

Theta AI™

Cohort Insight Report

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

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.

Sample cohort

120 members · 33 pages · Sample v6.1 · v1.1

Pre-calibration · descriptive

Scoring key 2026-09-23

scoring bw-s1-v1

No composite rank · no verdict

Missing stays missing

- | | | | |
|----|--|----|------------------------------|
| 01 | Executive summary | 22 | Ambiguity map |
| 02 | Visual dashboard | 23 | Gap · inversions |
| 03 | Comparison evidence section (pages 3–11) | 24 | Values priorities |
| 12 | Risk concentration | 25 | Decision style |
| 13 | Competency profile | 26 | Consistency · type stability |
| 14 | Development · onboarding | 27 | Response integrity |
| 15 | Interview questions | 28 | Response time in depth |
| 16 | Interviewer guide in depth | 29 | Roster |
| 17 | Judgment landscape | 30 | Method · limits |
| 18 | Choice distributions 12×5 | | |
| 19 | Distractor analysis | | |
| 20 | Item health · key check | | |
| 21 | Item discrimination story | | |

The one-minute conclusion for leaders.

Every number comes from a chart or table later in the report.

The conclusion in one sentence

53 of 120 are above the cut on both axes, 64 are borderline, and the remaining 3 are one-axis strength (1) and insufficient-evidence (2). Harmful Best picks cluster in Priority conflict, Schedule crisis, Feedback: start interviews there, and onboarding with 'Customer & stakeholders'.

3 strengths · situations the group judged best

- 09 Safety issue · mean 6.4/8 · model action = Best 60%
- 01 Client complaint · mean 6.3/8 · model action = Best 59%
- 10 Resource shortage · mean 6.2/8 · model action = Best 59%

3 points to check · where harmful Best picks cluster

- 04 Priority conflict · harmful Best 28 (23%) · mean 4.6/8
- 07 Schedule crisis · harmful Best 22 (18%) · mean 6.2/8
- 08 Feedback · harmful Best 20 (17%) · mean 5.7/8

3 questions to ask in interviews

- 04 “In a situation like Priority conflict, what did you do first, and why?” · strong answer leans toward “Align both execs on priorities together”
- 07 “In a situation like Schedule crisis, what did you do first, and why?” · strong answer leans toward “Tell the client first, offer a new timeline”
- 08 “In a situation like Feedback, what did you do first, and why?” · strong answer leans toward “Sort out the criticism, discuss fixes 1:1”

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

4 key findings

- 'Integrity & risk' ties 'Execution under pressure' at the top (median 6.5).
- 91 people (76%) picked a harmful option as Best at least once; far-misses cluster in Priority conflict, New hire, Client complaint.
- 16.3% of responses are inverted, highest in Priority conflict at 42%.
- Response speed and judgment score are barely related (item r -0.13–0.11).

Is the evidence enough?

- **120** people · Standard tier (50 or more)
- Split-half **0.82** (90% interval 0.75–0.87)
- Items working **12/12** (discrimination $r \geq 0.15$)
- **0** key-check items · auto rule ($r < 0.15$)
- Integrity signals: **4** members · for review, not proof
- Descriptive, pre-calibration (Scoring key 2026-09-23 · calibration unavailable)

What to do this week

- First onboarding topic: case reviews in 'Customer & stakeholders'
- Where interview questions go first: Priority conflict, New hire
- First topic = weakest competency median

Limits of the data

- 120-person synthetic data · descriptive before calibration
- No overall rank, 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

- **p12** Risk concentration
- **p14** Development · onboarding plan
- **p15** Interview question bank
- **p16** Interviewer guide in depth
- **p30** Method · tiers · limits

VISUAL DASHBOARD · CHARTS FOR LEADERS

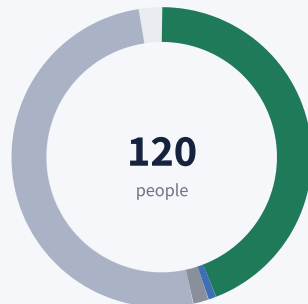
Of 120 people, 53 are above the cut on both axes and 64 are borderline. The other 3: one-axis strength (1), insufficient evidence (2). Risk concentrates most in the Priority conflict scenario.

Judgment landscape · 120 people



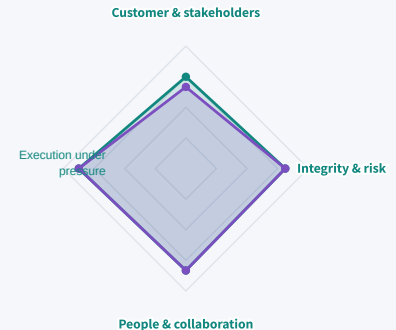
Dot color = type · faded = borderline (interval crosses the cut)
· green area = both axes at or above the 0.625 cut.

Type mix



- 53 Balanced
- 1 Strong on effective action, risk identification to develop
- 0 Strong on risk identification, effective action to develop
- 2 Insufficient evidence on both axes
- 64 Borderline · incl. 3 with review signals

Competency profile · cohort median



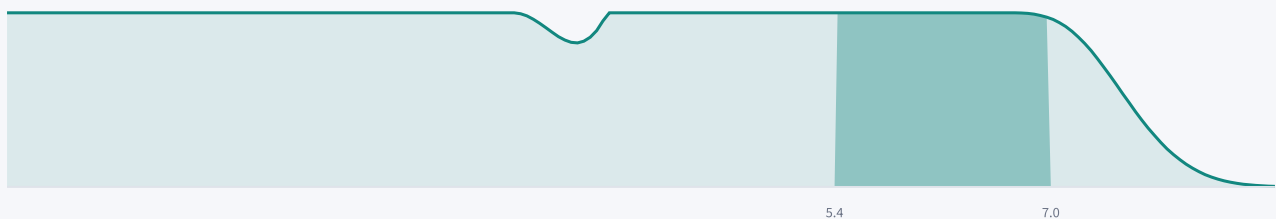
Teal = median Best credit, purple = median Worst credit (0–4) · competency tags are a sample mapping (to confirm).

All 12 situations · score · risk · discrimination

	01	02	03	04	05	06	07	08	09	10	11	12
Mean item score	6.3	6.3	6.3	4.6	6.3	6.2	6.2	5.7	6.4	6.2	6.0	6.0
Harmful as Best %	11	10	3	23	16	8	18	17	7	12	12	9
Discrimination r	0.64	0.52	0.56	0.17	0.51	0.57	0.39	0.39	0.62	0.66	0.57	0.41

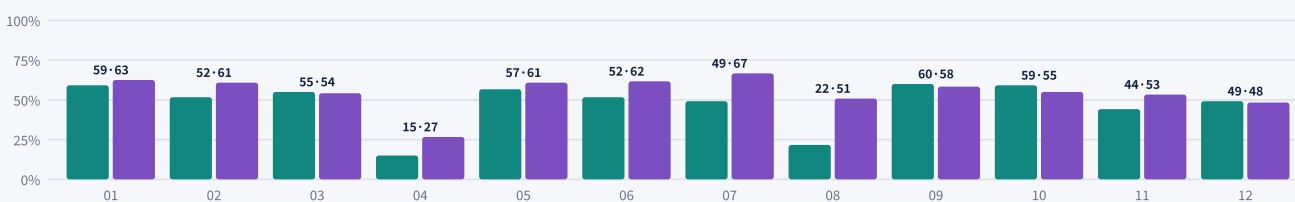
No low-discrimination items (all $r \geq 0.15$).

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

120 members · 1,440 pairs · α 0.84.

Type stability is 80%+ for 70%; 76% picked a harmful option at least once.

Headline indicators

Members

120

members

Response pairs

1,440

items × members

Cronbach α **0.84**

CI 0.78–0.87

Split-half

0.82

CI 0.75–0.87

Any harmful pick

75.8%

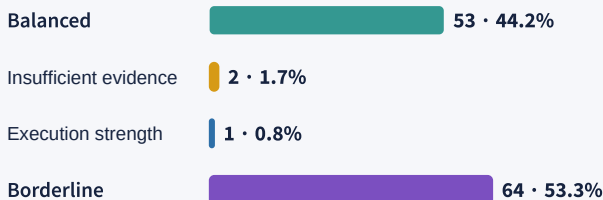
91 members

Type stability $\geq 80\%$ **84**

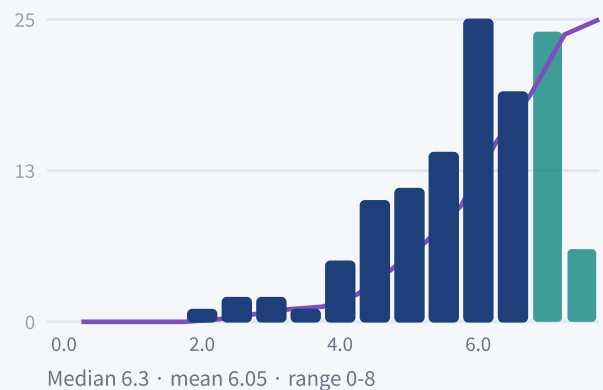
members · 70.0%

Type and score distributions

Judgment-type distribution (N · share)

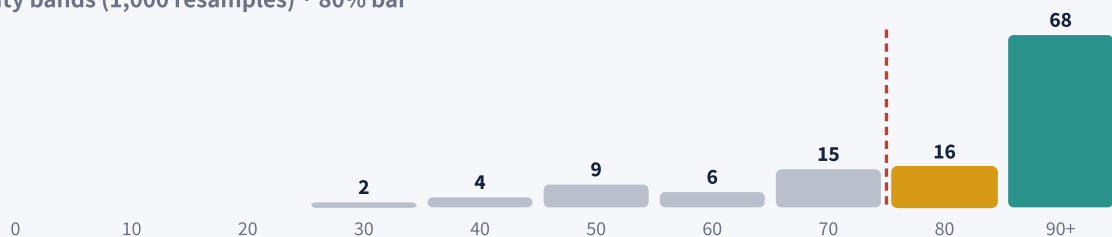


Mean-score distribution · purple line = cumulative share



Stability bands

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



How to read this dashboard

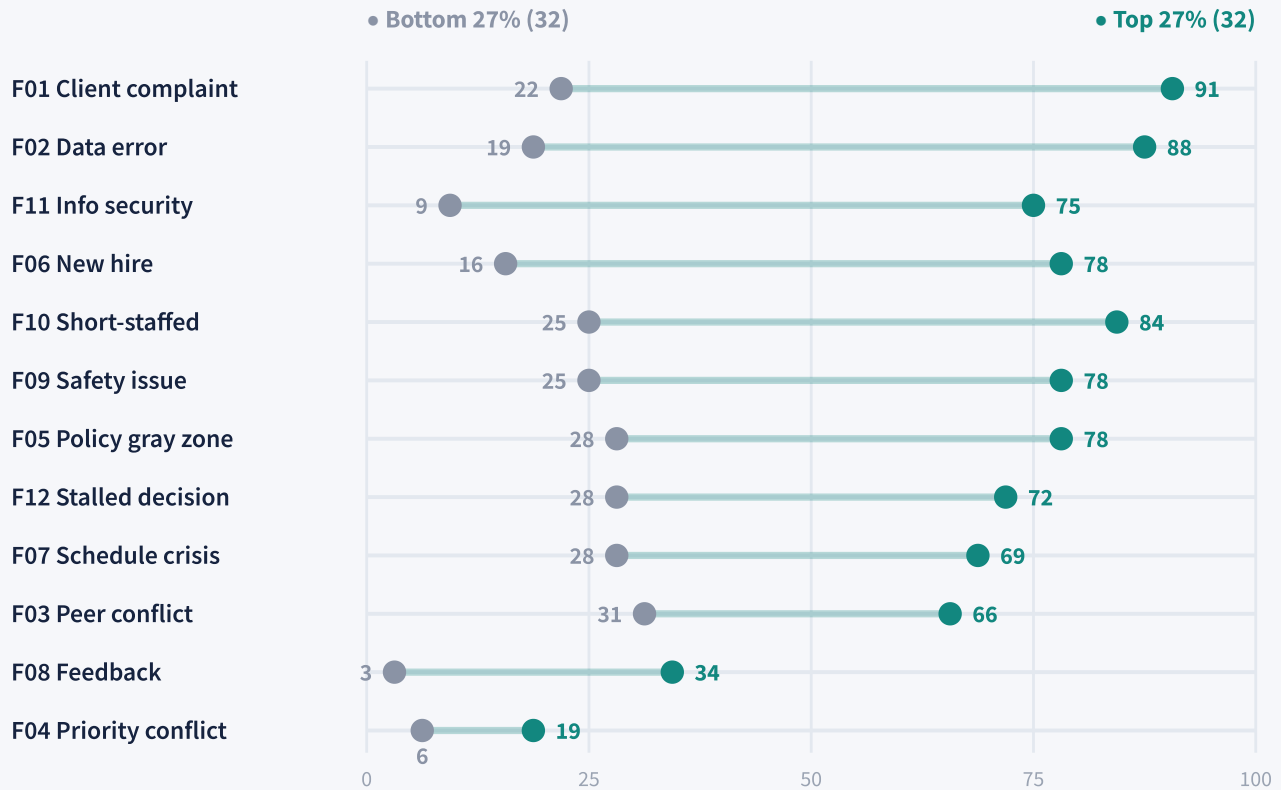
- This page summarizes the eight deep-dive sections that follow (difficulty, means, harmful picks & inversions, discrimination, stability, axis alternation, bands, synthesis). Every figure is computed directly from the 120 × 12 response records.
- The sample is pre-calibration synthetic data. Figures describe the cohort; they cannot grade any person or produce a selection ranking.

