

Theta AI™

Judgment Evidence Report

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

What this person judged most effective and most harmful across 12 work situations, where their judgment is solid or shaky by competency, and what to check in the interview.

Candidate C-0417 (pseudonymous)

12/12 answered · 29 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 | 14 Interview · development guide |
| 02 Visual dashboard · 30-second view | 15 Interviewer deep questions |
| 03 Overview dashboard | 16 Development 30/60/90 |
| 04 Difficulty & candidate scores | 17 Two-axis judgment map |
| 05 Candidate vs means | 18 Choice fingerprint 12×5 |
| 06 Candidate harmful & inversions | 19 Choice quality · gap |
| 07 Discrimination & candidate group | 20 Attractive distractors |
| 08 Candidate stability | 21 Difficulty-adjusted story |
| 09 Candidate alternations | 22 Consistency · type stability |
| 10 Candidate bands | 23 Response pattern |
| 11 Candidate nine-metric synthesis | 24 Response time in depth |
| 12 Competency profile | 25 Values fingerprint · decision style |
| 13 Risky picks · harm areas | 26 Measurement notes · evidence |

EXECUTIVE SUMMARY · DECISION PAGE

The one-minute conclusion before the interview.

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

The conclusion in one sentence

Mean 6.1/8, slightly above the cohort mean (6.05), with harm spotting (Worst credit 0.79) clearer than spotting effective action (0.73). Check Feedback, Info security, Policy gray zone first in the interview.

3 strengths · situations judged better than the cohort

- 04 Priority conflict · 7/8 · cohort 4.6 (+2.4)
- 06 New hire · 8/8 · cohort 6.2 (+1.8)
- 01 Client complaint · 8/8 · cohort 6.3 (+1.7)

3 points to check

- 08 Feedback · far below the cohort 2/8 · cohort 5.7
- 11 Info security · risky pick 1/8 · cohort 6.0
- 05 Policy gray zone · in line with the cohort 6/8 · cohort 6.3

For follow-up, not reasons to reject.

3 questions to ask in the interview

- 08 “In Feedback, why did ‘Sort out the criticism, discuss fixes 1:1’ look like the worst choice?”
- 11 “Have you acted like ‘Send the file first, report later’ in a situation like Info security? What did you weigh first?”
- 05 “In Policy gray zone, why did ‘Apply the discount without approval to hit the target’ look like the worst choice?”

Follow-ups and signs of strong and weak answers are on the ‘Interviewer deep questions’ page.

30-minute interview plan

- 0–5 min · open on the most solid competency, ‘Customer & stakeholders’
- 5–20 min · check the 1 key-risk pick with the deep-question page (reason → alternative → outcome)
- 20–27 min · ask for one real example in ‘Integrity & risk’
- 27–30 min · candidate questions · write down the reasoning for the decision meeting

Is the evidence enough?

- 12/12 answered · nothing missing
- Mean item score 90% interval **4.9–7.0** (out of 8)
- Type stability **89%** (1,000 item resamples)
- Response integrity: **no signal**
- Comparison: the 120-person cohort in this campaign · descriptive, pre-calibration (Scoring key 2026-09-23)

One line per competency · Best credit + Worst credit (0–8)

Customer & stakeholders · 7.0/8 · above the cohort median (5.7) · lowest situation 12 Stalled decision (6/8)

Integrity & risk · 5.5/8 · below the cohort median (6.5) · lowest situation 11 Info security (1/8)

People & collaboration · 5.7/8 · below the cohort median (6.3) · lowest situation 08 Feedback (2/8)

Execution under pressure · 6.5/8 · in line with the cohort median (6.5) · lowest situation 07 Schedule crisis (6/8)

Competency tags are a sample mapping · a comparison with the cohort median · relative standing in Appendix R

What this report says

- How well effective and harmful actions were spotted in 12 situations, with intervals
- Position against the cohort · relative standing in Appendix R
- Specific situations and questions to check in the interview

What it does not say

- Whether to hire · composite score · rank
- A calibrated ability score (before key and calibration approval)
- Personality · misconduct

Pages to read before deciding

- p19 Risky picks · harm areas
- p23 Interviewer deep questions · rubric
- p24 Development 30/60/90 days
- p26 Measurement notes · evidence

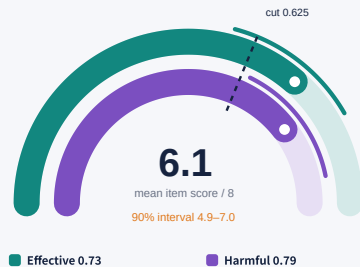
VISUAL DASHBOARD · 30-SECOND VIEW

Mean item score 6.1/8 across 12 situations.

Harm spotting is the clearer strength;

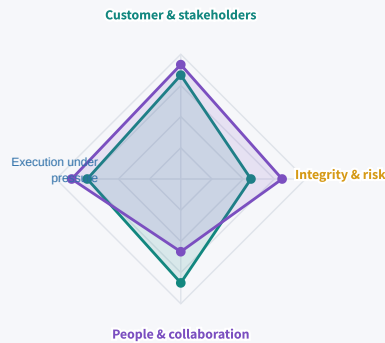
1 key-risk pick (critical) goes to the interview.

Two-sided judgment gauge



Outer = spotting effective action, inner = spotting harmful action (0-1). Thin arc = 90% interval. Descriptive before calibration, not θ .

Competency profile · Best vs Worst



Teal = Best credit, purple = Worst credit (mean 0-4 per competency) · dashed = cohort median · competency tags are a sample mapping.

Item score mix



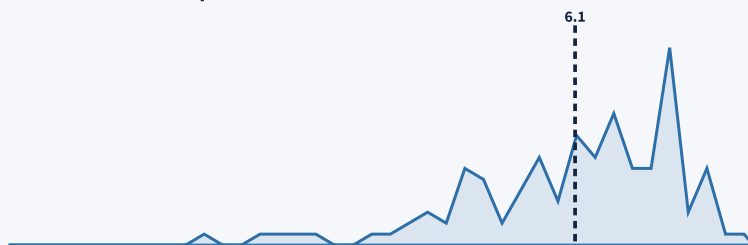
7-8 solid 7
5-6 mostly right 3
0-3 needs a check 2
Under S1, a score of 4 cannot occur.

All 12 situations at a glance · Best / Worst / total

Item score = Best credit key(B)-1 + Worst credit 5-key(W). Orange cell = 3 or below.

	01	02	03	04	05	06	07	08	09	10	11	12
Best credit (0-4)	4	3	4	3	2	4	3	2	4	3	0	3
Worst credit (0-4)	4	4	3	4	4	4	3	0	4	4	1	3
Item score (0-8)	8	7	7	7	6	8	6	2	8	7	1	6

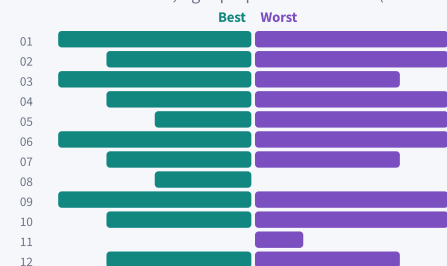
Position in the 120-person cohort · mean item score



Blue shape = cohort spread (dark = middle 50%) · orange = this candidate with 90% interval. A position on the spread; exact relative standing is in Appendix R.

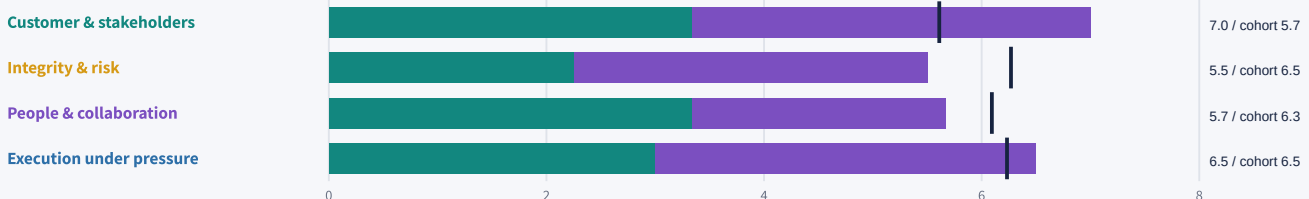
Best · Worst credit by situation

Left teal = Best credit, right purple = Worst credit (0-4 each).



Competency snapshot · Best credit + Worst credit (0-8)

Teal = Best credit, purple = Worst credit · navy line = cohort median · competency tags are a sample mapping.



RELATIVE-COMPARISON EVIDENCE · OVERVIEW (CANDIDATE MARKED)

The candidate's mean is 6.1, above the cohort mean of 6.05.

Type: borderline, stability 88.7%.

Candidate C-0417 at a glance

Candidate mean

6.1

cohort mean 6.05 · 0-8

Judgment type

Borderline

resample stability 88.7%

Axis alternations

4

cohort mean 3.0

Harmful picks

1

F11 Info security

Inversions · near

2 · 0

situations with gap < 0 · gap = 1

Answered

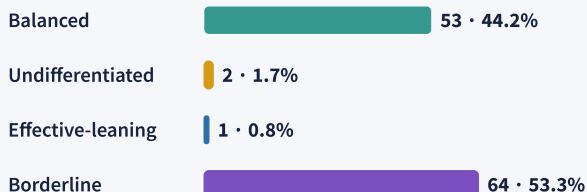
12/12

all situations answered

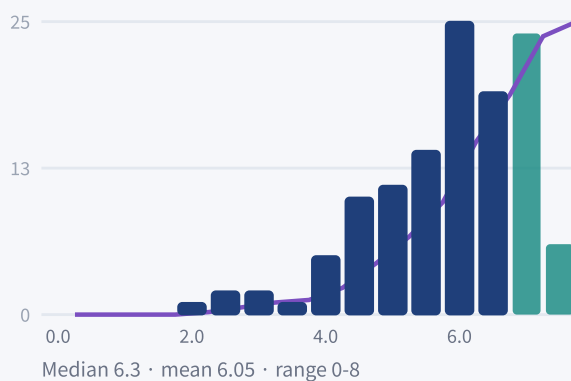
Cohort distributions (comparison base)

Type and score distributions

Judgment-type distribution (N · share)

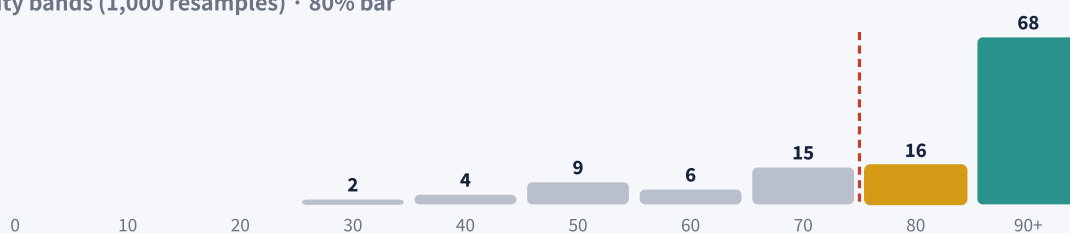


Mean-score distribution · purple line = cumulative share



Stability bands

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



RELATIVE-COMPARISON EVIDENCE · DIFFICULTY (CANDIDATE MARKED)

