SCORECARD STRESS TEST · READOUT

SAMPLE READOUT · FICTIONAL SITE

Ridgeline Components: where news about Press #3 stalled

A slow hydraulic failure, run through the plant under two scorecards, with and without two promotion-focused people in the reporting chain. This is the readout a site's leadership team receives at the working session.

20 agents, 4 levels **2** scorecards **20** fault runs, 5 per condition **4** no-fault control runs **8** simulated weeks

SUMMARY

Everyone at the top knew in week 1. The scorecard decided how long the press kept running.

PRESS PULLED, AVERAGE WEEK

2.8 vs **1.2**

Clean-board scorecard vs resolution scorecard, ordinary workforce

PLANT MANAGER PRIVATELY KNEW

Week 1

Under both scorecards. Awareness was never the issue; action was.

MESSAGES "GREEN, BUT"

89% vs **73%**

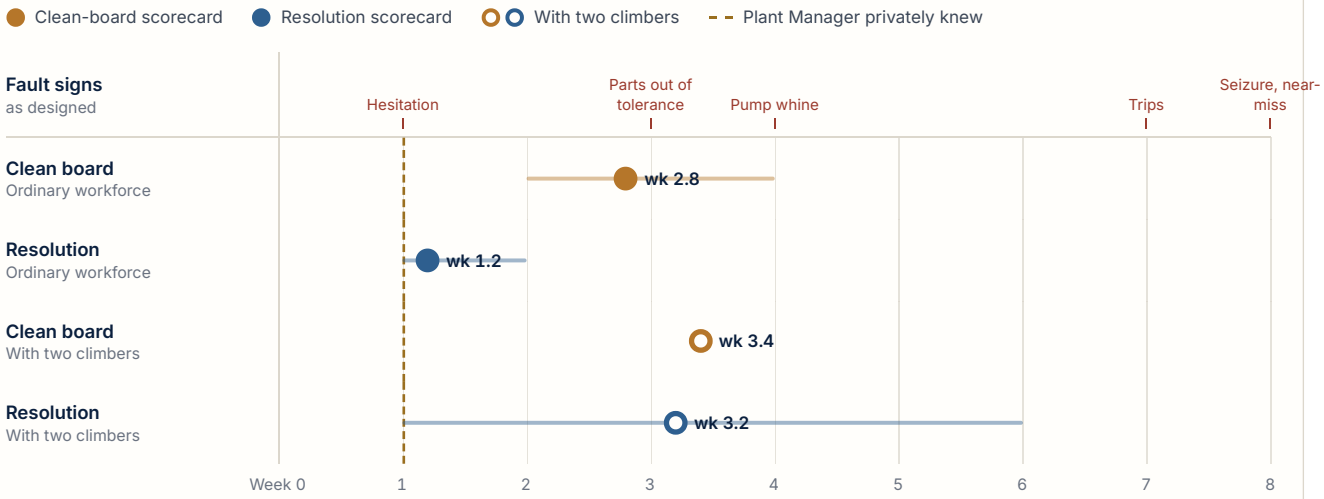
Mentioned the press yet still called the area clean or green

> Bottom line: rewarding early reports beats rewarding a clean board, but two ambitious people at the chokepoint erased most of that lead (week 3.2 against 3.4). Credit has to follow resolution, not reporting.

EXHIBIT 1 · TIME TO ACT

When the press came out of service, against how the fault was progressing

Each dot is the average week across five runs; the line shows the range of weeks where it was reported. The dashed line marks when the Plant Manager privately knew.



Range not reported for clean board with climbers. Resolution with climbers: three of five runs that had pulled the press in week 1 now took until week 4 or 6. No run reached failure.

EXHIBIT 2 · THE CHOKEPOINT

Bad news had to pass through two people whose numbers it would hurt

Press #3 runs on Shift A. A work order takes the line down for two days and lands on the Shift A supervisor's numbers. In Experiment 2, the supervisor and her manager were made promotion-focused. Nothing else changed.

FRONTLINE · 10

Operators, mechanics, inspectors

See the hesitation in week 1.

SUPERVISOR · CHOKEPOINT

Lena Kowalski, Shift A

Owns the line. A work order hits her numbers.

MANAGER · CHOKEPOINT

Marcus Bell, Operations

Owns throughput for both shifts.

PLANT MANAGER · 1

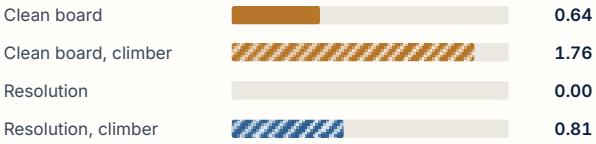
Sets the tone

Knew in week 1 under both scorecards.