RELATIVE-COMPARISON EVIDENCE · 12-SITUATION DIFFICULTY

The top 27% picked the model action at 90%+ in 1 of 12 situations.

The widest between-group gap is Client complaint and Data error (69pp).

Model-pick dumbbells · by situation



Figures

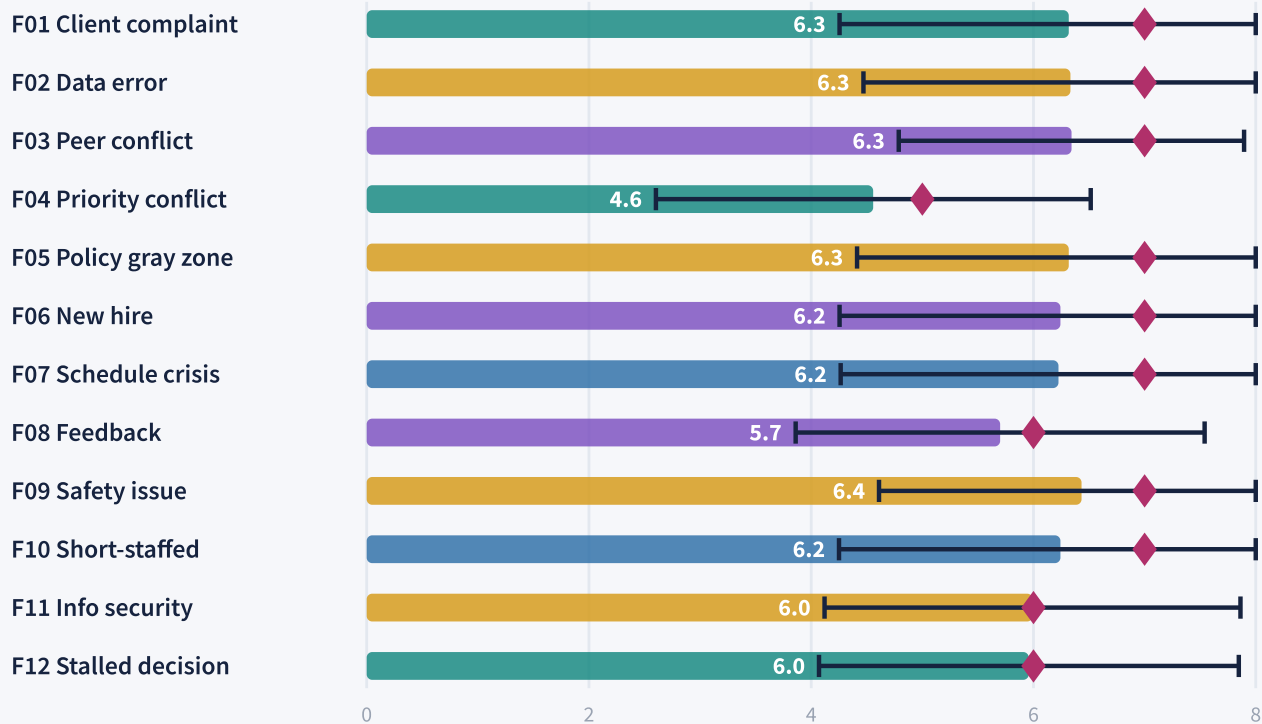
Situation (widest gap first)	Bottom 27%	Top 27%	Gap	Mean /8
F01 Client complaint	22%	91%	69pp	6.3
F02 Data error	19%	88%	69pp	6.3
F11 Info security	9%	75%	66pp	6.0
F06 New hire	16%	78%	63pp	6.2
F10 Short-staffed	25%	84%	59pp	6.2
F09 Safety issue	25%	78%	53pp	6.4
F05 Policy gray zone	28%	78%	50pp	6.3
F12 Stalled decision	28%	72%	44pp	6.0
F07 Schedule crisis	28%	69%	41pp	6.2
F03 Peer conflict	31%	66%	34pp	6.3
F08 Feedback	3%	34%	31pp	5.7
F04 Priority conflict	6%	19%	13pp	4.6

How to read, and what shows

- Dot = share of the group picking the model action (key 5) as Best. Gray = bottom 27% (32), teal = top 27% (32) · groups split by total mean score.
- Wider gaps mark situations where judgment separates most - widest Client complaint and Data error 69pp, narrowest Priority conflict 13pp.

RELATIVE-COMPARISON EVIDENCE · PER-SITUATION MEANS

Situation means span 1.9 points, from 4.6 (Priority conflict) to 6.4 (Safety issue).
Judgment splits most on Client complaint (SD 2.1).

Score distribution by situation · mean $\pm$ SD · medianBar = mean · whiskers = ± 1 SD · diamond = median

Figures

Situation	Mean	SD	Median	Q1-Q3	Range
F01 Client complaint	6.3	2.1	7.0	5.0-8.0	1-8
F02 Data error	6.3	1.9	7.0	6.0-8.0	1-8
F03 Peer conflict	6.3	1.6	7.0	5.0-8.0	1-8
F04 Priority conflict	4.6	2.0	5.0	3.0-6.0	1-8
F05 Policy gray zone	6.3	1.9	7.0	5.0-8.0	1-8
F06 New hire	6.2	2.0	7.0	5.0-8.0	1-8
F07 Schedule crisis	6.2	2.0	7.0	5.0-8.0	1-8
F08 Feedback	5.7	1.8	6.0	5.0-7.0	1-8
F09 Safety issue	6.4	1.8	7.0	5.0-8.0	1-8
F10 Short-staffed	6.2	2.0	7.0	5.0-8.0	1-8
F11 Info security	6.0	1.9	6.0	5.0-7.0	1-8
F12 Stalled decision	6.0	1.9	6.0	5.0-8.0	1-8

How to read, and what shows

- Score = 0-8 per situation (Best rank + (5 - Worst rank)). Every figure is recomputed from the 120 members' raw item scores.
- Highest mean Safety issue 6.4/8, lowest Priority conflict 4.6/8. The widest spread (SD) is Client complaint (2.1) - the situation where this cohort disagrees most.
- Bar colour = competency axis. Low means with long whiskers are cohort-level development priorities.

RELATIVE-COMPARISON EVIDENCE · BOTTOM-KEY PICKS & INVERSIONS

Bottom-key Best picks exceed 10% in 12 situations, peaking at Priority conflict (37.5%). Inversions concentrate on Priority conflict (41.7%).

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



Figures

Situation	Bottom-key	Inversion	Near-inversion	Model Best
F01 Client complaint	23.3%	12.5%	18.3%	59.2%
F02 Data error	10.8%	10.8%	10%	51.7%
F03 Peer conflict	19.2%	7.5%	22.5%	55%
F04 Priority conflict	37.5%	41.7%	24.2%	15%
F05 Policy gray zone	17.5%	15.8%	13.3%	56.7%
F06 New hire	26.7%	12.5%	19.2%	51.7%
F07 Schedule crisis	17.5%	18.3%	11.7%	49.2%
F08 Feedback	12.5%	22.5%	5.8%	21.7%
F09 Safety issue	18.3%	8.3%	17.5%	60%
F10 Short-staffed	15%	14.2%	14.2%	59.2%
F11 Info security	17.5%	15.8%	13.3%	44.2%
F12 Stalled decision	18.3%	15.8%	17.5%	49.2%

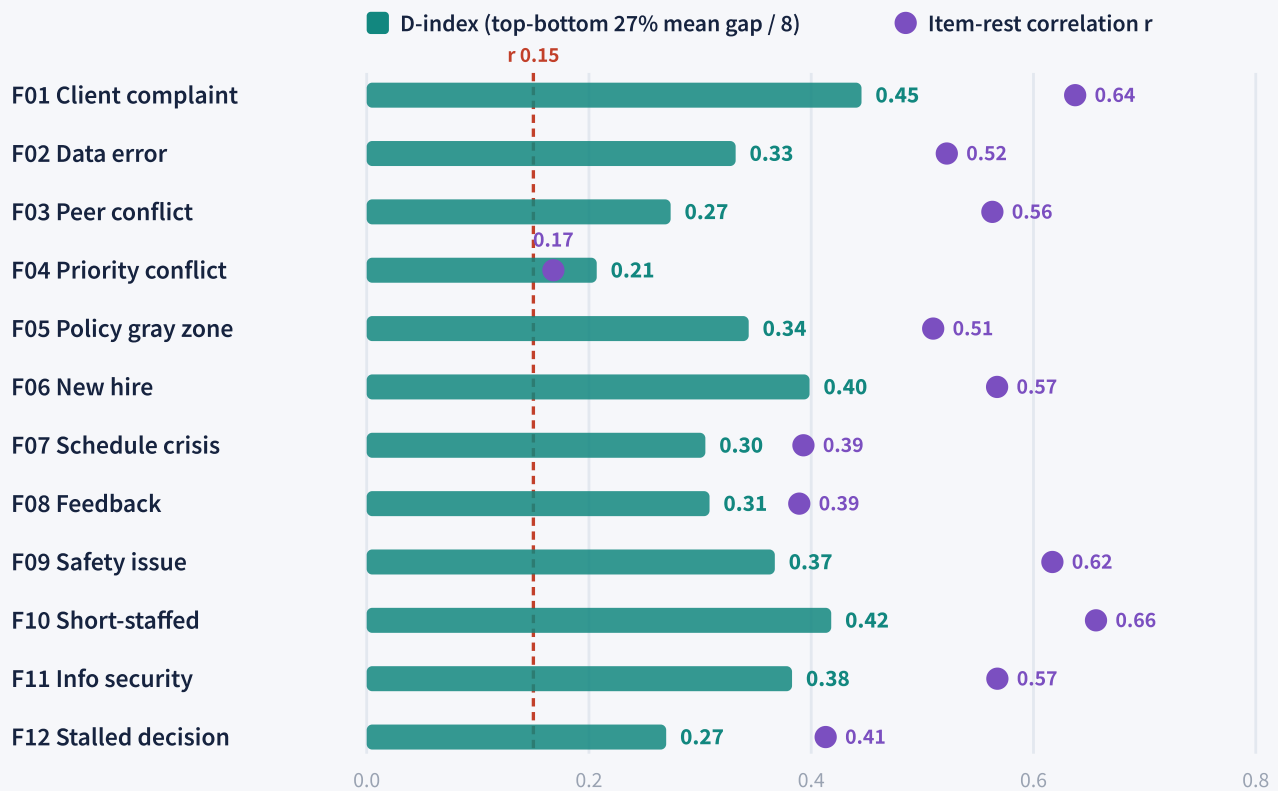
How to read, and what shows

- 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 1,440 response pairs: 235 inversions (16.3%), 225 near-inversions (15.6%). 91 members (75.8%) picked at least one bottom-key Best.
- Situations high on both inversion and near-inversion are key/option check 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 ($r \geq 0.15$).
The strongest separation is Client complaint (D 0.45).

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



Figures

Situation	D-index	Item-rest r	α if removed	Flag
F01 Client complaint	0.45	0.64	-0.024	working
F02 Data error	0.33	0.52	-0.014	working
F03 Peer conflict	0.27	0.56	-0.016	working
F04 Priority conflict	0.21	0.17	0.013	working
F05 Policy gray zone	0.34	0.51	-0.013	working
F06 New hire	0.40	0.57	-0.018	working
F07 Schedule crisis	0.30	0.39	-0.004	working
F08 Feedback	0.31	0.39	-0.004	working
F09 Safety issue	0.37	0.62	-0.021	working
F10 Short-staffed	0.42	0.66	-0.026	working
F11 Info security	0.38	0.57	-0.018	working
F12 Stalled decision	0.27	0.41	-0.006	working

How to read, and what shows

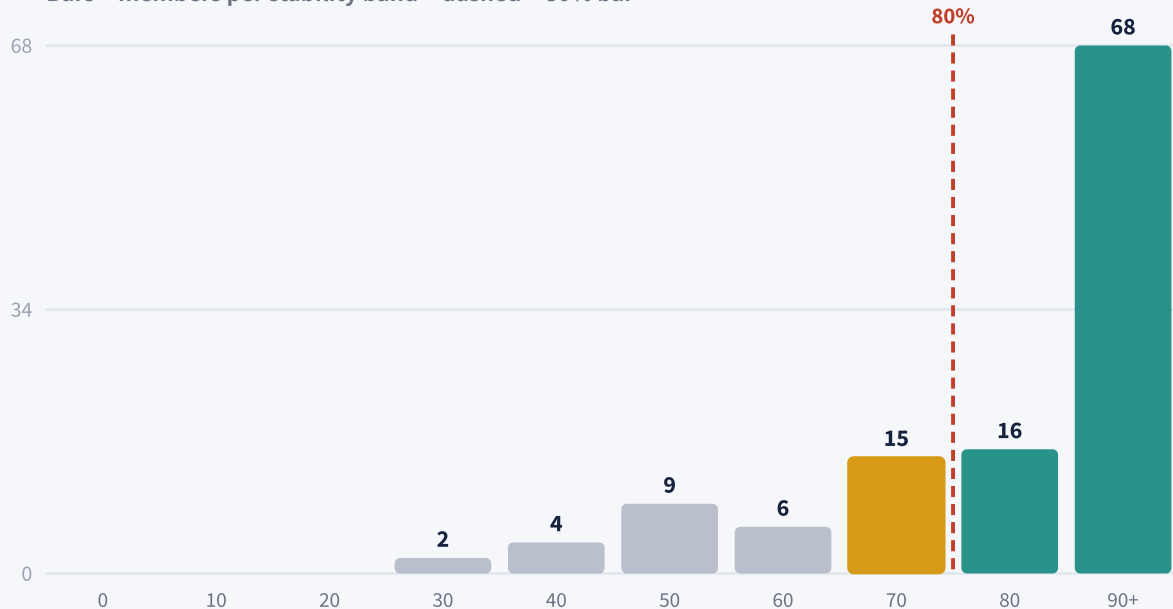
- D-index = (top 27% mean - bottom 27% mean) / 8 · item-rest r = correlation of the item with the rest total · working bar $r \geq 0.15$ (dashed).
- Overall reliability $\alpha = 0.836$. An alpha-if-removed above the overall α marks an item whose removal would raise reliability.
- All 12 items sit above the working bar.