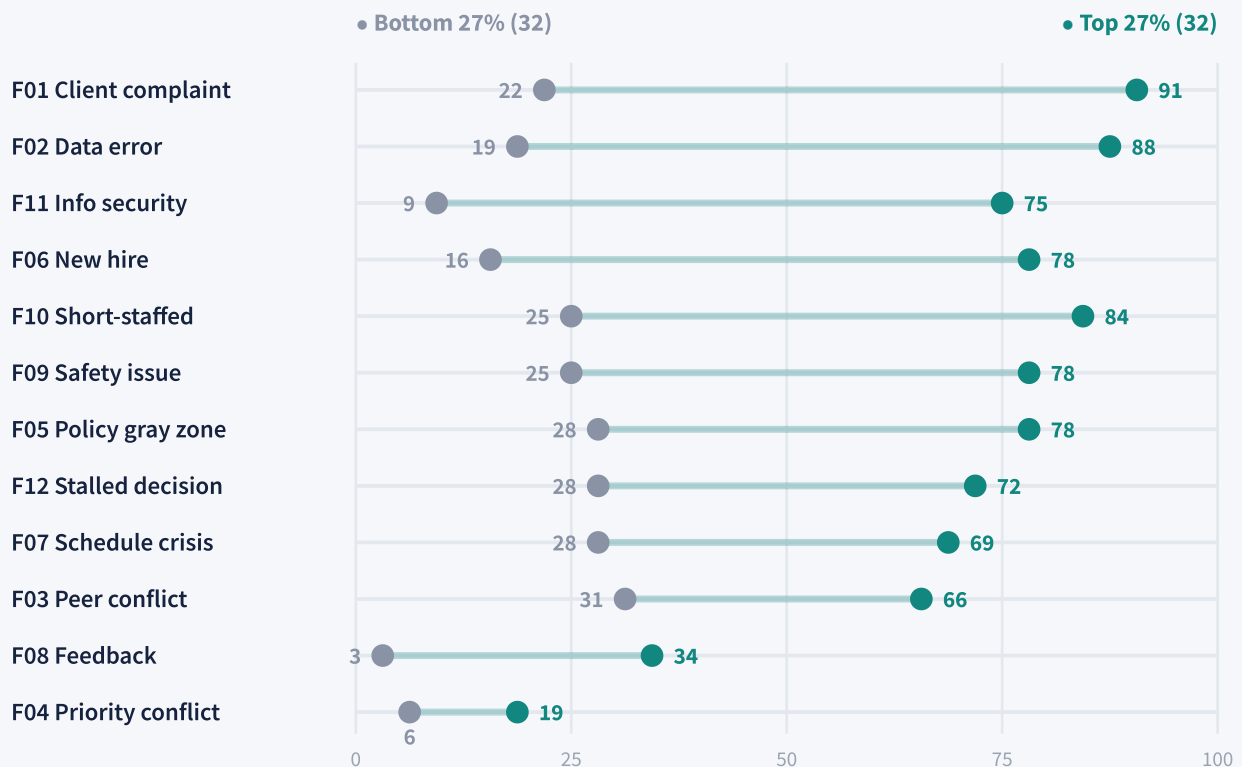
**Highest for the candidate: Client complaint and New hire (8.0);
lowest: Info security (1.0).**

Candidate per-situation scores

Situation	Score	Pick
F01 Client complaint	8.0	model pick
F02 Data error	7.0	best rank 4
F03 Peer conflict	7.0	model pick
F04 Priority conflict	7.0	best rank 4
F05 Policy gray zone	6.0	best rank 3
F06 New hire	8.0	model pick
F07 Schedule crisis	6.0	best rank 4
F08 Feedback	2.0	best rank 3
F09 Safety issue	8.0	model pick
F10 Short-staffed	7.0	best rank 4
F11 Info security	1.0	best rank 1
F12 Stalled decision	6.0	best rank 4

Cohort difficulty dumbbells (comparison base)

Model-pick dumbbells · by situation



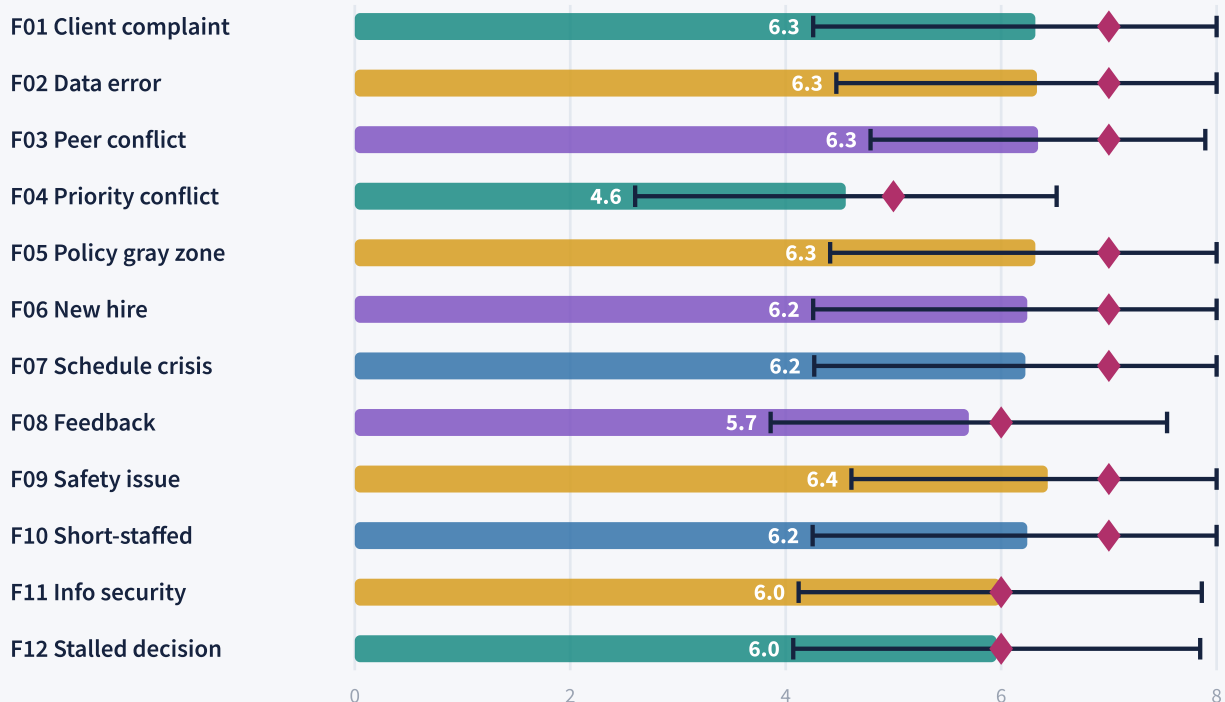
RELATIVE-COMPARISON EVIDENCE · MEANS (CANDIDATE MARKED)

The widest gap from the cohort mean is Info security (-5.0).

Candidate vs cohort means

Situation	Cohort mean	Candidate	Delta
F01 Client complaint	6.32	8.0	+1.7
F02 Data error	6.33	7.0	+0.7
F03 Peer conflict	6.34	7.0	+0.7
F04 Priority conflict	4.56	7.0	+2.4
F05 Policy gray zone	6.32	6.0	-0.3
F06 New hire	6.24	8.0	+1.8
F07 Schedule crisis	6.22	6.0	-0.2
F08 Feedback	5.70	2.0	-3.7
F09 Safety issue	6.43	8.0	+1.6
F10 Short-staffed	6.24	7.0	+0.8
F11 Info security	5.99	1.0	-5.0
F12 Stalled decision	5.96	6.0	+0.0

Cohort mean back-trace (comparison base)

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

RELATIVE-COMPARISON EVIDENCE · HARMFUL & INVERSIONS (CANDIDATE MARKED)

The candidate shows harmful picks in 1, inversions in 2, and near-inversions in 0 situations.

Candidate harmful & inverted situations

Harmful picks: 1

Situation	Detail	Score
F11 Info security	harmful option as Best	1.0

Inversions: 2 · near: 0

Situation	gap	Score
F08 Feedback	gap -2	2.0
F11 Info security	gap -3	1.0
none		

Cohort harmful & inversion analysis (comparison base)

Harmful picks · inversions · near-inversions by situation



RELATIVE-COMPARISON EVIDENCE · DISCRIMINATION (CANDIDATE MARKED)

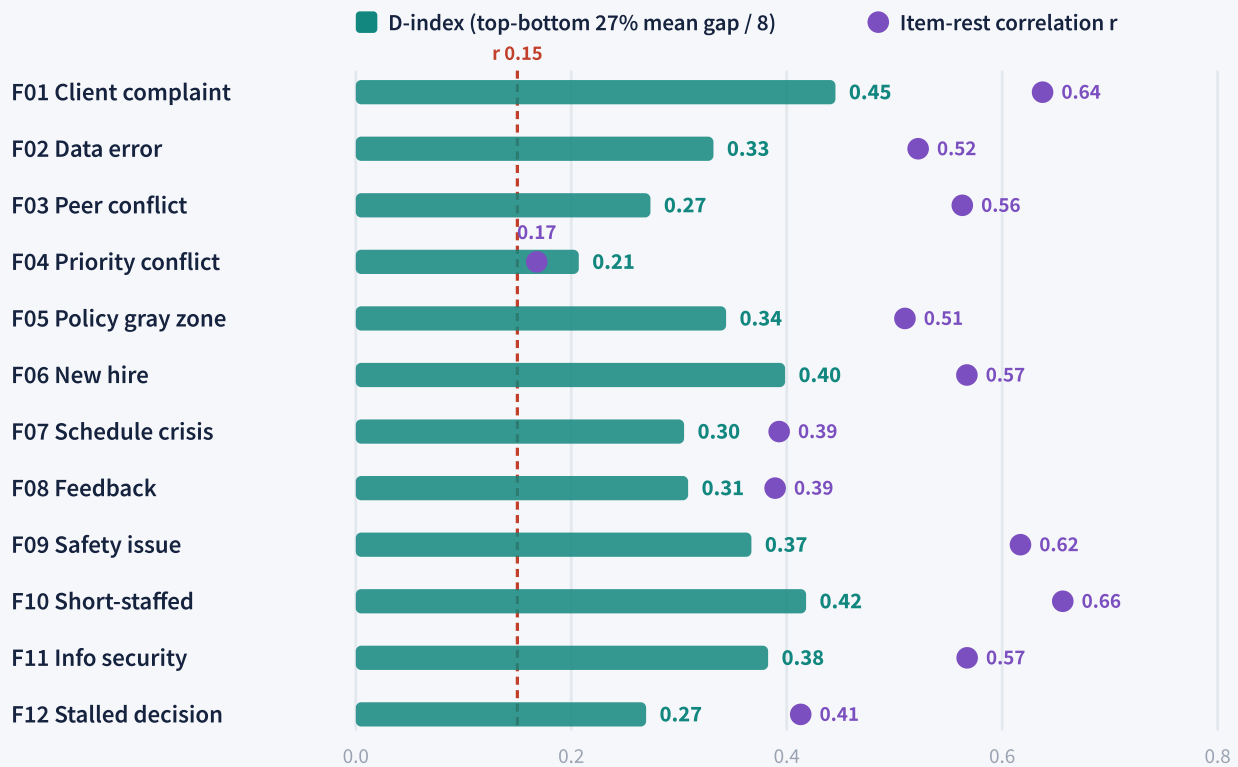
On the discrimination split the candidate falls in the middle group.

