

Four distributions · Best + Worst credit / 8 Descriptive



Band = middle half; line = median; All 24 records completed the 20 situations.

Publication gate

Provisional cohort

24 / 24 eligible | provisional floor 20

n ≥ 20 required

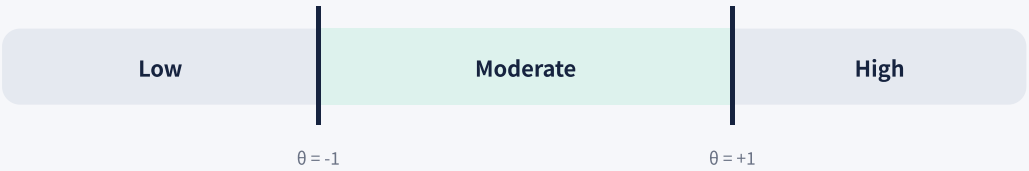
The provisional-synthetic report is blocked below n=20. The 20–49 Small tier needs a separately approved production-frozen asset.

Classification, in scope

Existing Best/Worst type labels describe two provisional cohort judgment axes. They are descriptive, not IRT Ability Bands. Read the type alongside its stability interval and response evidence; do not rank people by type.

IRT band thresholds · no candidate value

PRE-CALIBRATION · CANDIDATE POSITION SHOWN AFTER CALIBRATION APPROVAL



This scale diagrams source thresholds, not a measured result. Hypothetical θ=+0.42 is separate from this sample.

Boundary before interpretation

No θ, frozen calibration or approved reference version exists for this sample, so no actual Ability Band or Ability Percentile is available. Source thresholds: Low θ < -1, Moderate -1 ≤ θ ≤ 1, High θ > 1. Only a frozen item calibration without local refit allows a percentile. A separate hypothetical θ=+0.42 would be Moderate and 66.3, not this candidate's result or a real norm.

No system-made hiring verdict, candidate order or composite percentile. People decide.

Appendix R · Relative-standing reference

Quartiles and ranks · the cohort comparison baseline.

Axis score quantiles · exact from the 24-member roster

Metric	P10	P25	P50 (median)	P75	P90
Effective-action ID	0.24	0.36	0.52	0.74	0.80
Risk ID	0.40	0.54	0.67	0.75	0.79
Axis mean	0.32	0.49	0.61	0.71	0.75

Cut 0.625 sits between P50 and P75 - see the p3 legend for band edges

Competency quartiles · from the p10 band summary

Competency	P25	Median	P75
Customer & stakeholders	3.7	4.8	5.8
Integrity & risk	3.9	4.4	5.5
People & collaboration	3.8	4.8	5.6
Execution under pressure	4.0	5.0	6.0

Calculation and audience

Each axis quantile is computed from all 24 roster members, including Borderline members. P10, P25, P50, P75 and P90 use linear interpolation between sorted values. Competency quartiles use the same cohort and assigned source items. The cut is a descriptive classification boundary, not a population percentile. These reference values are specific to this cohort and form version.

Appendix R · Relative-standing reference

Quartiles and ranks · the cohort comparison baseline. · continued

Situation ranks · by mean, ties share a rank

Rank	Situation	Mean	SD	Median	Q1–Q3	6–8 band
1	**CB-11-B**	**5.4**	1.9	6.0	5.0–7.0	54.2%
2	**CB-10-A**	**5.3**	2.1	6.0	4.5–7.0	54.2%
3	**CB-02-C**	**5.2**	2.0	5.0	3.0–7.0	41.7%
T-4	**CB-07-A**	**5.0**	2.3	5.5	5.0–6.3	50.0%
T-4	**CB-12-C**	**5.0**	2.7	5.5	3.0–7.3	50.0%
6	**CB-02-A**	**4.9**	2.2	6.0	3.0–7.0	54.2%
7	**CB-08-A**	**4.8**	2.4	6.0	3.0–7.0	54.2%
8	**CB-12-A**	**4.7**	2.3	5.0	2.8–6.0	41.7%
9	**CB-03-A**	**4.6**	2.7	5.0	2.8–6.3	45.8%
T-10	**CB-01-B**	**4.5**	2.3	5.0	3.0–6.0	41.7%
T-10	**CB-04-B**	**4.5**	1.9	5.0	3.0–5.3	25.0%
T-12	**CB-05-A**	**4.4**	2.5	5.0	2.8–7.0	41.7%
T-12	**CB-06-D**	**4.4**	2.4	5.0	2.8–7.0	41.7%
T-12	**CB-09-A**	**4.4**	2.3	5.0	3.0–6.3	37.5%
T-15	**CB-03-C**	**4.3**	2.4	5.0	2.0–6.0	41.7%
T-15	**CB-06-A**	**4.3**	2.2	5.0	2.8–6.0	33.3%
T-15	**CB-07-C**	**4.3**	2.4	5.0	2.0–6.0	41.7%
18	**CB-01-C**	**4.2**	2.6	5.0	2.8–6.3	37.5%
19	**CB-11-C**	**4.1**	2.3	5.0	2.5–6.0	29.2%
20	**CB-05-D**	**4.0**	2.2	4.0	2.0–5.3	25.0%

Definitions · P10–P90 are linearly interpolated quantiles of all 24 roster members incl. borderline · situation ranks are competition ranks by mean (ties share) · competency quartiles cite the p10 band summary

Reference baseline for within-cohort comparison only - not for individual grading