RELATIVE-COMPARISON EVIDENCE · TYPE STABILITY RESAMPLING

84 of 120 members hold their type in 80%+ of 1,000 resamples.
Borderline is the least stable type (mean 74.2%).

Type stability distribution · 1,000 resamples

Bars = members per stability band · dashed = 80% bar



By type

Judgment type	N	Mean stability	≥80%
Balanced	53	99.3%	53
Insufficient evidence	2	98.8%	2
Execution strength	1	99.4%	1
Borderline	64	74.2%	28

How to read, and what shows

- Stability = share of 1,000 deterministic resamples (seed bw-v6|<memberID>) 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 85.9%, borderline (interval touches the cut) 64 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.

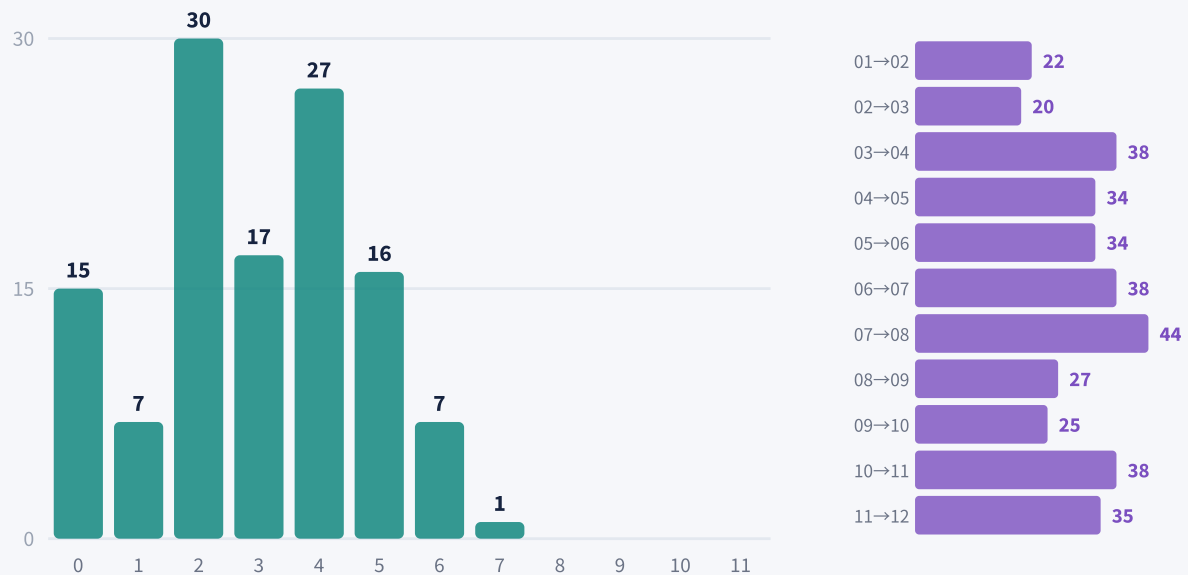
RELATIVE-COMPARISON EVIDENCE · AXIS ALTERNATION

Members alternate axes 3.0 times on average; 15 never alternate.

The busiest boundary is F07→F08 (44 members).

Axis alternation · by member and boundary

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



Figures

Alternations N Share Cumulative

0	15	12.5%	12.5%
1	7	5.8%	18.3%
2	30	25.0%	43.3%
3	17	14.2%	57.5%
4	27	22.5%	80.0%
5	16	13.3%	93.3%
6	7	5.8%	99.2%
7	1	0.8%	100.0%
8	0	0.0%	100.0%
9	0	0.0%	100.0%
10	0	0.0%	100.0%
11	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: 15 members answered consistently on one axis. More alternations = responses led by different axes across situations - a pattern description, not an integrity verdict.
- Busiest boundary: F07→F08 (Schedule crisis → Feedback), 44 members.

RELATIVE-COMPARISON EVIDENCE · BAND MATRIX

68.1% of responses sit in the 6-8 band.
Only 4.3% land in the lowest band (0-2).

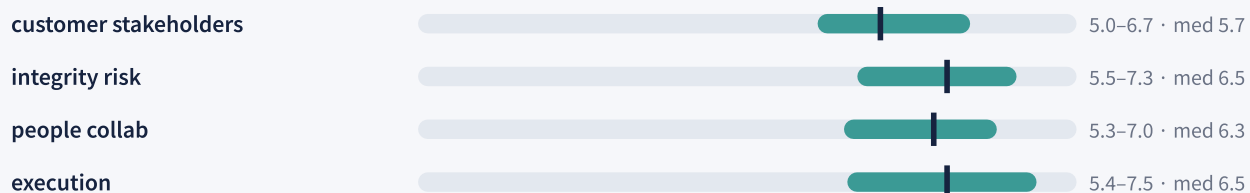
Situation × band matrix

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

	0-2	2-4	4-6	6-8
F01 Client complaint	8	7	22	83
F02 Data error	7	6	12	95
F03 Peer conflict	1	8	27	84
F04 Priority conflict	9	41	29	41
F05 Policy gray zone	3	16	16	85
F06 New hire	7	8	23	82
F07 Schedule crisis	3	19	14	84
F08 Feedback	1	26	7	86
F09 Safety issue	6	4	21	89
F10 Short-staffed	7	10	17	86
F11 Info security	5	14	16	85
F12 Stalled decision	5	14	21	80
Band totals	62	173	225	980

Competency bands

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



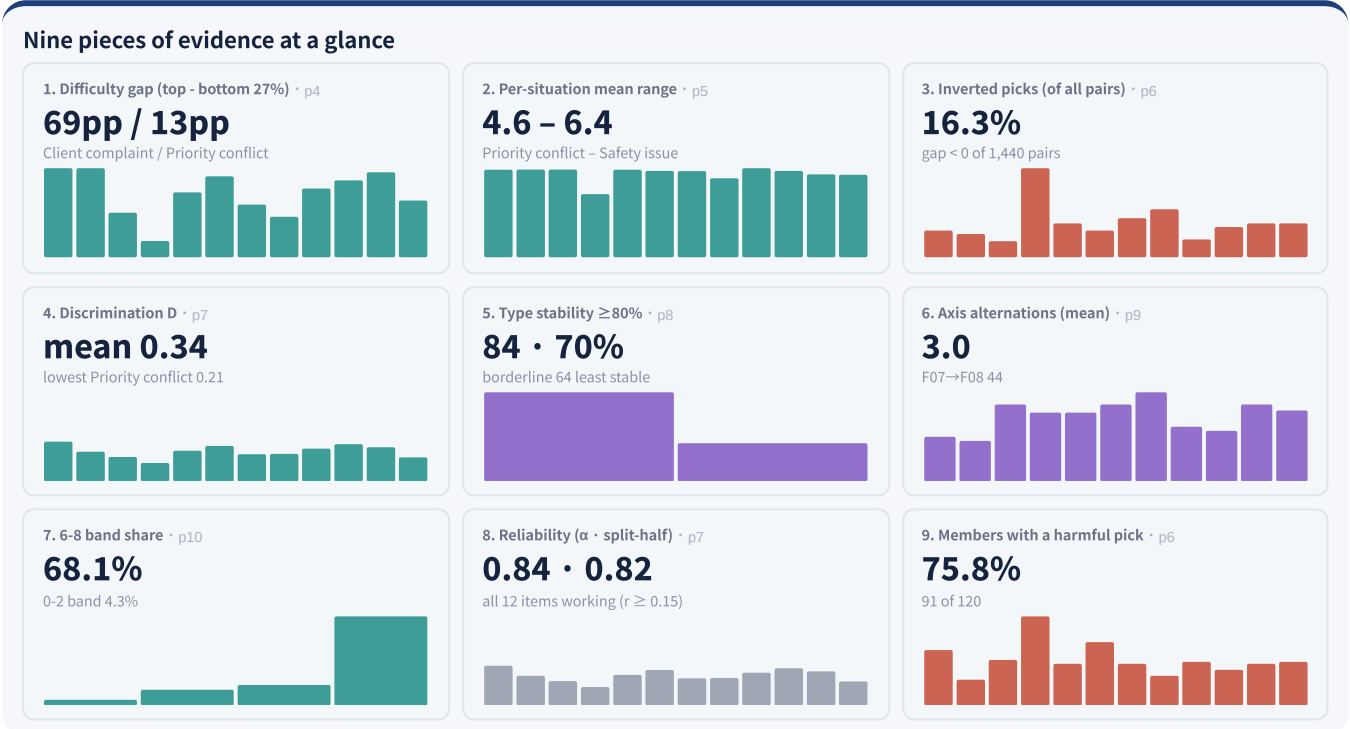
Figures

Situation	0-2	2-4	4-6	6-8	Top band	Share
F01 Client complaint	8	7	22	83	6-8	69.2%
F02 Data error	7	6	12	95	6-8	79.2%
F03 Peer conflict	1	8	27	84	6-8	70.0%
F04 Priority conflict	9	41	29	41	2-4 · 6-8	34.2%
F05 Policy gray zone	3	16	16	85	6-8	70.8%
F06 New hire	7	8	23	82	6-8	68.3%
F07 Schedule crisis	3	19	14	84	6-8	70.0%
F08 Feedback	1	26	7	86	6-8	71.7%
F09 Safety issue	6	4	21	89	6-8	74.2%
F10 Short-staffed	7	10	17	86	6-8	71.7%
F11 Info security	5	14	16	85	6-8	70.8%
F12 Stalled decision	5	14	21	80	6-8	66.7%

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 1440 pairs, 4.3% sit in the 0-2 band and 68.1% 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.

All 12 situations work (mean D 0.34), yet separation spans from Client complaint and Data error at 69pp down to Priority conflict at 13pp.



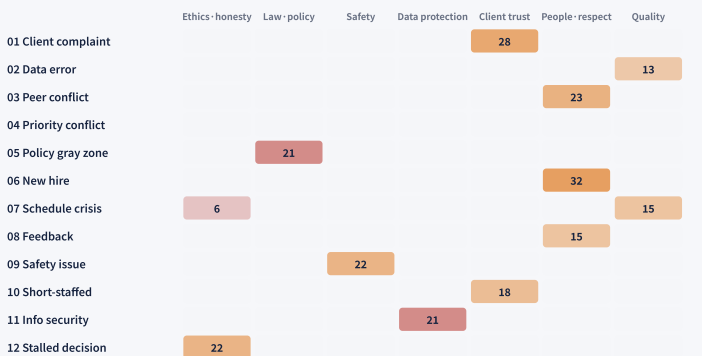
Synthesis in three sentences

① All 12 situations work (mean D 0.34, every item-rest r at or above 0.15), and 68% of responses sit in the 6-8 band - a high-scoring cohort. ② Separation varies widely by situation: Client complaint shows a 69pp top-bottom gap while Priority conflict manages only 13pp, leaving it nearly flat. ③ Types are stable for 70%, but the 64 borderline members can shift with the item mix, and the 76% harmful-pick rate with 16.3% inversions marks interview territory.

RISK CONCENTRATION

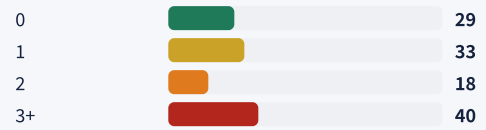
Harmful picks are not spread evenly. They pile up most in Priority conflict and Schedule crisis.