Where the candidate sits in the discrimination math

In the top/bottom 27% split the candidate falls in the **middle group** (mean 6.1; groups are cut by mean score). This metric is a property of the items, not an evaluation of the candidate.

Cohort discrimination (comparison base)

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



RELATIVE-COMPARISON EVIDENCE · STABILITY (CANDIDATE MARKED)

Candidate type stability is 88.7%, on the firm side.

Candidate type stability

Candidate stability

88.7%

1,000 deterministic resamples

Type

Borderline

cohort borderline: 64

Position in cohort

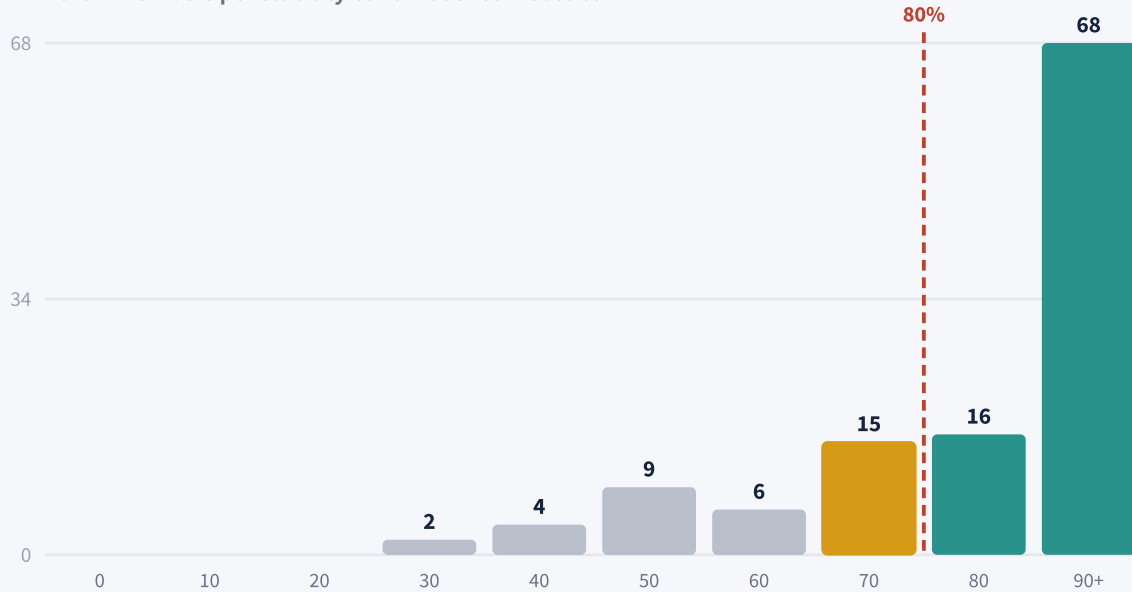
40.0% of members below

pseudo-standardized (descriptive, not an evaluation)

Cohort stability distribution (comparison base)

Type stability distribution · 1,000 resamples

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



RELATIVE-COMPARISON EVIDENCE · ALTERNATION (CANDIDATE MARKED)

The candidate alternates axes 4 times, at F02→F03, F03→F04, F07→F08, F09→F10.

Candidate axis alternations

Alternations

4

cohort mean 3.0 · max 11

Boundaries

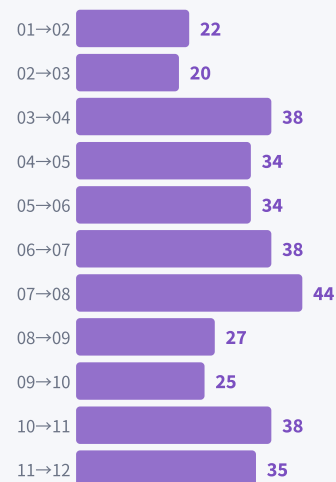
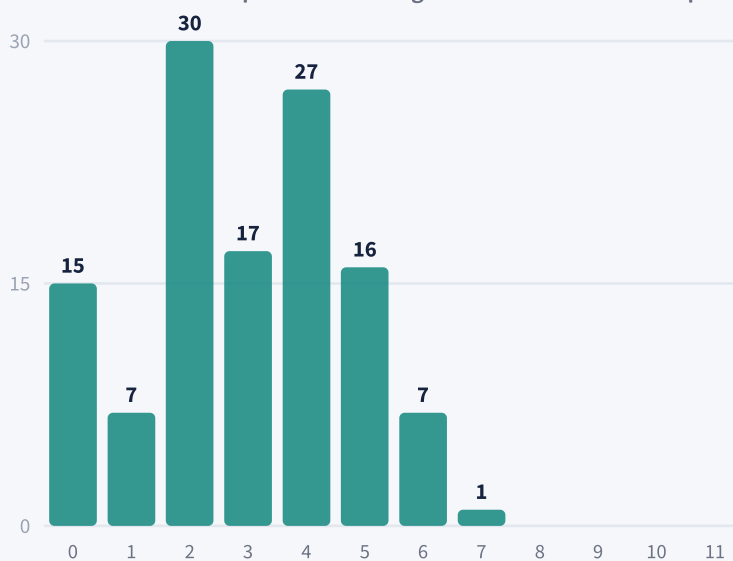
4

F02→F03, F03→F04, F07→F08, F09→F10

Cohort alternation distribution (comparison base)

Axis alternation · by member and boundary

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



RELATIVE-COMPARISON EVIDENCE · BANDS (CANDIDATE MARKED)

10 of the candidate's responses sit in the 6-8 band; 2 fall below 4.

Candidate band placement

Of 12 situations: 6-8 band 10, 4-6 0, 2-4 1, 0-2 1. Below 4: F08 Feedback, F11 Info security.

Situation	Score	Band
F01 Client complaint	8.0	6-8
F02 Data error	7.0	6-8
F03 Peer conflict	7.0	6-8
F04 Priority conflict	7.0	6-8
F05 Policy gray zone	6.0	6-8
F06 New hire	8.0	6-8
F07 Schedule crisis	6.0	6-8
F08 Feedback	2.0	2-4
F09 Safety issue	8.0	6-8
F10 Short-staffed	7.0	6-8
F11 Info security	1.0	0-2
F12 Stalled decision	6.0	6-8

Cohort band matrix (comparison base)

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

RELATIVE-COMPARISON EVIDENCE · SYNTHESIS (CANDIDATE MARKED)

Candidate: mean 6.1, stability 88.7%, 4 alternations, harmful picks in 1 situation.

Candidate summary (nine metrics)

1. Mean score

6.1

cohort 6.05

2. Best · worst situation

8.0 / 1.0

0-8 scale

3. Inversions

2

16.3% of cohort pairs

4. Model picks

4

Best rank 5

5. Stability

88.7%

80% bar

6. Alternations

4

cohort mean 3.0

7. 6-8 band

10

68.1% of cohort pairs

8. Type

Borderline

cohort borderline 64

9. Harmful picks

1

75.8% of cohort

Cohort nine-graph synthesis (comparison base)

Nine pieces of evidence at a glance

1. Difficulty gap (top - bottom 27%) · p4

69pp / 13pp

Client complaint / Priority conflict



2. Per-situation mean range · p5

4.6 – 6.4

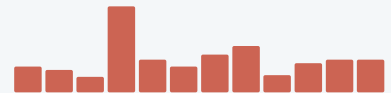
Priority conflict – Safety issue



3. Inverted picks (of all pairs) · p6

16.3%

gap < 0 of 1,440 pairs



4. Discrimination D · p7

mean 0.34

lowest Priority conflict 0.21



5. Type stability ≥ 80% · p8

84 · 70%

borderline 64 least stable



6. Axis alternations (mean) · p9

3.0

F07→F08 44



7. 6-8 band share · p10

68.1%

0-2 band 4.3%



8. Reliability (α · split-half) · p7

0.84 · 0.82

all 12 items working (r ≥ 0.15)



9. Members with a harmful pick · p6

75.8%

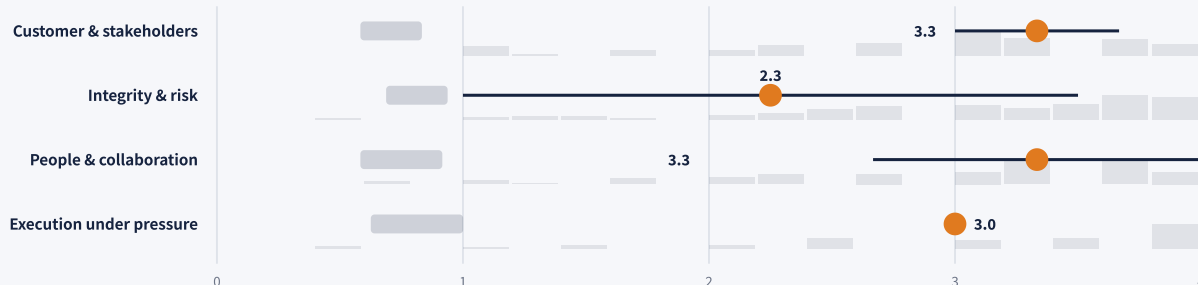
91 of 120



COMPETENCY PROFILE

**Judgment is most solid in ‘Customer & stakeholders’;
‘Integrity & risk’ has a wide interval and is the point to
check in the interview.**

Best credit by competency · this candidate on the cohort spread



Worst credit by competency · this candidate on the cohort spread



Most solid competency

Customer & stakeholders

highest Best+Worst mean

Check in the interview

Integrity & risk

lowest mean, widest interval

How to read

Shape = cohort spread, dark bar = middle 50%, orange dot = this candidate, navy line = 90% bootstrap interval. With 2–4 situations per competency the intervals are wide. A starting point for the conversation, not a ranking.

RISKY PICKS · HARM AREAS

There are 2 flagged picks and no repeated pattern.
One data protection scenario is rated Critical.