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



Number = people · color = highest severity in the cell (red Critical · orange High · yellow Caution) · darker = more · blank = no option in that area

Risky picks per person



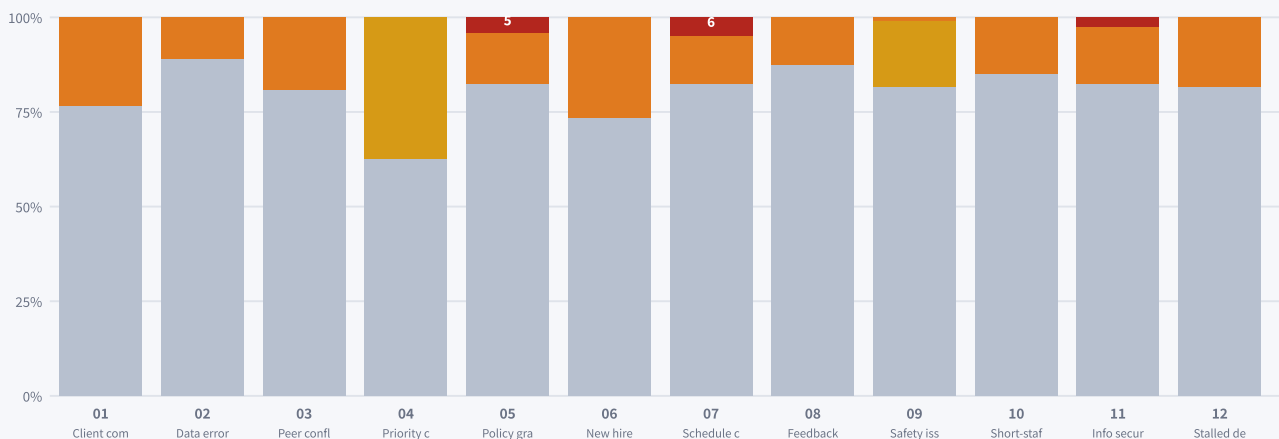
Repeated pattern (same area 2+) 24 · Critical at least once 12

Most attractive harmful options · top 6

④ Priority conflict	Do only the senior exec's task	key 2	Caution	38%
⑥ New hire	Take them off the work	key 2	High	27%
① Client complaint	Promise the timeline they asked for	key 2	High	23%
③ Peer conflict	Confront them in the meeting	key 2	High	19%
⑫ Stalled decision	Keep waiting for consensus	key 2	High	18%
⑨ Safety issue	Give only a verbal warning	key 2	Caution	18%

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 'Integrity & risk' and most spread out in 'Execution under pressure'.
Use it to set the onboarding focus.

Best credit by competency

Customer & stakeholders

median 3.00 · middle 50% 2.33–3.33

Integrity & risk

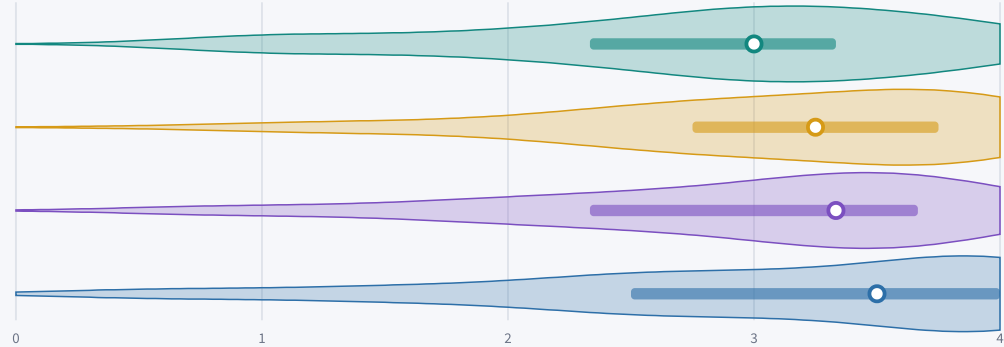
median 3.25 · middle 50% 2.75–3.75

People & collaboration

median 3.33 · middle 50% 2.33–3.67

Execution under pressure

median 3.50 · middle 50% 2.50–4.00



Worst credit by competency

Customer & stakeholders

median 2.67 · middle 50% 2.33–3.33

Integrity & risk

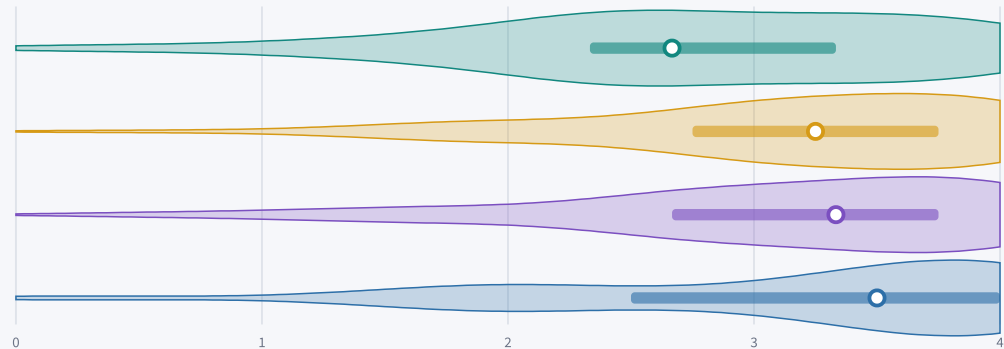
median 3.25 · middle 50% 2.75–3.75

People & collaboration

median 3.33 · middle 50% 2.67–3.75

Execution under pressure

median 3.50 · middle 50% 2.50–4.00



How to read

Shape = spread, dark bar = middle 50%, white dot = median. Competency tags come from the item bank and are auto-checked. With 2–4 situations per competency, use this for group trends rather than individual reads.

DEVELOPMENT · ONBOARDING PLAN

‘Execution under pressure’ has the widest score spread;
include it in differentiated onboarding.
Each suggestion is tied to a situation this cohort actually
missed.

Cohort spread by competency · Best + Worst credit

Customer & stakeholders

median 5.67 · middle 50% 5.00–6.67

Integrity & risk

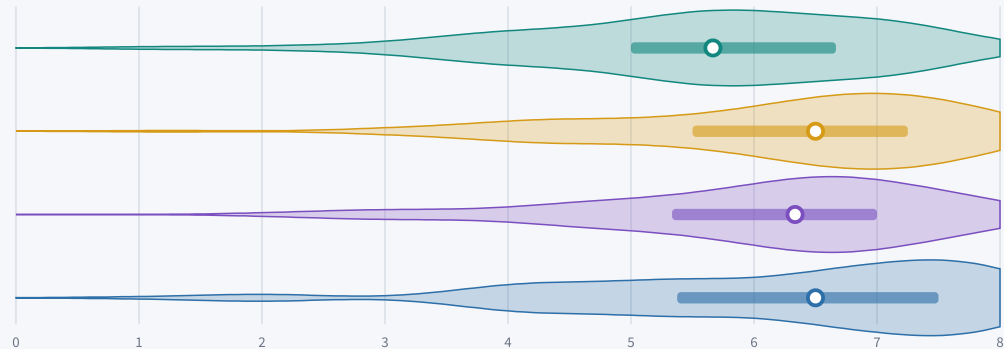
median 6.50 · middle 50% 5.50–7.25

People & collaboration

median 6.33 · middle 50% 5.33–7.00

Execution under pressure

median 6.50 · middle 50% 5.38–7.50



Customer & stakeholders · median 5.7/8

Evidence · most-missed situation Priority conflict (mean 4.6/8 · far-miss 40% · inverted 42%)

Who · 39 people (33%) at 5 or below in this competency

Week 1 · Priority conflict case workshop: model action “Align both execs on priorities together” 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)

People & collaboration · median 6.3/8

Evidence · most-missed situation Feedback (mean 5.7/8 · far-miss 25% · inverted 23%)

Who · 26 people (22%) at 5 or below in this competency

Week 1 · Feedback case workshop: model action “Sort out the criticism, discuss fixes 1:1” 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)

Integrity & risk · median 6.5/8

Evidence · most-missed situation Info security (mean 6.0/8 · far-miss 32% · inverted 16%)

Who · 22 people (18%) at 5 or below in this competency

Week 1 · Info security case workshop: model action “Share via an encrypted channel after security approval” 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 6.5/8

Evidence · most-missed situation Schedule crisis (mean 6.2/8 · far-miss 24% · inverted 18%)

Who · 30 people (25%) at 5 or below in this competency

Week 1 · Schedule crisis case workshop: model action “Tell the client first, offer a new timeline” 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. Competency tags come from the item bank and are auto-checked.

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.

④ Priority conflict · 28 people

“Two executives give you urgent tasks at once. How do you set priority?”

Follow-up: What if both say 'right now'?

Strong answer

Uses impact evidence to get both to agree

Concern

Goes by rank or promises both

IDs (ID order)

C-0127, C-0133, C-0157, C-0177, C-0213, C-0247, C-0328, C-0330, C-0365 +19 more

⑦ Schedule crisis · 22 people · Critical 6

“A core test fails just before launch. Who do you tell first?”

Follow-up: What if you are pushed to hold the date?

Strong answer

Tells clients and stakeholders openly; adjusts scope

Concern

Hides results or 'ship now, patch later'

IDs (ID order)

C-0110, C-0127, C-0128, C-0140, C-0182, C-0259, C-0286, C-0410, C-0476 +13 more

⑤ Policy gray zone · 19 people · Critical 5

“Have you been asked to skip an approval step to hit a target?”

Follow-up: Does it change if the amount is very small?

Strong answer

Keeps the process and looks for a fast-track

Concern

'Small amount, apply first' reasoning

IDs (ID order)

C-0128, C-0133, C-0139, C-0177, C-0213, C-0247, C-0259, C-0283, C-0286 +10 more

⑧ Feedback · 20 people

“How did you react when a manager criticized your report in public?”

Follow-up: Was there a time pushing back on the spot was right?

Strong answer

Stays calm, follows up 1:1 (ambiguous item: calm on-the-spot explanation is also acceptable)

Concern

Blames others or reacts emotionally

IDs (ID order)

C-0110, C-0128, C-0213, C-0247, C-0255, C-0362, C-0365, C-0373, C-0417 +11 more

⑩ Short-staffed · 14 people

“How did you rebalance work when a teammate suddenly left?”

Follow-up: When do you tell people you may miss a deadline?

Strong answer

Redistributes and tells stakeholders early

Concern

Takes it all alone or stays silent

IDs (ID order)

C-0140, C-0177, C-0200, C-0225, C-0259, C-0286, C-0410, C-0563, C-0664 +5 more

Top 5 scenarios by risky picks (fatal/high) · generated the same way as page 14 (Interview · development guide) of each individual report · questions for all 12 scenarios are in the HTML version and audit appendix · the interviewer makes the final choice.

Item score bands by situation · 120 people

01 Client complaint
02 Data error
03 Peer conflict
04 Priority conflict
05 Policy gray zone
06 New hire
07 Schedule crisis
08 Feedback
09 Safety issue
10 Short-staffed
11 Info security
12 Stalled decision



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.

INTERVIEWER GUIDE IN DEPTH

The 4 situations with the most risky picks, and the interview focus for each type: who to ask what, on one page.

Interview focus by type

Type · people	Interview focus	Sign of a strong answer
53 Balanced	Confirm strengths · reasoning behind the hardest decision	Can they explain with a real case why their judgment holds
1 Strong on effective action, risk identification to develop	Harm spotting · how they recognise what not to do	Do they name the most harmful option first
2 Insufficient evidence on both axes	Both axes · start by checking how they read the situation	Do they identify the key stakeholders
64 Borderline	Interval touches the cut · focus on the risky-pick situations	Does the judgment change when they re-explain the situation

04 Priority conflict · 28 people

Ask · “In a situation like Priority conflict, what did you do first, and why?”

Strong · in the direction of the model action: “Align both execs on priorities together”

Weak · defends the most common risky pick: “Do only the senior exec’s task”

Who · C-0127, C-0133, C-0157, C-0177, C-0213, C-0247 ...

07 Schedule crisis · 22 people

Ask · “In a situation like Schedule crisis, what did you do first, and why?”

Strong · in the direction of the model action: “Tell the client first, offer a new...”

Weak · defends the most common risky pick: “Ship now, patch later”

Who · C-0110, C-0127, C-0128, C-0140, C-0182, C-0259 ...

05 Policy gray zone · 19 people

Ask · “In a situation like Policy gray zone, what did you do first, and why?”

Strong · in the direction of the model action: “Follow the approval step, share the...”

Weak · defends the most common risky pick: “Apply first since it is a small amount”

Who · C-0128, C-0133, C-0139, C-0177, C-0213, C-0247 ...

08 Feedback · 20 people

Ask · “In a situation like Feedback, what did you do first, and why?”

Strong · in the direction of the model action: “Sort out the criticism, discuss fixes...”

Weak · defends the most common risky pick: “Raise it with HR”

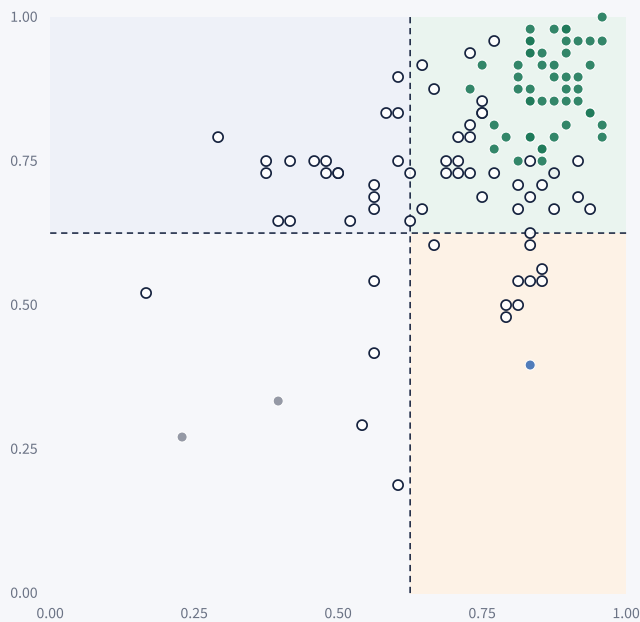
Who · C-0110, C-0128, C-0213, C-0247, C-0255, C-0362 ...

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.

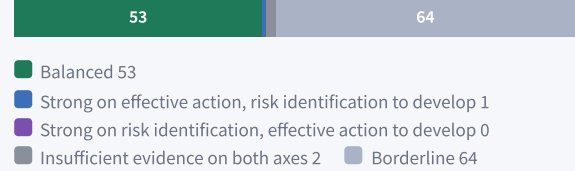
JUDGMENT LANDSCAPE · INTERVAL STRUCTURE

The 64 borderline people reflect too few items, not a skill gap.
Half would resolve with about 50 items.

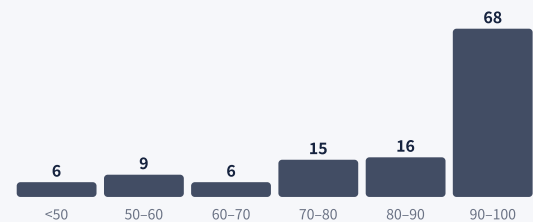


Color = type · hollow = borderline (interval crosses a cut) · one pseudonymous person per dot · not a ranking

Type distribution

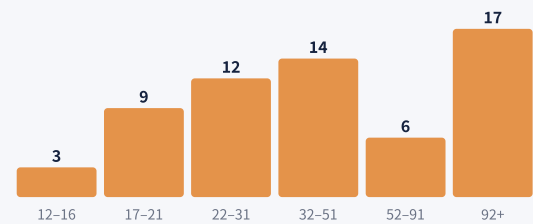


Type stability (1,000 resamples)



People by % of resamples staying in the same zone · 80%+: 84

Borderline · items needed to set a type (estimate)



Items needed $\approx 12 \times (\text{interval half-width} \div \text{distance to cut})^2$ · median 43 items · resolve with extra items or interview checks

Balanced

53

mean effective-action 0.87 · mean harm-ID 0.88

Strong on effective action, risk identification to develop

1

mean effective-action 0.83 · mean harm-ID 0.40

Insufficient evidence on both axes

2

mean effective-action 0.31 · mean harm-ID 0.30

Borderline

64

mean effective-action 0.67 · mean harm-ID 0.69

CHOICE DISTRIBUTIONS · 12 SITUATIONS × 5 OPTIONS

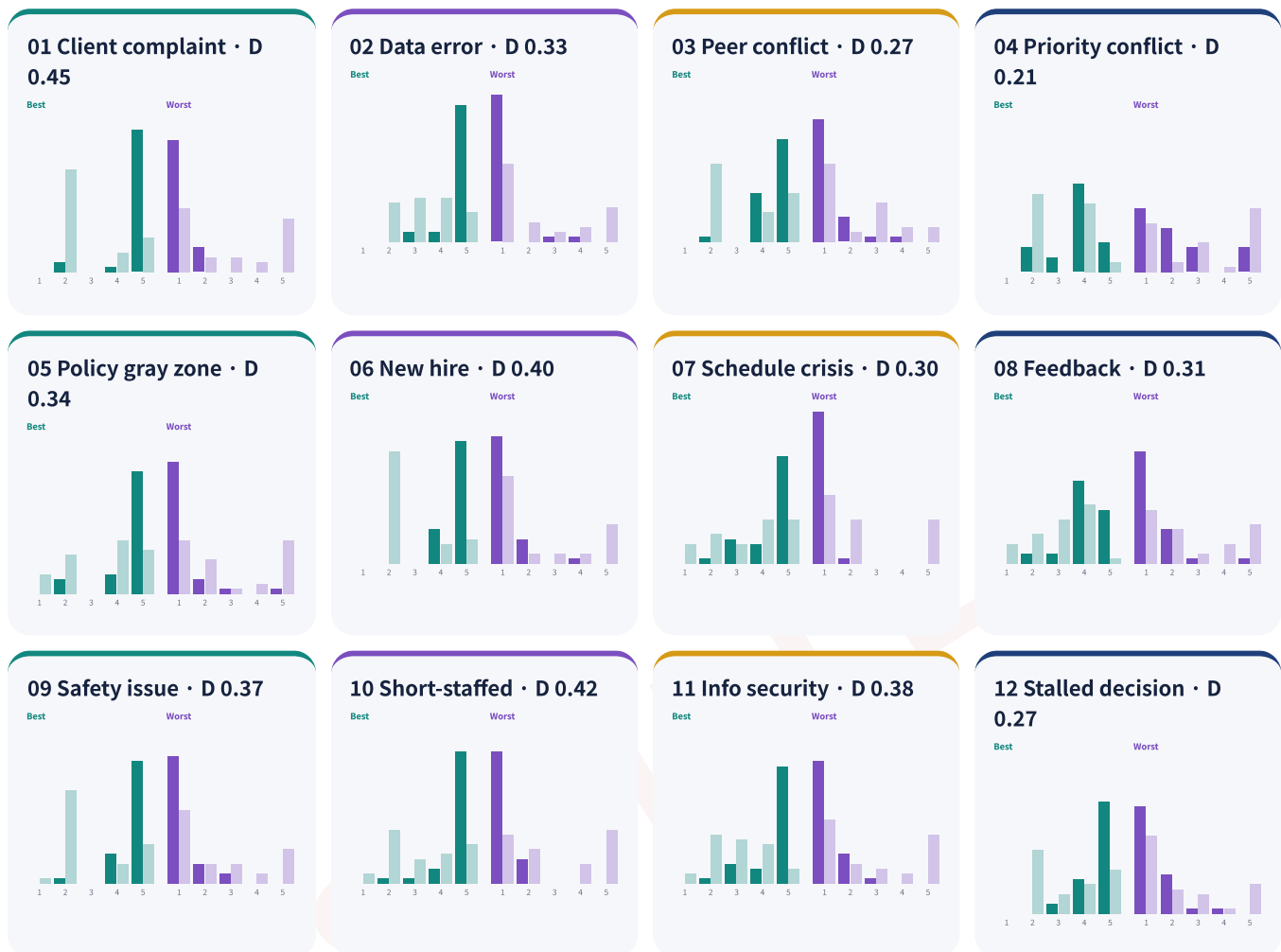
In most situations the model action is the top Best pick and the most harmful action the top Worst pick. The exceptions are covered on the item-health page.



Options in key order (key 5 = model ... key 1 = most harmful), with the share of 120 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 all 12 situations the top group picked the model action as Best and the most harmful action as Worst more often. The options work as intended.



How to read, and what shows

- Bars = pick share by key 1...5 · solid = top 27% (32 people) · light = bottom 27% (32).
- 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 2, 02 · key 3, 02 · key 4, 03 · key 2, 04 · key 2, 05 · key 2, 05 · key 4, 06 · key 2, 07 · key 2, 07 · key 4, 08 · key 2, 08 · key 3, 09 · key 2, 10 · key 2, 10 · key 3 + 4

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 Client complaint	59%	63%	0.64 [0.52, 0.73]	key 2 · 23%	11%		Keep
02 Data error	52%	61%	0.52 [0.38, 0.64]	key 4 · 22%	10%		Keep
03 Peer conflict	55%	54%	0.56 [0.43, 0.67]	key 4 · 26%	3%	Competing option	Keep · check competing option wording
04 Priority conflict	15%	27%	0.17 [-0.01, 0.34]	key 4 · 45%	23%	Competing option Risk hotspot	Keep · auto-suggest in interview bank · check competing option wording
05 Policy gray zone	57%	61%	0.51 [0.36, 0.63]	key 4 · 24%	16%		Keep
06 New hire	52%	62%	0.57 [0.43, 0.68]	key 2 · 27%	8%	Competing option	Keep · check competing option wording
07 Schedule crisis	49%	67%	0.39 [0.23, 0.53]	key 4 · 22%	18%		Keep
08 Feedback	22%	51%	0.39 [0.23, 0.53]	key 4 · 50%	17%	Competing option	Keep · check competing option wording
09 Safety issue	60%	58%	0.62 [0.49, 0.72]	key 4 · 21%	7%		Keep
10 Short-staffed	59%	55%	0.66 [0.54, 0.75]	key 4 · 14%	12%		Keep
11 Info security	44%	53%	0.57 [0.43, 0.68]	key 4 · 22%	12%		Keep
12 Stalled decision	49%	48%	0.41 [0.25, 0.55]	key 4 · 21%	9%		Keep

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%+ · disagreement = items marked F2. Key changes apply only as a new key_version.

Mean discrimination

0.50

mean corrected item-total r, 12 items

Agreement with key (Best)

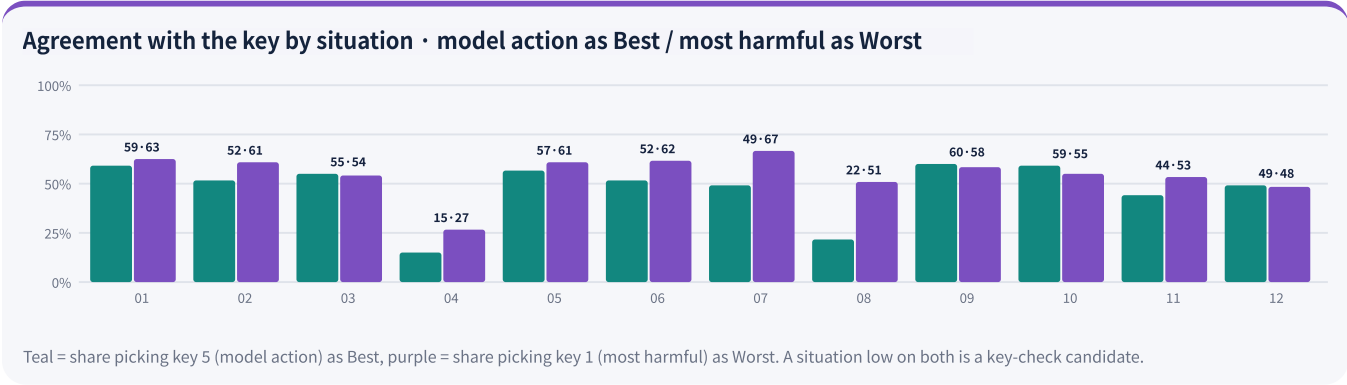
48%

share of all responses with key 5 as Best

Next-round actions

0

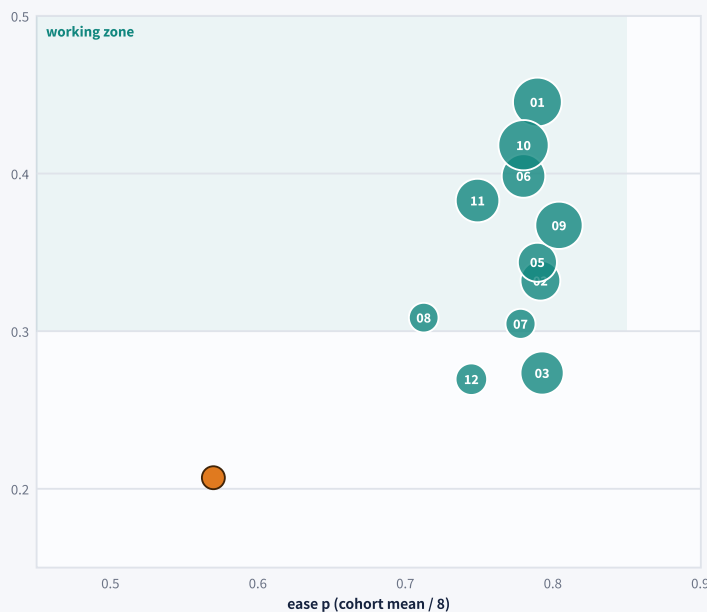
below the auto threshold ($r < 0.15$)
F04 weak on D-index — see next page



ITEM DISCRIMINATION STORY

12 of 12 items pass the automatic key screen (item-rest $r \geq 0.15$).
One item (F04) is weak on the D-index.