Risky-pick timeline

bar = item score (0-8) · circle = risky pick (red critical · orange high · gold model action as Worst) · flagged here = 1 key-risk · 1 model action as Worst



Judgment by harm area

	Harmful options shown	Rejected as Worst	Endorsed as Best
Ethics	3	2	0
Law & policy	2	1	0
Safety	2	1	0
Data protection	2	0	1
Client trust	4	2	0
People	6	2	0
Quality	3	1	0

Numbers = count of picks. Deeper purple = harmful actions rejected; orange = a harmful action endorsed as Best.

Risky picks to follow up

08 Feedback · model action picked as Worst (key-conflict item · caution level)

11 Info security · harmful option (severity 3) picked as Best

Risky picks are for interview follow-up, not reasons to reject. No repeated pattern (2+ in one area).

INTERVIEW · DEVELOPMENT GUIDE

Three things to check in the interview, and development notes ready to use after hire.

Structured interview guide · questions built from this candidate's evidence

Question 1 · Critical risky pick ⑪

“If a partner urgently asks for a customer list, what would you check before sending it?”

Follow-up: What if your manager says 'just send it'?

Strong answer

Approval, minimum fields and a secure channel first

Concern

'Send first, report later' or thinks masking alone is enough

Evidence

Info security: B key 1 · W key 4

Question 2 · E2 on ambiguous item ⑧ - check whether it is the person or the item

“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

Evidence

Feedback: B key 3 · W key 5

Question 3 · Action axis borderline - picked key 3 on 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

Evidence

Policy gray zone: B key 3 · W key 1

Development notes (for onboarding after hire)

Data-protection judgment

Build the habit of checking approval, minimum fields and a secure channel before any outside request. Suggest a data-protection exercise in month one.

Effective-action axis

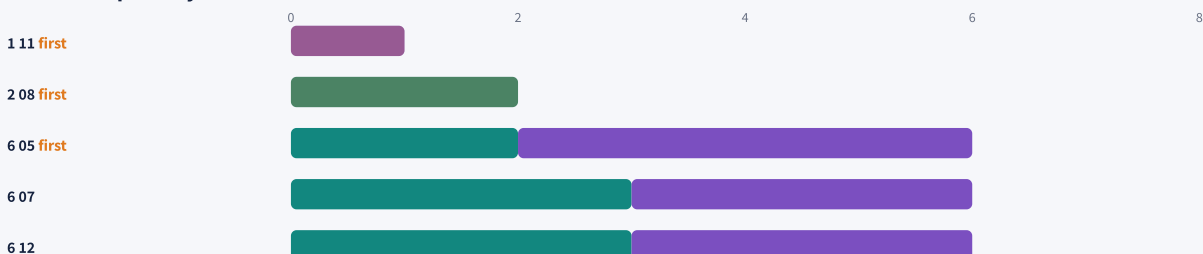
Some picks were 'safe' options (key 3–4) instead of the model action. Stretch tasks that require bringing stakeholders to agreement would help.

Strength to use

The most harmful actions in law · policy, safety, client trust were all rejected. This is a strength to use right away in risk-review work.

Questions are candidates picked automatically from this test's evidence; the interviewer makes the final choice. Development notes are not a basis for the hiring decision.

Interview priority · items from lowest score



Bar = item score (teal Best credit below + purple Worst credit above, 0–8). Orange = 3 or below. Starting with the three on the left makes best use of interview time.

INTERVIEWER DEEP QUESTIONS · RUBRIC

For the 2 flagged picks and one more situation to probe: the question, follow-ups, and signs of a strong and a weak answer.

08 Feedback · key-conflict item · caution level

“In Feedback, why did ‘Sort out the criticism, discuss fixes 1:1’ look like the worst choice?”

Follow-ups

“Who was affected by that decision?”
 “What would you change if you did it again?”
 “How did you check the rules or approval process?”

Signs of a strong answer

Same direction as the model action: “Sort out the criticism, discuss fixes 1:1”
 Names the risk and the stakeholders first
 A concrete case · outcome · lesson

Signs of a weak answer

Defends the pick as is: “Sort out the criticism, discuss fixes 1:1”
 Gives only speed or convenience as the reason
 General statements with no real case

11 Info security · critical · severity 3

“Have you acted like ‘Send the file first, report later’ in a situation like Info security? What did you weigh first?”

Follow-ups

“Who was affected by that decision?”
 “What would you change if you did it again?”
 “How did you check the rules or approval process?”

Signs of a strong answer

Same direction as the model action: “Share via an encrypted channel after security approval”
 Names the risk and the stakeholders first
 A concrete case · outcome · lesson

Signs of a weak answer

Defends the pick as is: “Send the file first, report later”
 Gives only speed or convenience as the reason
 General statements with no real case

05 Policy gray zone · at the cohort mean (6/8)

“In Policy gray zone, why did ‘Apply the discount without approval to hit the target’ look like the worst choice?”

Follow-ups

“Who was affected by that decision?”
 “What would you change if you did it again?”
 “How did you check the rules or approval process?”

Signs of a strong answer

Same direction as the model action: “Follow the approval step, share the timeline”
 Names the risk and the stakeholders first
 A concrete case · outcome · lesson

Signs of a weak answer

Defends the pick as is: “Apply the discount without approval to hit the target”
 Gives only speed or convenience as the reason
 General statements with no real case

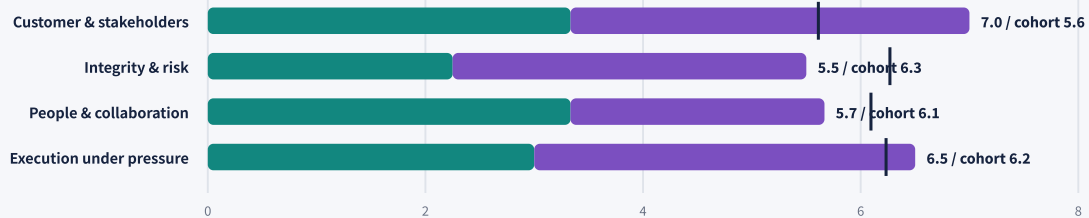
How to record the answer

- For each question, tick the signs you heard and write one sentence of what the candidate said.
- Do not create a new score. Bring this record and the report to the decision meeting together.

DEVELOPMENT SUGGESTIONS · 30/60/90 DAYS

If hired, start the first 90 days with ‘Integrity & risk’.
Every suggestion comes from choices this person actually made.

Priority by competency · lowest first



Teal = Best credit, purple = Worst credit · navy line = cohort median.

Integrity & risk · 5.5/8

Evidence · Info security (1/8) · Best “Send the file first, report later”

30 days · review one similar real case with a senior colleague, against the model action “Share via an encrypted channel after security approval”

60 days · before a ‘Integrity & risk’ decision, write down the most harmful option first (think Worst first)

90 days · check two cases with the manager · see whether choices changed in the same kind of situation

People & collaboration · 5.7/8

Evidence · Feedback (2/8) · Best “Share a revision right away”

30 days · review one similar real case with a senior colleague, against the model action “Sort out the criticism, discuss fixes 1:1”

60 days · before a ‘People & collaboration’ decision, write down the most harmful option first (think Worst first)

90 days · check two cases with the manager · see whether choices changed in the same kind of situation

Execution under pressure · 6.5/8

Evidence · Schedule crisis (6/8) · Best “Ship the core, delay the rest”

30 days · review one similar real case with a senior colleague, against the model action “Tell the client first, offer a new timeline”

60 days · before a ‘Execution under pressure’ decision, write down the most harmful option first (think Worst first)

90 days · check two cases with the manager · see whether choices changed in the same kind of situation

Customer & stakeholders · 7.0/8

Evidence · Stalled decision (6/8) · Best “Summarize issues, ask the sponsor to decide”

30 days · review one similar real case with a senior colleague, against the model action “Set decision criteria and a deadline”

60 days · before a ‘Customer & stakeholders’ decision, write down the most harmful option first (think Worst first)

90 days · check two cases with the manager · see whether choices changed in the same kind of situation

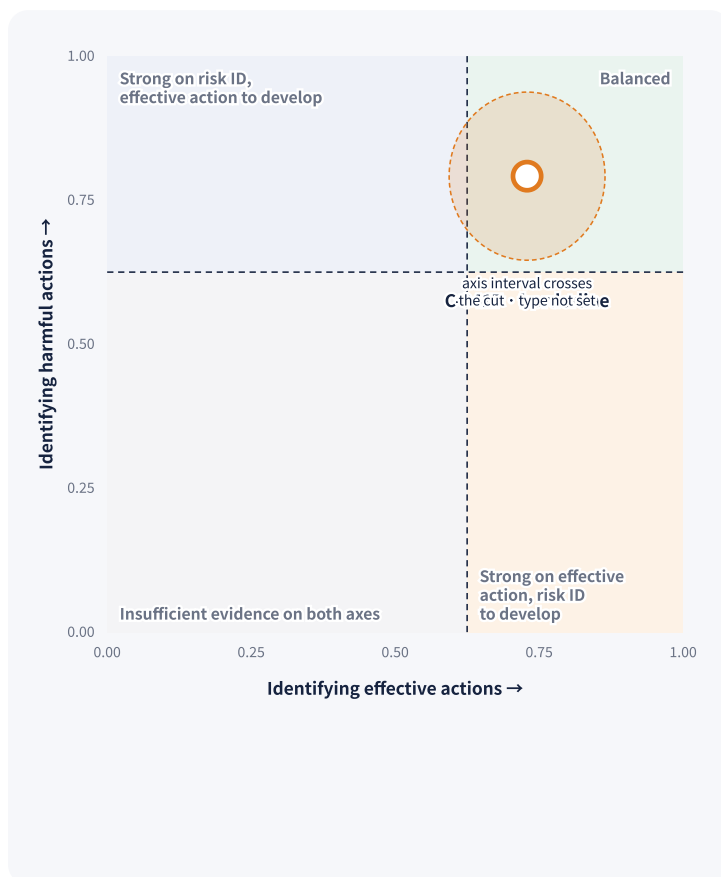
Basis and limits of these suggestions

- Each card comes from the lowest-scoring situation in that competency and its model action. It is tied to this person's actual choices, not general advice.
- The 30/60/90-day activities are onboarding suggestions, not a validated training program.
- Competency tags are a sample mapping, auto-checked.

JUDGMENT MAP · 2 AXES

Spotting what works and spotting what harms are read separately.

Harm spotting clears the bar with room to spare; effective-action spotting sits on the bar, so the interview should confirm it.