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

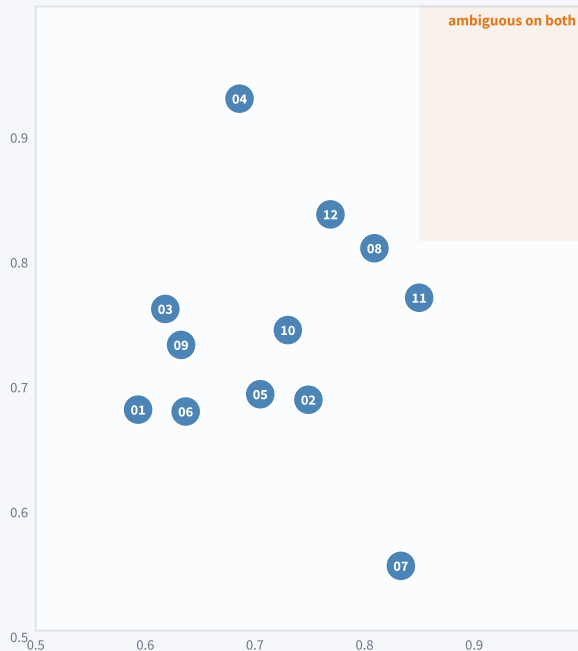
Situation	p	D	Item-rest r [95%]	Entropy B / W	Reading
01 Client complaint	0.79	0.45	0.64 [0.52, 0.73]	0.59 / 0.68	easier and stable → checks basic judgment
02 Data error	0.79	0.33	0.52 [0.38, 0.64]	0.75 / 0.69	easier and stable → checks basic judgment
03 Peer conflict	0.79	0.27	0.56 [0.43, 0.67]	0.62 / 0.76	easier and stable → checks basic judgment
04 Priority conflict	0.57	0.21	0.17 [-0.01, 0.34]	0.69 / 0.93	weak · small gap between top and bottom
05 Policy gray zone	0.79	0.34	0.51 [0.36, 0.63]	0.70 / 0.69	easier and stable → checks basic judgment
06 New hire	0.78	0.40	0.57 [0.43, 0.68]	0.64 / 0.68	easier and stable → checks basic judgment
07 Schedule crisis	0.78	0.30	0.39 [0.23, 0.53]	0.83 / 0.56	easier and stable → checks basic judgment
08 Feedback	0.71	0.31	0.39 [0.23, 0.53]	0.81 / 0.81	hard and separates well → core item
09 Safety issue	0.80	0.37	0.62 [0.49, 0.72]	0.63 / 0.73	easier and stable → checks basic judgment
10 Short-staffed	0.78	0.42	0.66 [0.54, 0.75]	0.73 / 0.75	easier and stable → checks basic judgment
11 Info security	0.75	0.38	0.57 [0.43, 0.68]	0.85 / 0.77	hard and separates well → core item
12 Stalled decision	0.74	0.27	0.41 [0.25, 0.55]	0.77 / 0.84	hard and separates well → core item

Readings are rule-based ($D < 0.25 \rightarrow$ weak, entropy $\geq 0.85 \rightarrow$ spread, $p < 0.76 \rightarrow$ hard). Replacing an item is a separate decision.

AMBIGUITY MAP · PICK ENTROPY

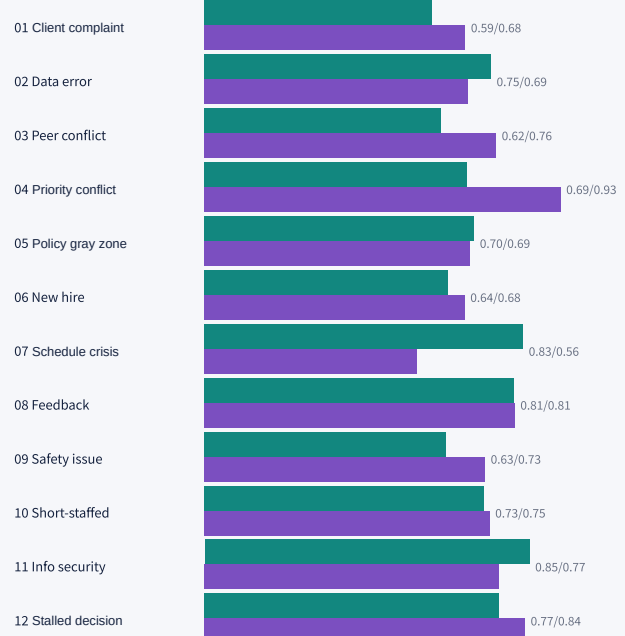
Picks spread widely in 1 situation.
The item check should tell whether this is people's judgment or item ambiguity.

Best × Worst entropy



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

Entropy by situation · Best / Worst



Key-check candidates · entropy 0.85+

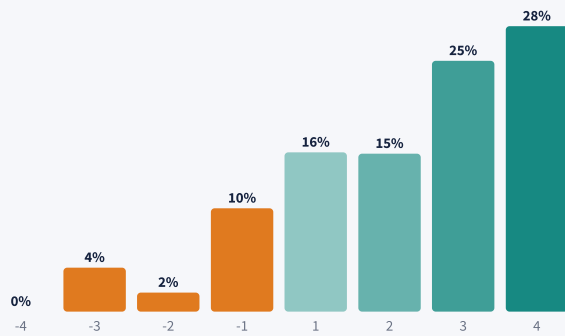
04 Priority conflict · Best 0.69 · Worst 0.93 · top Best key 4 (45%) · top Worst key 3 (28%)

High entropy can mean the item is ambiguous or several actions all look reasonable. Flagged by the automated key check. Keys are not changed from this data.

CHOICE QUALITY · GAP AND INVERSIONS

53% of responses have a gap of +3 or more.
Inversions are most frequent in Priority conflict (42%).

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

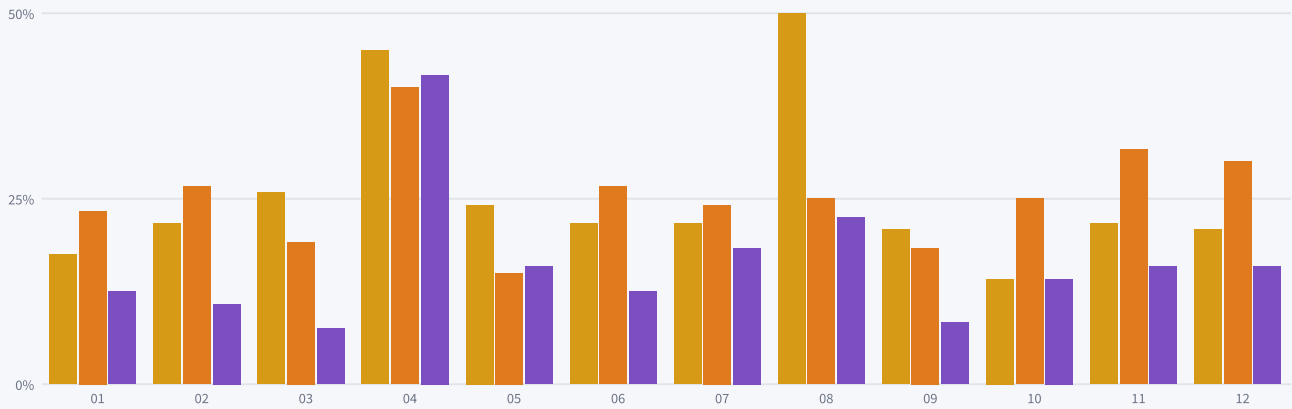


120 × 12 = 1440 responses. Orange = inverted (negative).

Gap summary

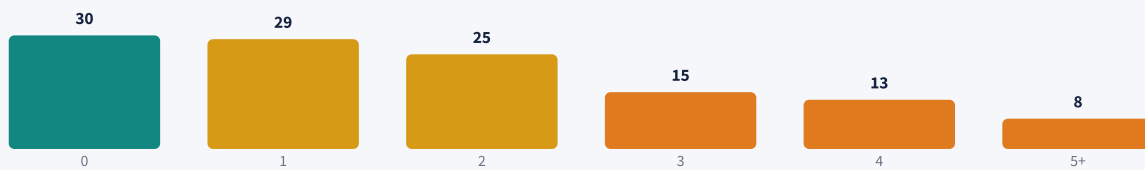
- +4 (perfect order): 28%
- +3 or more: 53%
- Inverted: 16% · 2.0 per person
- People with no inversion: 30
- 3+ inversions: 36 people → interview check

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



VALUE TENSION MAP

Among company priority values, 'Autonomy' shows the most tension: 17 people picked 25%+ of the reasonable options linked to it as Worst.

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	86	29	4	1	100%
Customer	70	38	10	2	100%
Team	57	60	3	0	83%
Results · speed	104	0	13	3	100%
Process	54	58	8	0	67%
Autonomy	103	0	0	17	0%

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

Autonomy

17

people reject 25%+

Results · speed

16

people reject 25%+

Customer

12

people reject 25%+

Process

8

people reject 25%+

Integrity

5

people reject 25%+

Team

3

people reject 25%+

How to read this, and limits

- The priority value with the most tension is 'Autonomy' (17 people). This is a point to explain in onboarding why the company values it, not a hiring criterion.
- Value tags are the G4 default draft and not yet validated. This page is not shown to clients.
- 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

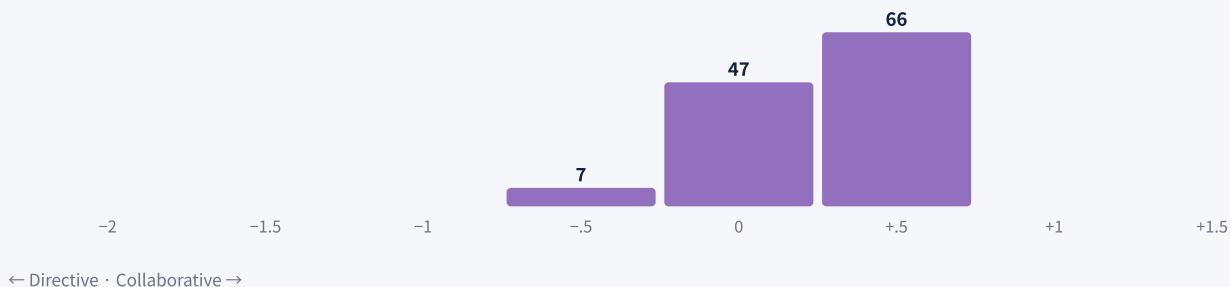
In week one, share one real decision that shows why autonomy comes first here and how it plays out.

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

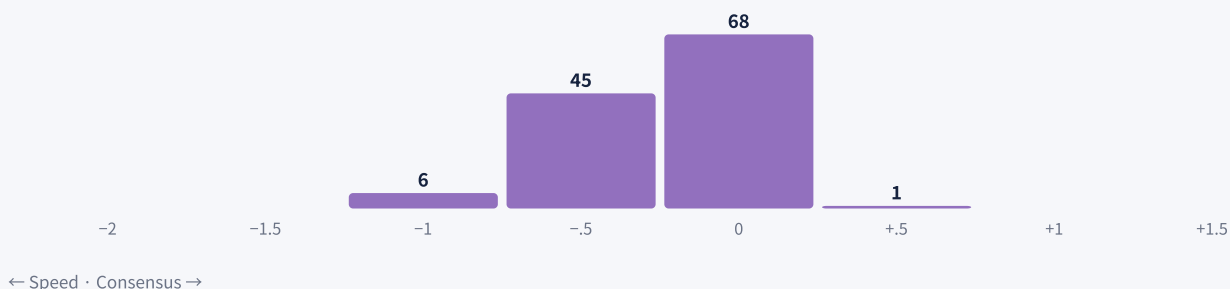
DECISION STYLE DISTRIBUTION

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

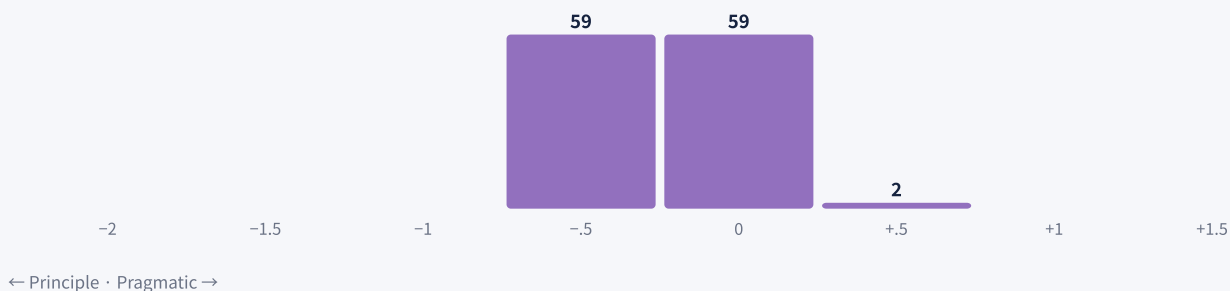
Directive ↔ Collaborative · median +0.50



Speed ↔ Consensus · median +0.00



Principle ↔ Pragmatic · median +0.00



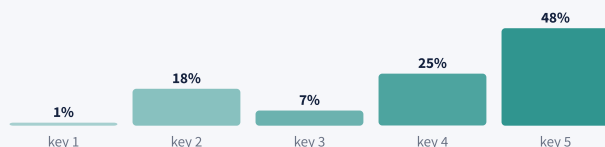
Style means by type

Type	Directive↔Collaborative	Speed↔Consensus	Principle↔Pragmatic	N
Balanced	+0.56	+0.08	-0.04	53
Strong on effective action, risk identification to develop	+0.08	+0.42	-0.33	1
Insufficient evidence on both axes	-0.33	-0.38	+0.13	2

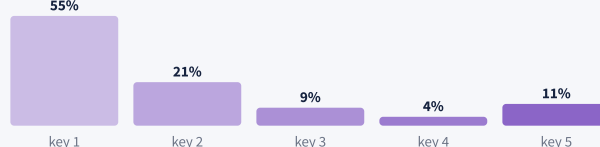
Score -2 to +2 = the approach in options picked as Best vs Worst · style is a difference in approach, not better or worse · use it to see team-mix diversity · faking resistance not yet validated. · borderline members (64) are excluded from type means (type not set).

Where Best and Worst picks land on the key · all 12 situations

key of Best picks



key of Worst picks



120 people × 12 situations = 1440 responses. The more Best picks sit near key 5 and Worst picks near key 1, the more the cohort judges in the same direction as the the scoring key.

MEASUREMENT CONSISTENCY · TYPE STABILITY

Split-half 0.82, α 0.84.
70% of people keep the same type under resampling.

Split-half (odd/even, S-B)

0.82

90% 0.75–0.87

Cronbach α **0.84**

90% 0.78–0.87

Type stable 80%+

70%

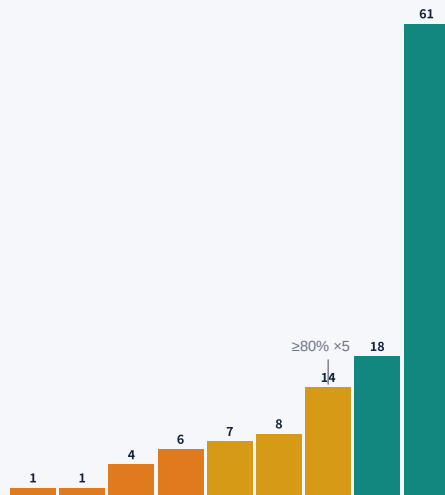
84 people

Shown as borderline

64

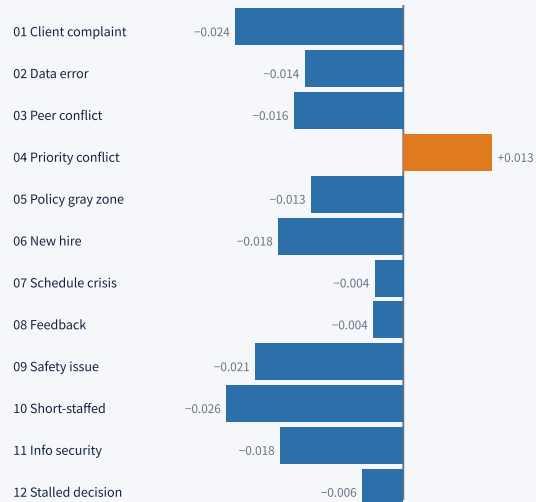
interval touches the cut

Type stability per person



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

84 people at 80%+ · 5 of them fall in the boundary bin (60–80% color).

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	53				31
Strong on effective action, risk identification to develop		1			9
Strong on risk identification, effective action to develop					19
Insufficient evidence on both axes				2	5

If the interval touches the cut, the report shows 'Borderline' whatever the point estimate. This guards against overconfidence.

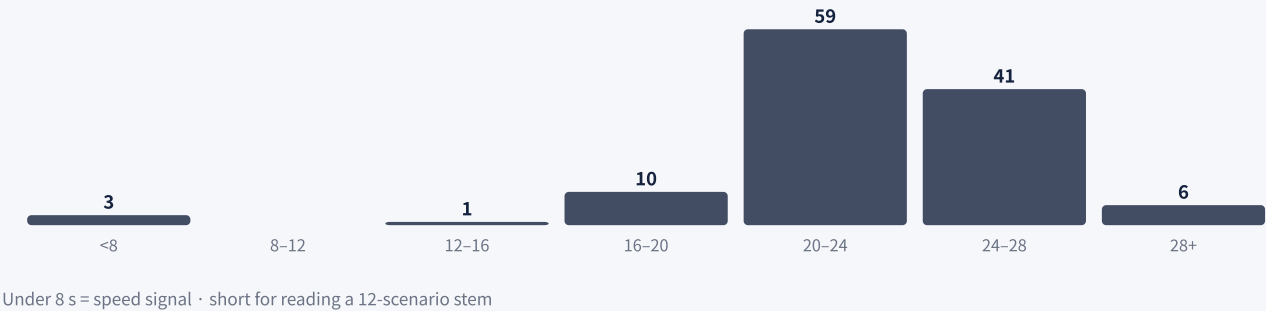
What these figures mean, and do not

- Descriptive values from 120-person synthetic data, not a real reliability or validity result.
- With real responses, the same method is rerun and published with a version.
- Based on 12 items; the values change if the item count changes.

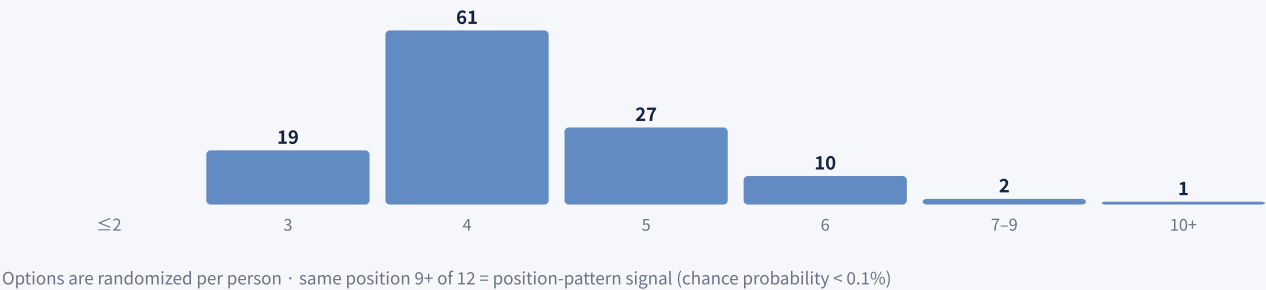
RESPONSE INTEGRITY OVERVIEW

4 of 120 members carry a review signal.
The rest are within normal range on both speed and position pattern.

Median response time per person (s)



Max times Best was picked at the same screen position

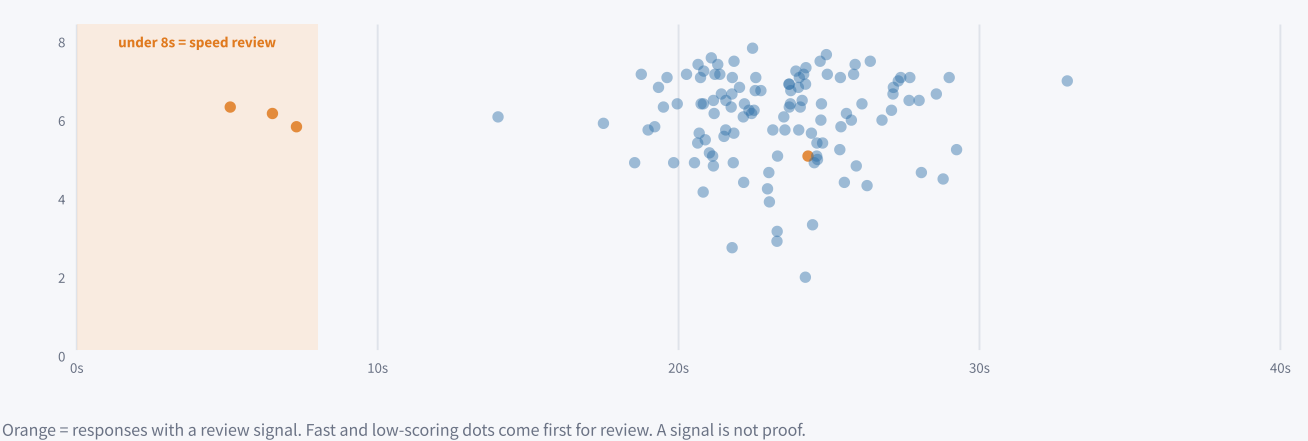


Responses with signals

ID	Signal	Median time	Position max	Type	Suggested
C-0213	Position pattern	24.3s	12/12	Borderline	Human review first · retest if needed
C-0386	Speed	6.5s	4/12	Borderline	Human review first · retest if needed
C-0489	Speed	7.3s	4/12	Borderline	Human review first · retest if needed
C-0972	Speed	5.1s	5/12	Balanced	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.

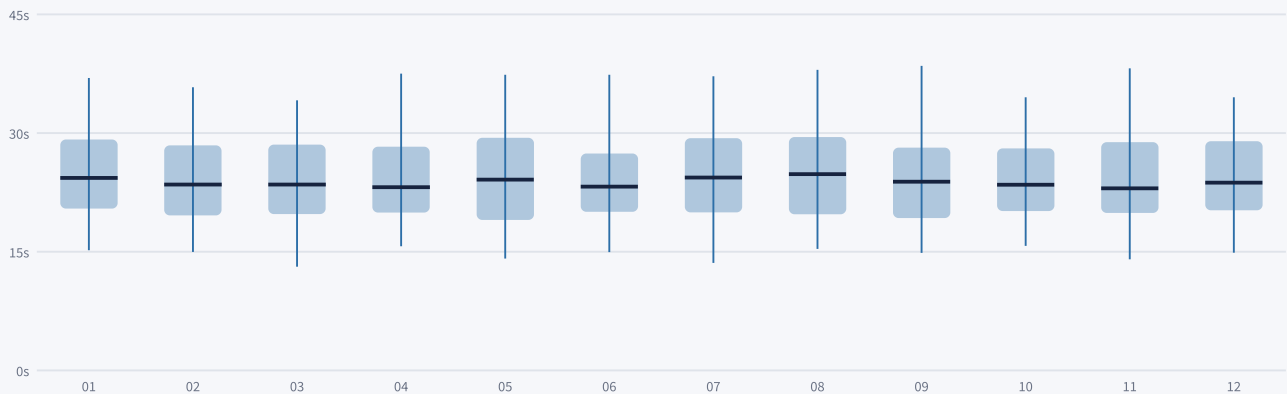
Response time × mean item score · 120 people



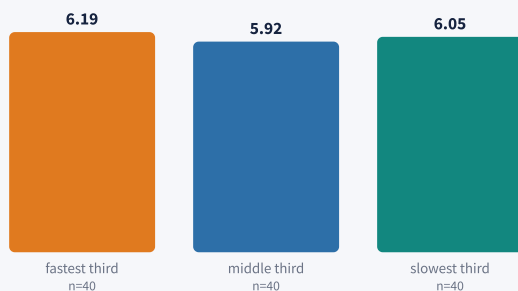
RESPONSE TIME IN DEPTH

The judgment gap between the fastest and slowest responders is only 0.14.
Speed is not a screening criterion.

Response time by item · 120 people

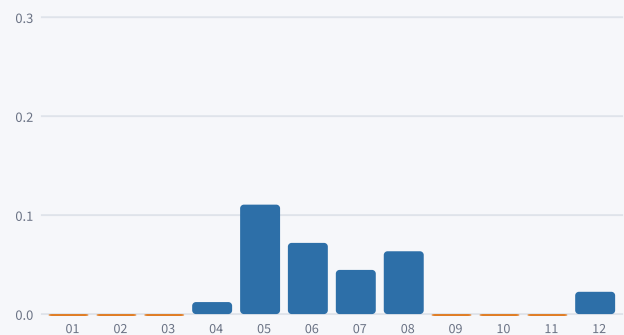


Mean item score by speed group



Split into thirds by median response time.

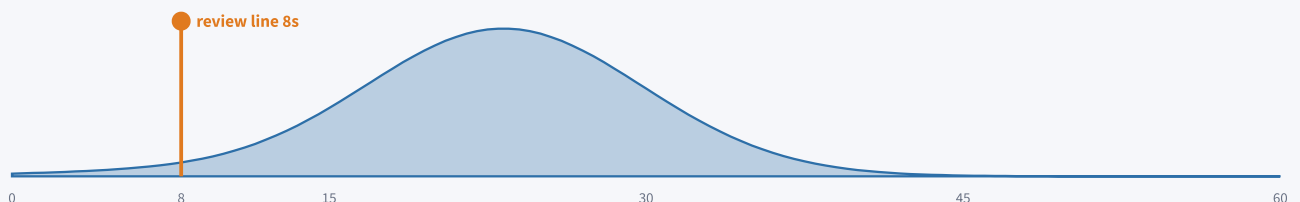
Time-score correlation by item



What can be said about speed

- Fastest third mean 6.19, slowest third 6.05. The difference is small, 0.14.
- Median below 8s (review signal): 3 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 120 people on one page.