Effective-action ID · contributor ①

③ Peer conflict

Best: Raise it 1:1, propose a joint plan (key 5)

Effective-action ID · contributor ②

⑨ Safety issue

Best: Stop the work, check the cause, log it (key 5)

Harmful-action ID · contributor ①

⑨ Safety issue

Worst: Mention it at the next routine check (key 1)

Harmful-action ID · contributor ②

⑥ New hire

Worst: Call it out in front of the team (key 1)

Pulled the action axis down ①

⑪ Info security

Picked key 1 as Best: Send the file first, report later · model (key 5): Share via an encrypted channel after security approval

Pulled the action axis down ②

⑤ Policy gray zone

Picked key 3 as Best: Ask the manager to decide · model (key 5): Follow the approval step, share the timeline

Four types and borderline

Balanced

Both axes at or above the cut

Strong on effective action, risk identification to develop

Only the effective-action axis at or above the cut

Strong on risk identification, effective action to develop

Only the harmful-action axis at or above the cut

Insufficient evidence on both axes

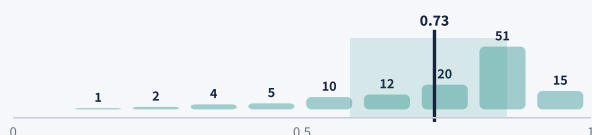
Both axes below the cut

Borderline

Interval crosses the cut · type not set

$x = \text{mean best_credit} = (\text{key}(B)-1)/4$ · $y = \text{mean worst_credit} = (5-\text{key}(W))/4$ · interval = 90% over 1,000 item resamples · before calibration this is descriptive, not θ · the cut is a criterion reference, not the cohort median.

Effective-action ID · cohort spread



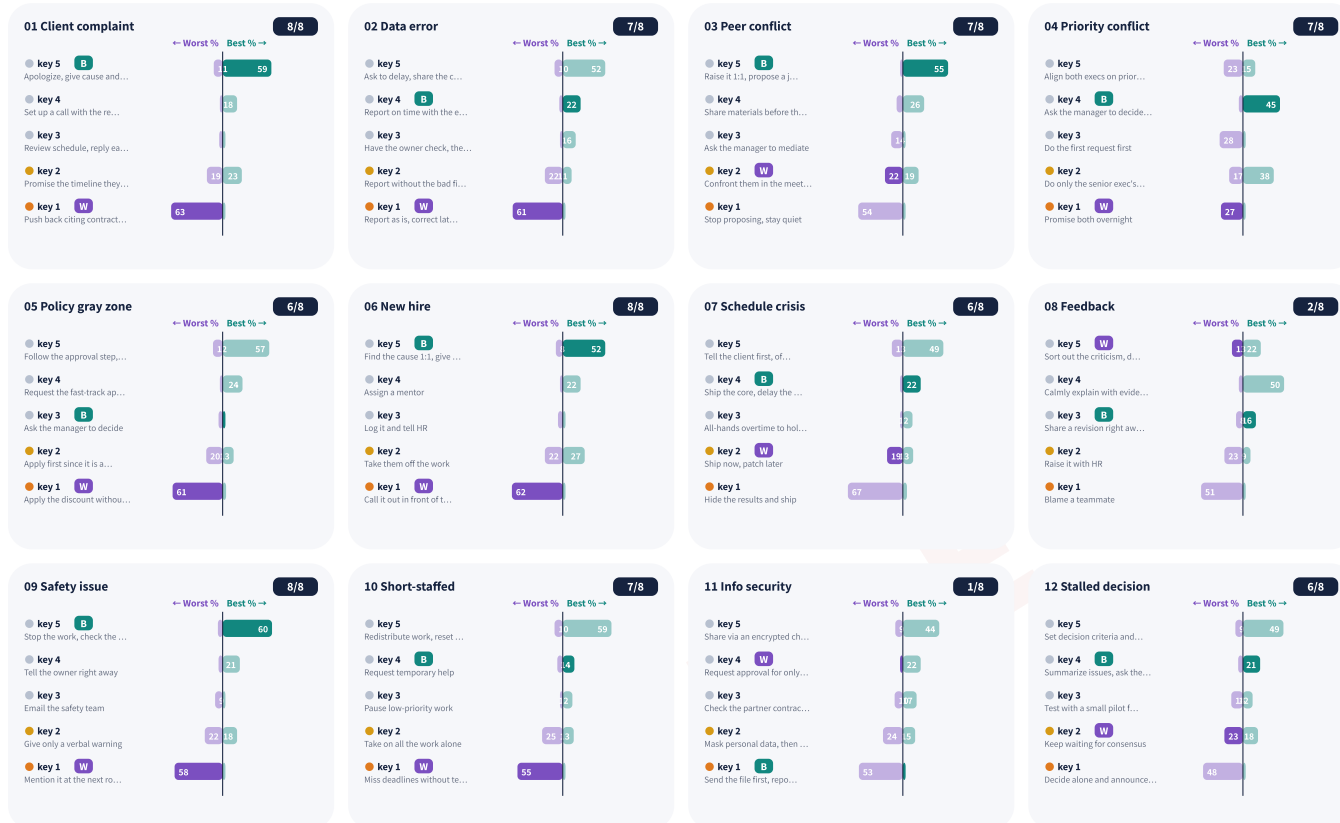
Harmful-action ID · cohort spread



CHOICE FINGERPRINT · 12 SITUATIONS × 5 OPTIONS

Model action picked as Best in 4 of 12; most harmful action picked as Worst in 7.

Where this person differs from the cohort shows at a glance.



Each situation's 5 options in key order (key 5 = model action ... key 1 = most harmful), with the share of the 120-person cohort picking each as Best (right, teal) and Worst (left, purple). B/W chips = this candidate's picks. Dot color = harm severity (orange high+ · gold low · grey none). Full option text is in the authorized appendix.

Rare choices (picked by under 15% of the cohort)

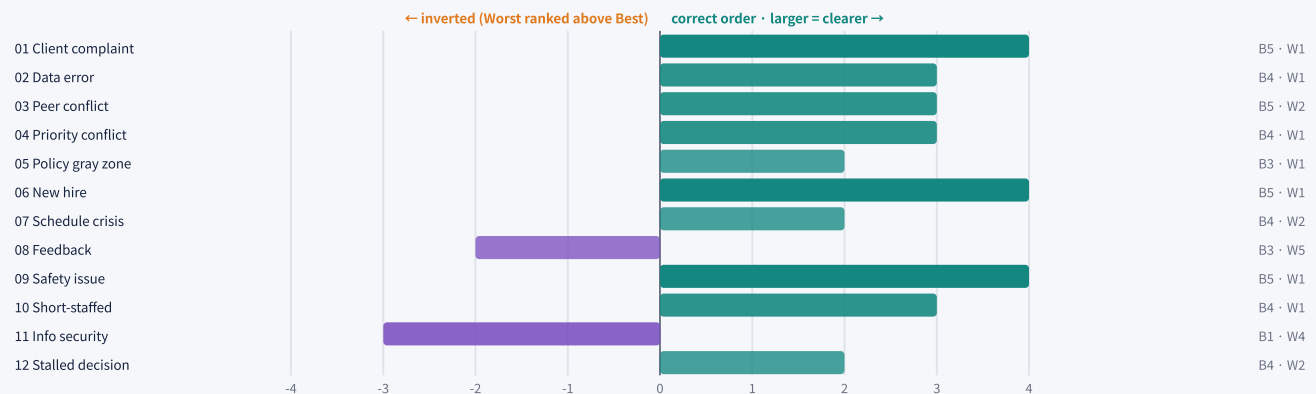
- B 05 Policy gray zone · key 3 · Ask the manager to decide · cohort 2%
- B 10 Short-staffed · key 4 · Request temporary help · cohort 14%
- B 11 Info security · key 1 · Send the file first, report later · cohort 3%
- W 08 Feedback · key 5 · Sort out the criticism, discuss fixes 1:1 · cohort 13%
- W 11 Info security · key 4 · Request approval for only the needed fields · cohort 3%

A rare choice is not a wrong choice. It is a good place to ask for the reasoning.

CHOICE QUALITY · JUDGMENT GAP

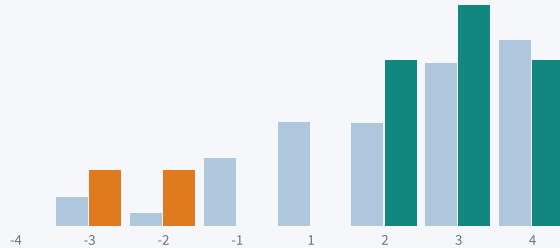
The gap is +3 or more in 7 of 12 situations.
The 2 inverted picks are the first thing to check.

Judgment gap by situation · key(Best) – key(Worst)

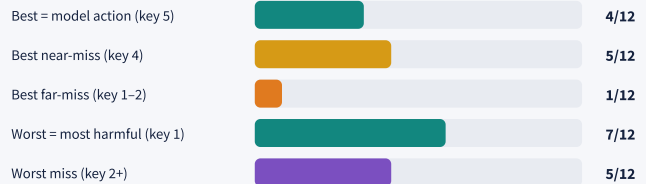


Gap distribution vs the cohort

light blue = cohort · solid = this candidate



Near-miss vs far-miss



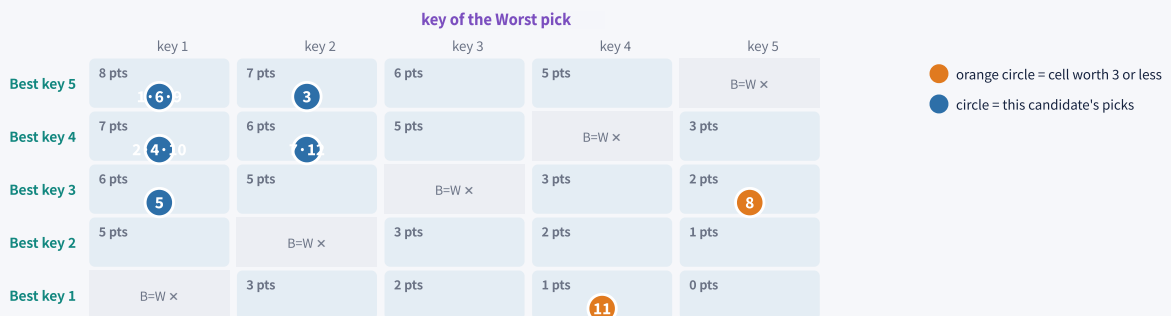
Near-miss = the second-best (or second-worst) option, a much lighter signal than a far-miss.

2 inverted picks · the Worst pick has a higher key than the Best pick

- 08 Feedback** · Best key 3 “Share a revision right away” · Worst key 5 “Sort out the criticism, discuss fixes 1:1” · cohort inversion rate 23%
- 11 Info security** · Best key 1 “Send the file first, report later” · Worst key 4 “Request approval for only the needed fields” · cohort inversion rate 16%

An inversion means the priorities in that situation were read opposite to the the scoring key. Ask first how the person understood the situation.

Where the 12 picks land on the S1 score table



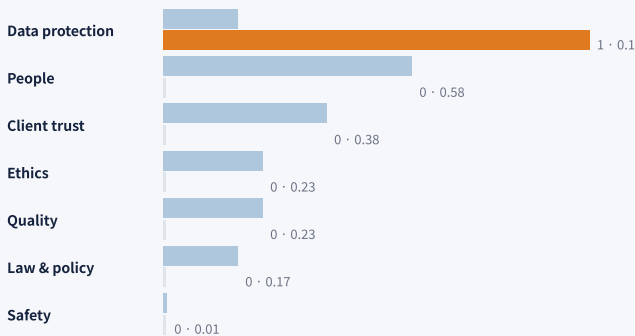
Cell = Best key × Worst key and its score (0–8) · % = share of the cohort's 120 × 12 responses in that cell · circle = this candidate's picks and situation numbers.

ATTRACTIVE DISTRACTORS · HARM AREAS

Of 8 Best picks that were not the model action, 5 were the second-best action.

The only area where a harmful option was picked as Best is Data protection.

Pull toward harmful options by area · this candidate vs cohort



Which wrong options were attractive

- Of 8 Best picks that were not the model action: 5 near-miss (key 4), 1 far-miss (key 1–2)
- Areas where a harmful option (severity 2+) was picked as Best: Data protection
- Cohort mean 1.79 per person; this candidate 1
- If much of the cohort made the same wrong pick, the option itself may be an attractive distractor, not just this person.

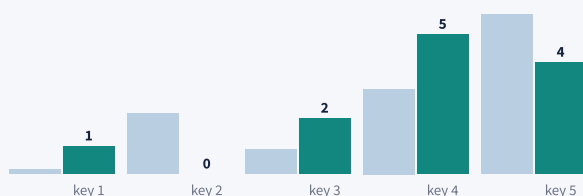
Every Best pick that was not the model action

Situation	key	Picked as Best	Area · severity	Cohort same pick	model action
02 Data error	key 4	Report on time with the error flagged	no harm area · sev 0	22%	Ask to delay, share the corrected version
04 Priority conflict	key 4	Ask the manager to decide with an impact table	no harm area · sev 0	45%	Align both execs on priorities together
05 Policy gray zone	key 3	Ask the manager to decide	no harm area · sev 0	2%	Follow the approval step, share the timeline
07 Schedule crisis	key 4	Ship the core, delay the rest	no harm area · sev 0	22%	Tell the client first, offer a new timeline
08 Feedback	key 3	Share a revision right away	no harm area · sev 0	16%	Sort out the criticism, discuss fixes 1:1
10 Short-staffed	key 4	Request temporary help	no harm area · sev 0	14%	Redistribute work, reset the schedule with stakeholders
11 Info security	key 1	Send the file first, report later	Data protection · sev 3	3%	Share via an encrypted channel after security approval
12 Stalled decision	key 4	Summarize issues, ask the sponsor to decide	no harm area · sev 0	21%	Set decision criteria and a deadline

Full option text is in the authorized appendix. Severity = draft tag (0 none ... 3 critical).

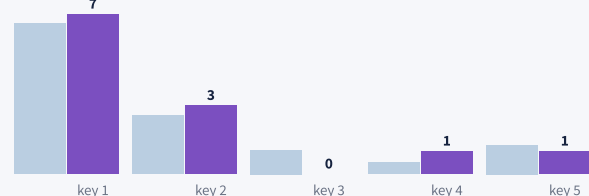
key of Best picks

light = cohort · solid = this candidate (number = count)



key of Worst picks

light = cohort · solid = this candidate (number = count)



DIFFICULTY-ADJUSTED ITEM STORY

Ahead of the cohort on 2 of 4 hard items. Missed 0 easy item(s).

Hard items mean

4.0

cohort 5.6 · 4 items

Easier items mean

7.1

cohort 6.3 · 8 items

Furthest ahead

04

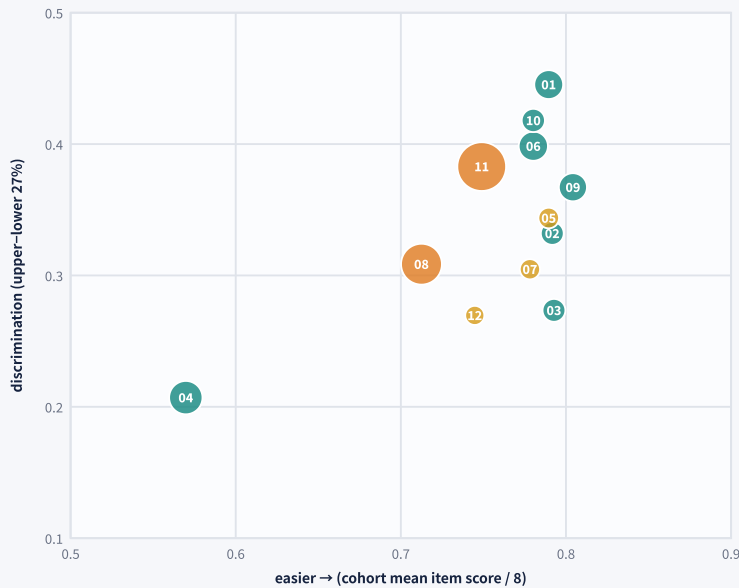
+2.4 · Priority conflict

Furthest behind

11

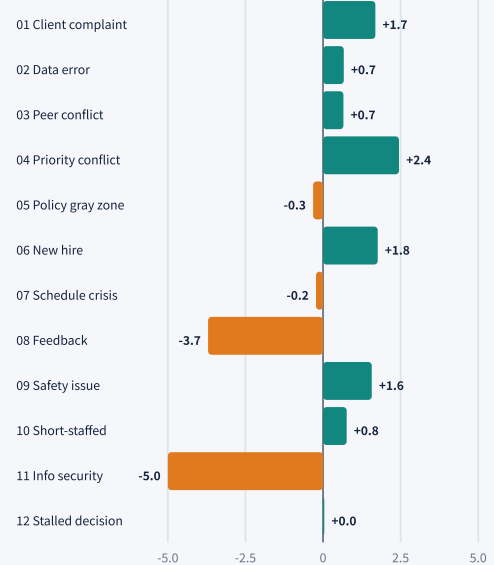
-5.0 · Info security

Item difficulty × discrimination · circle = this candidate vs cohort



Circle size = size of the difference from the cohort mean · teal = 0.5+ above · orange = 0.5+ below · gold = similar.

Difference from the cohort mean



Ahead on hard items

04 Priority conflict · cohort mean 4.6 → 7

12 Stalled decision · cohort mean 6.0 → 6

Missed on easy items

None.

How to read

Being ahead on a hard item is stronger evidence than being ahead on an easy one. Differences on high-discrimination items (top) are more trustworthy. Based on the 120-person cohort.

Item figures · the source of every dot and bar on this page

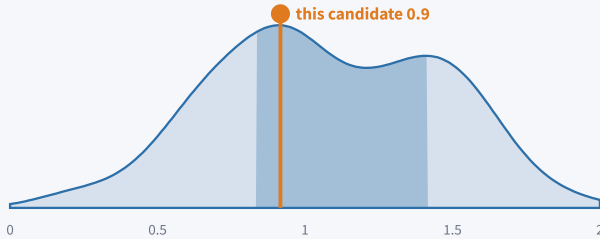
Situation	Competency	Cohort mean	Ease p	D	Item-rest r	This person	Diff	Hard
01 Client complaint	Customer & stakeholders	6.32	0.79	0.45	0.64	8	+1.68	
02 Data error	Integrity & risk	6.33	0.79	0.33	0.52	7	+0.67	
03 Peer conflict	People & collaboration	6.34	0.79	0.27	0.56	7	+0.66	
04 Priority conflict	Customer & stakeholders	4.56	0.57	0.21	0.17	7	+2.44	✓
05 Policy gray zone	Integrity & risk	6.32	0.79	0.34	0.51	6	-0.32	
06 New hire	People & collaboration	6.24	0.78	0.40	0.57	8	+1.76	
07 Schedule crisis	Execution under pressure	6.22	0.78	0.30	0.39	6	-0.22	
08 Feedback	People & collaboration	5.70	0.71	0.31	0.39	2	-3.70	✓
09 Safety issue	Integrity & risk	6.43	0.80	0.37	0.62	8	+1.57	
10 Short-staffed	Execution under pressure	6.24	0.78	0.42	0.66	7	+0.76	
11 Info security	Integrity & risk	5.99	0.75	0.38	0.57	1	-4.99	✓
12 Stalled decision	Customer & stakeholders	5.96	0.74	0.27	0.41	6	+0.04	✓

p = cohort mean / 8 · D = (upper 27% mean – lower 27% mean) / 8 · hard = p < 0.76 · 120-person cohort.

CONSISTENCY · TYPE STABILITY

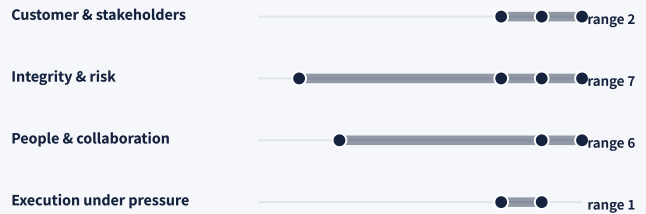
‘Balanced’ in 89% of resamples. The interval touches the cut, so the report shows ‘Borderline’

Agreement with the cohort's most common pick



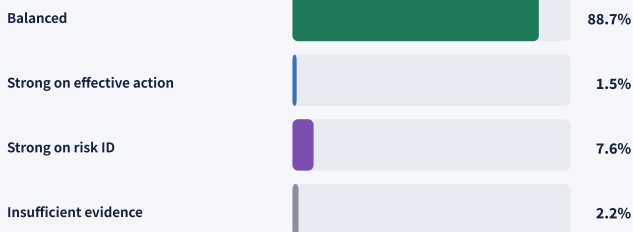
Best agreement 5/12 + Worst agreement 6/12. Blue shape = cohort spread. Differing from the majority is not the same as being wrong.

Consistency within each competency



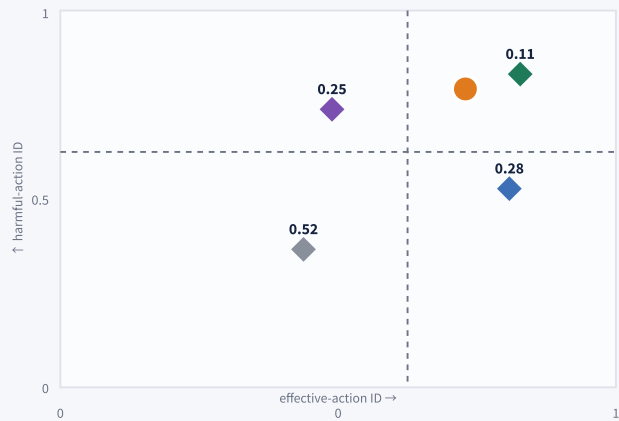
Dots = item scores in that competency · a narrower range means more consistent judgment.