ID	Eff.	Harm	Type	Stab.	Risky	Sig.	Interview focus	ID	Eff.	Harm	Type	Stab.	Risky	Sig.	Interview focus
C-0110	0.83	0.75	Borderline	91%	0			C-0569	0.96	1.00	Balanced	100%	0		
C-0119	0.92	0.85	Balanced	100%	0			C-0573	0.81	0.67	Borderline	71%	2		④ ⑫
C-0127	0.56	0.54	Borderline	55%	5		① ④ ⑤ ⑥ ⑪	C-0575	0.60	0.90	Borderline	54%	5●		① ③ ④ ⑦ ⑪
C-0128	0.50	0.73	Borderline	80%	5●		① ③ ⑥ ⑧ ⑨	C-0583	0.92	0.69	Borderline	76%	1		④
C-0133	0.79	0.48	Borderline	89%	2		① ⑥	C-0597	0.58	0.83	Borderline	61%	5		④ ⑥ ⑦ ⑪ ⑫
C-0135	0.83	0.69	Borderline	79%	1		④	C-0616	0.88	0.79	Balanced	99%	1		⑫
C-0139	0.56	0.69	Borderline	49%	5		① ④ ⑤ ⑥ ⑫	C-0618	0.69	0.75	Borderline	75%	3		① ② ⑥
C-0140	0.40	0.33	Low evid.	98%	8		① ② ④ ⑤ ⑥ ⑨ ⑩ ⑪	C-0627	0.90	0.98	Balanced	100%	0		
C-0157	0.96	0.81	Balanced	98%	0			C-0631	0.77	0.77	Balanced	97%	2		④ ⑥
C-0162	0.71	0.73	Borderline	80%	3		① ② ④	C-0634	0.56	0.42	Borderline	62%	6●		② ⑥ ⑧ ⑩ ⑪ ⑫
C-0165	0.85	0.94	Balanced	100%	1		④	C-0641	0.83	0.40	Action+	99%	1		⑨
C-0177	0.42	0.65	Borderline	62%	8		① ② ③ ④ ⑤ ⑦ ⑪ ⑫	C-0652	0.90	0.85	Balanced	99%	0		
C-0182	0.69	0.73	Borderline	69%	3		② ③ ⑤	C-0654	0.75	0.85	Borderline	91%	3		③ ⑥ ⑦
C-0200	0.83	0.54	Borderline	79%	1		④	C-0664	0.60	0.19	Borderline	55%	5		① ⑥ ⑧ ⑨ ⑫
C-0213	0.56	0.71	Borderline	60%	4●	■	① ⑧ ⑨ ⑫	C-0665	0.96	0.79	Balanced	100%	0		
C-0214	0.88	0.92	Balanced	100%	1		①	C-0675	0.52	0.65	Borderline	58%	4		① ④ ⑥ ⑨
C-0221	0.83	0.94	Balanced	99%	1		⑤	C-0679	0.83	0.85	Balanced	99%	2		④ ⑥
C-0225	0.38	0.75	Borderline	82%	8		① ② ③ ④ ⑤ ⑦ ⑨ ⑫	C-0696	0.67	0.60	Borderline	30%	4●		③ ④ ⑩ ⑪
C-0247	0.54	0.29	Borderline	78%	6		③ ⑤ ⑥ ⑦ ⑨ ⑫	C-0704	0.71	0.79	Borderline	89%	3		④ ⑥ ⑫
C-0255	0.85	0.56	Borderline	72%	0			C-0705	0.94	0.67	Borderline	74%	0		
C-0259	0.42	0.75	Borderline	87%	8●		① ② ③ ⑥ ⑧ ⑨ ⑩ ⑪	C-0738	0.81	0.54	Borderline	78%	1		⑨
C-0269	0.88	0.73	Balanced	99%	1		④	C-0753	0.90	0.94	Balanced	100%	1		⑤
C-0275	0.75	0.83	Borderline	92%	3		④ ⑩ ⑫	C-0754	0.46	0.75	Borderline	86%	6		① ③ ④ ⑨ ⑩ ⑫
C-0283	0.83	0.79	Balanced	98%	2●		⑤ ⑧	C-0760	0.83	0.60	Borderline	53%	1		⑥
C-0286	0.29	0.79	Borderline	95%	9●		① ③ ④ ⑤ ⑥ ⑦ ⑨ ⑩ ⑫	C-0764	0.83	0.79	Balanced	98%	2		④ ⑩
C-0293	0.73	0.88	Balanced	100%	0			C-0779	0.77	0.73	Borderline	84%	2		④ ⑦
C-0328	0.83	0.88	Balanced	99%	1		⑧	C-0786	0.81	0.71	Borderline	91%	3		① ④ ⑦
C-0330	0.92	0.88	Balanced	99%	1		④	C-0788	0.77	0.96	Borderline	98%	2		④ ⑨
C-0342	0.83	0.96	Balanced	99%	1		④	C-0793	0.85	0.75	Balanced	100%	0		
C-0362	0.90	0.88	Balanced	100%	0			C-0795	0.90	0.81	Balanced	98%	1		⑫
C-0365	0.79	0.50	Borderline	84%	2		① ④	C-0797	0.60	0.83	Borderline	56%	4		① ④ ⑩ ⑫
C-0370	0.90	0.96	Balanced	100%	1		③	C-0809	0.83	0.96	Balanced	100%	0		
C-0373	0.85	0.71	Borderline	89%	1		⑦	C-0826	0.23	0.27	Low evid.	100%	11●		① ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪ ⑫
C-0386	0.67	0.88	Borderline	79%	2	■	④ ⑤	C-0844	0.92	0.96	Balanced	100%	1		⑧
C-0393	0.90	0.90	Balanced	100%	0			C-0845	0.88	0.90	Balanced	100%	0		
C-0408	0.83	0.98	Balanced	100%	2		⑤ ⑧	C-0858	0.65	0.92	Borderline	67%	2		⑥ ⑩
C-0410	0.38	0.73	Borderline	84%	9		① ② ④ ⑥ ⑦ ⑨ ⑩ ⑪ ⑫	C-0863	0.85	0.85	Balanced	100%	1		⑪
C-0417	0.73	0.79	Borderline	89%	1●		⑪	C-0869	0.48	0.75	Borderline	84%	6		① ③ ④ ⑥ ⑦ ⑨
C-0425	0.81	0.88	Balanced	99%	2		③ ⑧	C-0871	0.73	0.73	Borderline	80%	4		② ③ ④ ⑦
C-0455	0.92	0.75	Borderline	93%	0			C-0873	0.83	0.96	Balanced	100%	1		⑪
C-0458	0.88	0.67	Borderline	70%	1		⑫	C-0876	0.81	0.92	Balanced	100%	2		⑦ ⑩
C-0469	0.94	0.83	Balanced	98%	0			C-0877	0.63	0.73	Borderline	47%	4●		③ ⑥ ⑧ ⑨
C-0472	0.94	0.83	Balanced	99%	0			C-0884	0.94	0.96	Balanced	100%	1		④
C-0474	0.83	0.94	Balanced	100%	1		①	C-0885	0.83	0.63	Borderline	61%	1		⑫
C-0476	0.75	0.92	Balanced	100%	0			C-0895	0.40	0.65	Borderline	59%	7●		③ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪
C-0482	0.81	0.90	Balanced	100%	1		⑪	C-0907	0.90	0.98	Balanced	100%	0		
C-0484	0.56	0.67	Borderline	49%	4		③ ⑤ ⑥ ⑫	C-0918	0.73	0.81	Borderline	91%	2		④ ⑦
C-0489	0.71	0.75	Borderline	82%	2	■	④ ⑥	C-0928	82%	0.65	Borderline	36%	4		④ ⑤ ⑨ ⑪
C-0490	0.92	0.90	Balanced	100%	0			C-0929	0.88	0.85	Balanced	100%	0		
C-0534	0.85	0.92	Balanced	100%	0			C-0931	0.85	0.77	Balanced	100%	1		⑤
C-0540	0.88	0.73	Borderline	93%	1		④	C-0934	0.96	0.96	Balanced	100%	0		
C-0545	0.75	0.69	Borderline	75%	3		② ⑦ ⑨	C-0941	0.81	0.75	Balanced	100%	0		
C-0548	0.83	0.85	Balanced	99%	2		⑤ ⑪	C-0942	0.83	0.79	Balanced	97%	1		⑪
C-0551	0.48	0.73	Borderline	81%	5●		① ④ ⑤ ⑥ ⑫	C-0950	0.65	0.67	Borderline	44%	4		③ ⑥ ⑧ ⑪
C-0553	0.73	0.94	Borderline	89%	3		④ ⑤ ⑪	C-0972	0.77	0.81	Balanced	99%	0	■	
C-0555	0.79	0.79	Balanced	96%	1●		⑦	C-0973	0.50	0.73	Borderline	81%	6●		① ② ⑥ ⑦ ⑩ ⑫
C-0561	0.60	0.75	Borderline	50%	4		① ③ ⑤ ⑨	C-0982	0.88	0.98	Balanced	100%	0		
C-0563	0.17	0.52	Borderline	78%	12●		① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪ ⑫	C-0984	0.85	0.54	Borderline	91%	1		④
C-0564	0.75	0.83	Borderline	95%	2		③ ④	C-0985	0.94	0.92	Balanced	100%	0		
C-0566	0.81	0.50	Borderline	91%	1		⑥	C-0992	0.85	0.77	Balanced	98%	0		

Eff./Harm = 0–1 axes (cut 0.625) · Stab. = same-type share over 1,000 resamples · Risky = bottom-key picks, ● = Critical risky pick · ■ = integrity review signal · interview focus = scenarios with risky picks · type shows Borderline when the interval touches the cut.
Stab. is rounded to a whole percent — the 80%+ stability metric (84 members) uses unrounded values. Risky picks = Best on the two lowest-ranked key options (ranks 1–2).

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=120, Standard tier (green row).

Definitions

Two-axis scores	$x = \text{mean}(\text{key}(B)-1)/4$, $y = \text{mean}(5-\text{key}(W))/4$. Descriptive before calibration, not θ .
Intervals · borderline	90% interval over 1,000 item resamples. If it crosses the cut (0.625, fixed by in_key_version), no type is set.
Risky picks	E1 = harmful option (severity 1–3) as Best · E2 = model action (key 5) as Worst. key-conflict items are lowered one level.
Discrimination	Corrected item-total correlation, Fisher z 95% CI.
Integrity	Median time < 8 s, Best at the same screen position ≥ 9/12. Never combined with scores.
Versions	Scoring key 2026-09-23 · 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

Scoring key 2026-09-23

Reference cut 0.625

No overall rank · no composite

Missing stays missing

Synthetic sample

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.
- Keys (bw-s1-v1) pass an automated quality gate.

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.

Illustrative cohort

120

same synthetic assessment

Competency lenses

4

Best + Worst credits

Provisional floor

50

below 50: report blocked

Situations

12

5 options, Best + Worst

Four distributions · Best + Worst credit / 8

Descriptive

Customer & stakeholders

5.00–6.67 · 5.67

Integrity & risk

5.50–7.25 · 6.50

People & collaboration

5.33–7.00 · 6.33

Execution under pressure

5.38–7.50 · 6.50

0 2 4 6 8

Band = middle half; line = median; All 120 synthetic records completed the 12 situations. Competency tags are auto-checked.

Publication gate

Provisional cohort

120 / 120 eligible

 $n \geq 50$ required

provisional floor 50

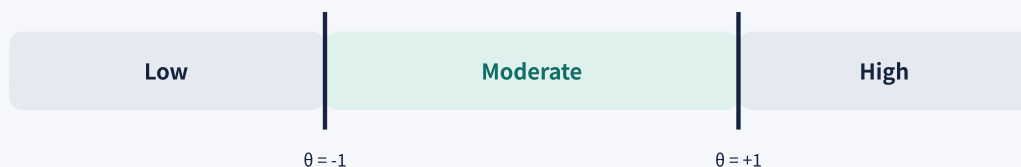
The provisional-synthetic report is blocked below $n=50$. The 20–49 Small tier needs a separately approved production-frozen asset.

Classification, in scope

Existing Best/Worst type labels describe two 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 $\theta=+0.42$ is separate from this sample.

Boundary before interpretation

No θ , frozen calibration or approved reference version exists for this Scoring key 2026-09-23 sample, so no actual Ability Band or Ability Percentile is available. Source thresholds: Low $\theta < -1$, Moderate $-1 \leq \theta \leq 1$, High $\theta > 1$. Only a frozen item calibration without local refit allows a percentile. A separate hypothetical $\theta=+0.42$ 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.

Quartiles and ranks · the cohort comparison baseline.

Axis score quartiles · exact from the 120-member roster

Metric	P10	P25	P50 (median)	P75	P90
Effective-action ID	0.50	0.67	0.81	0.88	0.92
Risk ID	0.54	0.69	0.79	0.88	0.96
Axis mean	0.58	0.68	0.79	0.87	0.90

Cut 0.625 sits between P10 and P25 - see the p3 legend for band edges

Competency quartiles · from the p10 band summary

Competency	P25	Median	P75
Customer & stakeholders	5.0	5.7	6.7
Integrity & risk	5.5	6.5	7.3
People & collaboration	5.3	6.3	7.0
Execution under pressure	5.4	6.5	7.5

Situation ranks · by mean, ties share a rank

Rank	Situation	Mean	SD	Median	Q1–Q3	6–8 band
1	F09 Safety issue	6.4	1.8	7.0	5.0–8.0	74.2%
T-2	F01 Client complaint	6.3	2.1	7.0	5.0–8.0	69.2%
T-2	F02 Data error	6.3	1.9	7.0	6.0–8.0	79.2%
T-2	F03 Peer conflict	6.3	1.6	7.0	5.0–8.0	70.0%
T-2	F05 Policy gray zone	6.3	1.9	7.0	5.0–8.0	70.8%
T-6	F06 New hire	6.2	2.0	7.0	5.0–8.0	68.3%
T-6	F07 Schedule crisis	6.2	2.0	7.0	5.0–8.0	70.0%
T-6	F10 Resource shortage	6.2	2.0	7.0	5.0–8.0	71.7%
T-9	F11 Info security	6.0	1.9	6.0	5.0–7.0	70.8%
T-9	F12 Decision delay	6.0	1.9	6.0	5.0–8.0	66.7%
11	F08 Feedback	5.7	1.8	6.0	5.0–7.0	71.7%
12	F04 Priority conflict	4.6	2.0	5.0	3.0–6.0	34.2%

Definitions · P10–P90 are linearly interpolated quantiles of all 120 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