Type stability · 1,000 item resamples



Share of 1,000 resamples (12 situations drawn with replacement) landing in each type. ‘Balanced’ 89%.

Distance to each type centre



Diamond = cohort centre for each type · number = distance to this candidate (orange) · dashed = 0.625 cut.

Agreement by situation · ✓ = same as the cohort's most common pick, % = cohort share making the same pick

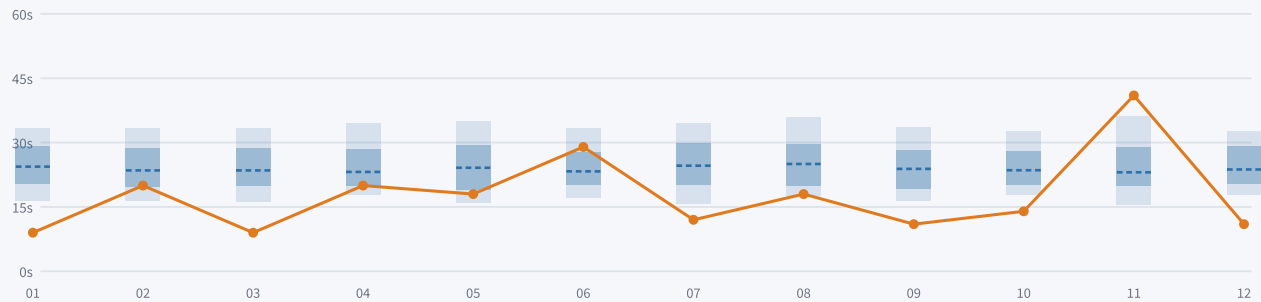
	01	02	03	04	05	06	07	08	09	10	11	12
Best agrees	✓	22%	✓	✓	2%	✓	22%	16%	✓	14%	3%	21%
Worst agrees	✓	✓	22%	27%	✓	✓	19%	13%	✓	✓	3%	23%

RESPONSE PATTERN · TIME · POSITION

Response speed and pick positions are both in normal range.

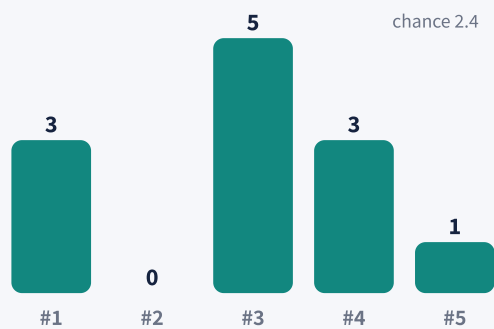
The situation with the longest deliberation is a good opening for the interview.

Response time per item · on the cohort band



Orange = this candidate · blue band = cohort middle 50% (dark) and 80% (light) · dashed = cohort median.

Screen position of Best picks



Options are shuffled per person. Max at one position 5/12 (review at 9+).

Response process summary

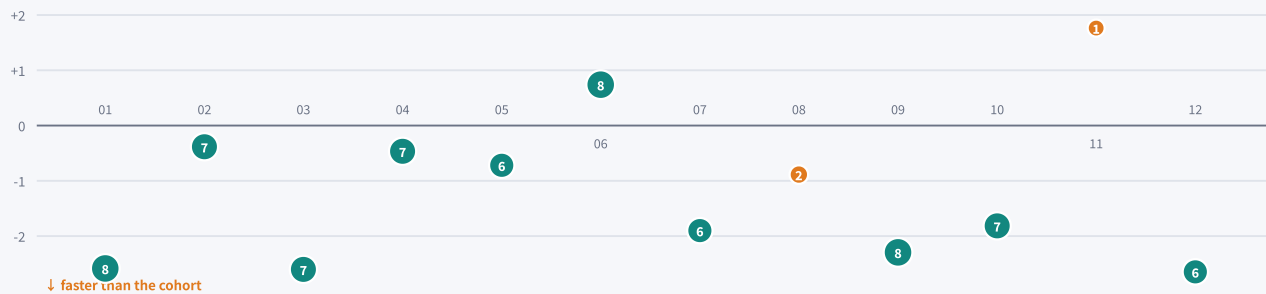
- Median response time 18s (review below 8s)
- Longest deliberation: 11 Info security (41s)
- Best = Worst answers: 0 (rejected by the server)
- Integrity signals: none
- Signals are for review, not proof of misconduct.

RESPONSE TIME IN DEPTH

Most answers were faster than the cohort, and longer deliberation went with lower scores ($r = -0.57$). These are the hard moments to hear about in the interview.

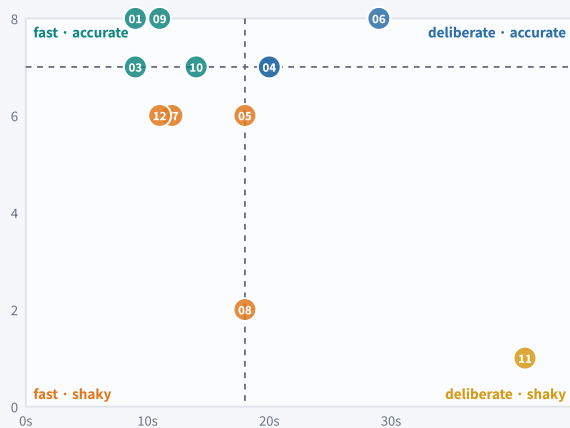
Response time per item · vs cohort (log z)

↑ longer than the cohort



Number = item score · teal = 6+ · orange = 5 or below · up = longer than the cohort.

Deliberation × accuracy quadrant

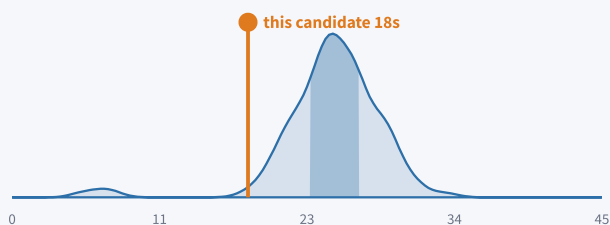


response time (s) · vertical dashed = own median · horizontal = score 6

How time relates to judgment

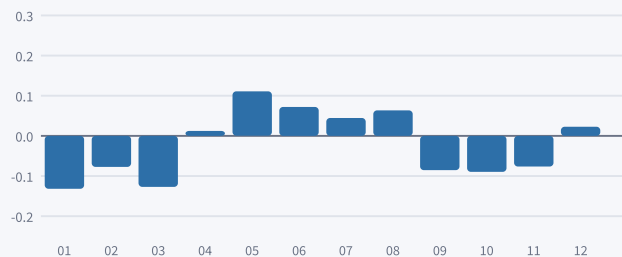
- This candidate's time-score correlation $r = -0.57$ (cohort median -0.02)
- Negative = items with longer deliberation scored lower. This may be confidence dropping in hard situations.
- Fast but shaky items: 4
- Fast response (cohort bottom 25%) + harmful Best: none
- A correlation over 12 points is unstable. Use it only as a conversation opener.
- Time signals are for review, not proof of misconduct.

Median response time · this candidate on the cohort spread



Review threshold: median below 8s.

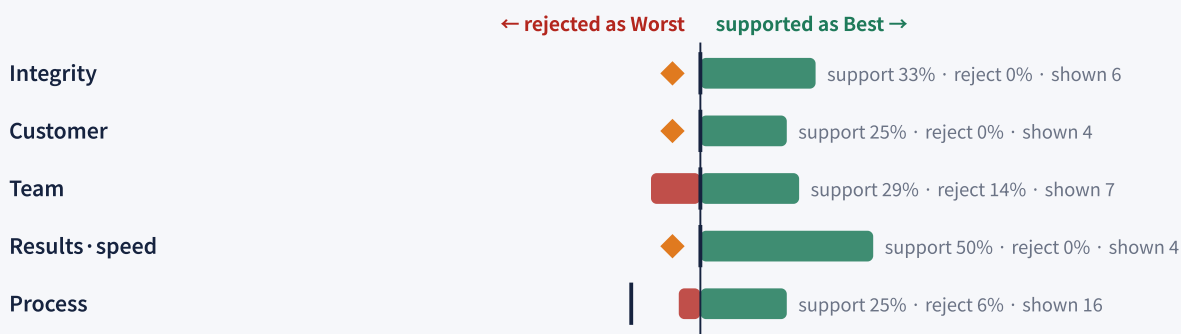
Time-score correlation by item · 120-person cohort

Whether people who took longer scored higher on that item (r). All are weak ($|r| < 0.2$), so time explains little about judgment quality.

VALUES FINGERPRINT · DECISION STYLE

No picks clash with company priority values. The value supported most often is 'Results · speed'

Value priorities shown in effectiveness judgments (in-house value framework draft, R4)

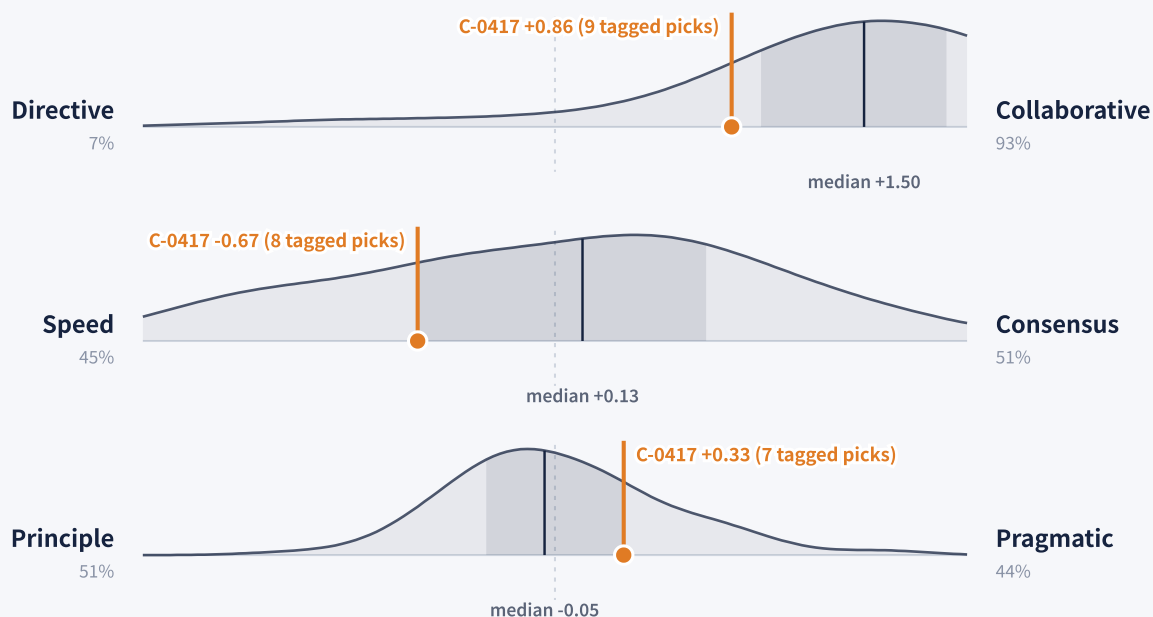


◆ = top value in the company culture profile · vertical line = cohort median reject rate · harmful options excluded · no total fit score · these are priorities shown in "what the candidate judges effective," not personal values.

No tension evidence The candidate never picked a reasonable option linked to a company priority value (integrity, customer, results) as Worst.

Not shown Shown fewer than 4 times: Autonomy

Decision style from actual picks (3 axes, R6)



Interview questions from this page

Directive ↔ Collaborative · score +0.86

"Tell me about a real case where you chose a more collaborative approach. When would the opposite approach have worked better?"

Speed ↔ Consensus · score -0.67

"Tell me about a real case where you chose a more speed approach. When would the opposite approach have worked better?"

The two axes furthest from 0 are shown · there is no right answer · use the answers to see whether the approach adapts to the situation.

MEASUREMENT NOTES · EVIDENCE TRAIL

Every number is computed from the 12 response records and can be traced on this page.

What this result says

- What the candidate judged most effective and most harmful in 12 work scenarios
- Descriptive scores on two axes with 90% bootstrap intervals
- Cases where a harmful action was endorsed, with severity
- Evidence to check in the interview

What it does not say

- Whether to hire, or an overall rank (no composite score)
- A calibrated ability θ (before key and calibration approval)
- Personality or personal values
- Misconduct (signals are only for review)

Evidence trail · where every number comes from

Scenario	Response ID	Picks and key	Screen position B/W	Time	Key · scoring version
01 Client complaint	resp_5b44...b44	B=key 5 · W=key 1	3 / 4	9s	Scoring key 2026-09-23 · S1
02 Data error	resp_5e44...ffd	B=key 4 · W=key 1	3 / 1	20s	Scoring key 2026-09-23 · S1
03 Peer conflict	resp_5d44...e6a	B=key 5 · W=key 2	1 / 2	9s	Scoring key 2026-09-23 · S1
04 Priority conflict	resp_5844...68b	B=key 4 · W=key 1	4 / 2	20s	Scoring key 2026-09-23 · S1
05 Policy gray zone	resp_5744...4f8	B=key 3 · W=key 1	4 / 5	18s	Scoring key 2026-09-23 · S1
06 New hire	resp_5a44...9b1	B=key 5 · W=key 1	3 / 5	29s	Scoring key 2026-09-23 · S1
07 Schedule crisis	resp_5944...81e	B=key 4 · W=key 2	1 / 2	12s	Scoring key 2026-09-23 · S1
08 Feedback	resp_5444...03f	B=key 3 · W=key 5	5 / 2	18s	Scoring key 2026-09-23 · S1
09 Safety issue	resp_5344...eac	B=key 5 · W=key 1	1 / 5	11s	Scoring key 2026-09-23 · S1
10 Short-staffed	resp_5642...4ce	B=key 4 · W=key 1	3 / 4	14s	Scoring key 2026-09-23 · S1
11 Info security	resp_5742...661	B=key 1 · W=key 4	3 / 5	41s	Scoring key 2026-09-23 · S1
12 Stalled decision	resp_5442...1a8	B=key 4 · W=key 2	4 / 3	11s	Scoring key 2026-09-23 · S1

Options are shown in a random order per person (screen position). The audit appendix (authorized users only) holds full option text, scoring rationale and the input snapshot sha256. Published results are never overwritten; re-scoring creates a new version.

Pre-calibration · descriptive

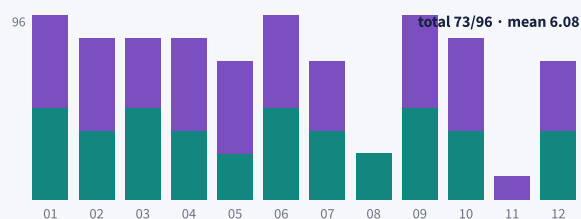
Scoring key 2026-09-23

Reference cut 0.625

No composite · no overall rank

Missing stays missing

Score build-up · Best + Worst credit over 12 situations



Best credit (teal) and Worst credit (purple) of each situation stacked in order. Every number in the table above adds up to this total. No weighting.

Share of the cohort making the same pick · Best / Worst



Teal = share of the 120-person cohort making the same Best pick as this candidate, purple = the same Worst pick. Lower means rarer, not wrong.

COMPETENCY INTERPRETATION · PRE-CALIBRATION

DESCRIPTIVE

A response pattern, not a hiring verdict

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

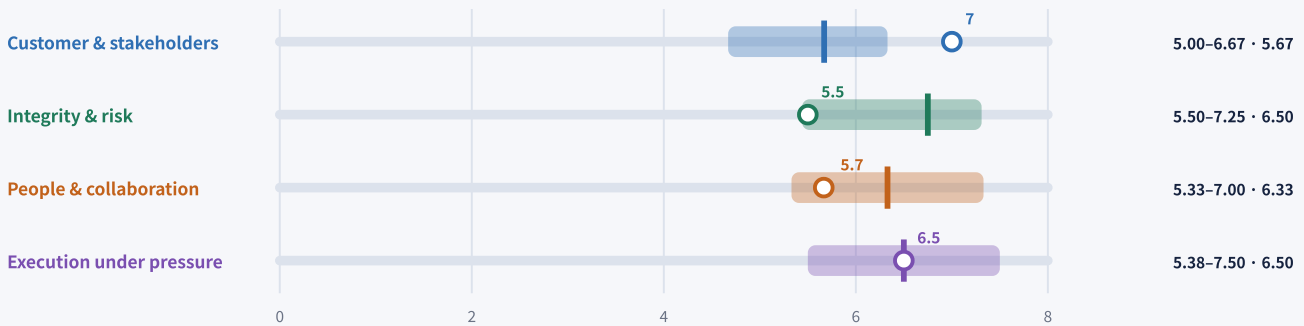
Situations

12

5 options, Best + Worst

Four distributions · Best + Worst credit / 8

Descriptive



Band = middle half; line = median; hollow dot = sample individual. 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

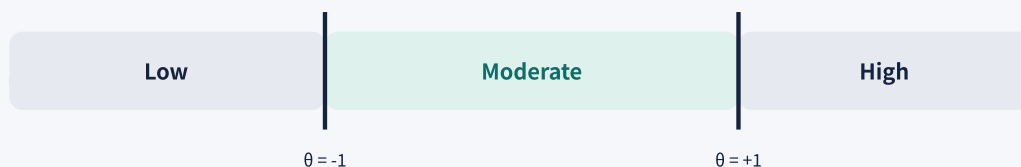
The provisional-synthetic report is blocked below $n=50$. The 20–49 Small tier needs a separately approved production-frozen asset.

Classification, in scope

Current Best/Worst type is descriptive, not an IRT Ability Band. The sample individual is Borderline: Balanced appears in 88.7% of 1,000 resamples, but its interval touches the 0.625 cut. Do not order candidates 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.

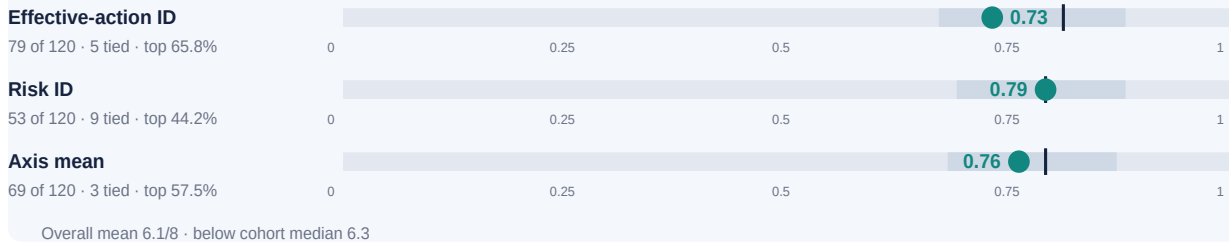
Boundary before interpretation

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.

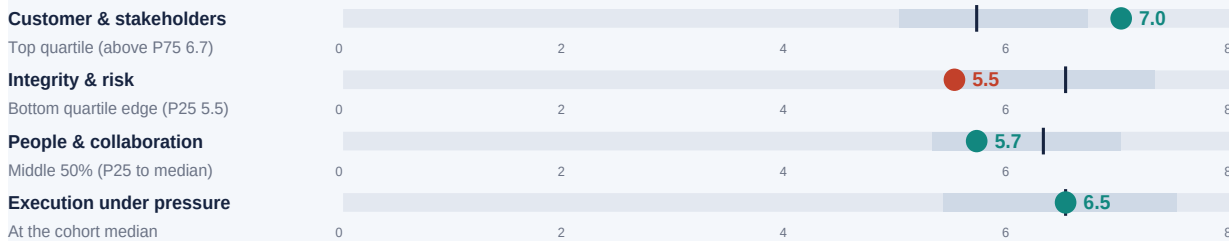
No system-made hiring verdict, candidate order or composite percentile. People decide.

Where the candidate stands in the 120-person cohort.

Two axis scores · exact ranks from the roster



Competency standing · vs published quartiles (P25-median-P75)



Situation standing · all 12 situations

Situation	Cand.	Cohort med.	Q1–Q3	Same-band count	Position
F01 Client complaint	8.0	7.0	5.0–8.0	83 of 120	Top-25% edge
F02 Data error	7.0	7.0	6.0–8.0	95 of 120	Middle 50%
F03 Peer conflict	7.0	7.0	5.0–8.0	84 of 120	Middle 50%
F04 Priority conflict	7.0	5.0	3.0–6.0	41 of 120	Top 25%
F05 Policy gray zone	6.0	7.0	5.0–8.0	85 of 120	Middle 50%
F06 New hire	8.0	7.0	5.0–8.0	82 of 120	Top-25% edge
F07 Schedule crisis	6.0	7.0	5.0–8.0	84 of 120	Middle 50%
F08 Feedback	2.0	6.0	5.0–7.0	1 of 120	Bottom 25%
F09 Safety issue	8.0	7.0	5.0–8.0	89 of 120	Top-25% edge
F10 Resource shortage	7.0	7.0	5.0–8.0	86 of 120	Middle 50%
F11 Info security	1.0	6.0	5.0–7.0	5 of 120	Bottom 25%
F12 Decision delay	6.0	6.0	5.0–8.0	80 of 120	Middle 50%

Rank definition · ties share a rank and the next rank skips (competition ranking) · denominator = all 120 members incl. borderline · top % = rank ÷ 120

Axis ranks are exact from roster source values (2 dp) · competency/situation positions are band judgments from published quartiles and band counts (member-level scores not published)

Overall mean 6.1/8 sits below the cohort median 6.3, inside the 6–8 band (84 of 120, 70.0%)