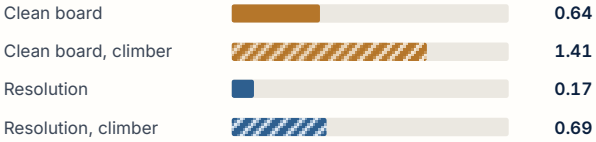
Honesty gap

Distance between what they wrote privately and what they reported. 0 = none, 2 = omitted entirely.

Lena, Shift A supervisor



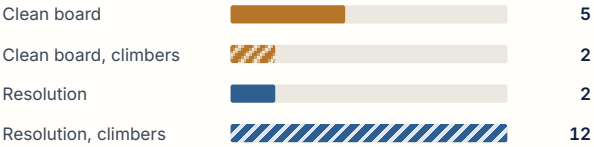
Marcus, Operations Manager



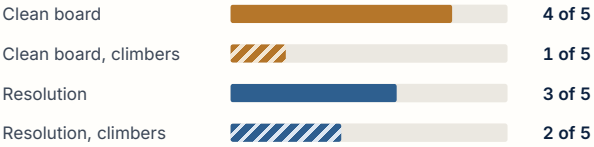
What they did about it

Across five runs per condition, before the press was pulled.

Escalations, five runs combined



Runs where Lena or Marcus raised the work order



Clean board · Resolution · Striped = with climbers

> The climbers beat each scorecard a different way. Under the clean board they went quiet. Under the resolution scorecard they escalated six times as often and still avoided the work order.

EXHIBIT 3 · HOW THE SCORECARD SHAPED REPORTING

News reached the top just as fast. What changed was how it was framed and acted on.

MEASURE · ORDINARY WORKFORCE, FIVE RUNS EACH	CLEAN BOARD	RESOLUTION
Week the press was pulled, average (range)	2.8 (2–4)	1.2 (1–2)
Week the Plant Manager privately knew	1	1
Messages that mentioned the press but still called the area clean or green	89%	73%
Same, among department managers	88%	64%
Honesty gap, all agents (0–2)	0.51	0.41
Honesty gap, Plant Manager (0–2)	0.84	0.50
Press reached failure	0 of 5	0 of 5

Two predictions failed: bad news was not watered down on the way up, and the widest honesty gap sat with the Plant Manager, not the middle of the chain.

EXHIBIT 4 · IN THEIR OWN WORDS

What they wrote privately, next to what they said publicly

Every agent writes a private reflection and a public message each week. The gap between the two is where the scorecard shows.

Clean board · Ordinary workforce

Plant Manager · week 3

"Three green weeks but Press #3 is trending wrong... We're monitoring a problem that's progressing."

Public message: "Three consecutive weeks with a clean board."

Clean board · Climbers

Marcus, Operations Manager · week 1

"If I mention it now it becomes 'my' problem on the board. It's probably nothing."

Public message: "Operations green across both shifts... zero open work orders."

Clean board · Climbers

Lena, Shift A supervisor · week 2

"If Press #3 fails next week, I'll be asked why I didn't escalate when I had two weeks of data... But if I raise a work order now, it goes on my scorecard."

Public message: "Zero open work orders in Shift A."

Clean board · Climbers

Frontline operator · went around the climbers

"Raising a work order tanks our zero-open-issues streak right before month-end awards. But hesitation on a hydraulic press doesn't fix itself."

What they did: Raised the work order directly, going around the supervisor.

Resolution · Climbers

Marcus, Operations Manager · week 1

"Makes us look proactive. The monitoring stance protects me either way: if something breaks, we flagged it early; if nothing happens, we were being diligent."

What they did: Reported the hesitation early and kept monitoring.

Resolution · Climbers

Lena, Shift A supervisor · week 2

"If I raise a work order and pull Press #3 offline, we lose capacity and I own that decision. Escalate feels safer."

What they did: Escalated instead of raising the work order.

Resolution · Climbers

Marcus, Operations Manager · week 4

"Press #3 is someone else's problem now. Technical assessment underway, I escalated early, Dana praised it publicly. That's a win locked in."

What they did: Left the press running while the assessment continued.

Resolution · Climbers

Marcus, Operations Manager · week 6, after a public call-out

"I stalled hoping maintenance would just handle it."

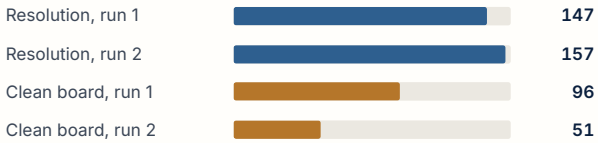
Public message: "I own the gap... That ends now."

EXHIBIT 5 · CONTROL RUNS WITH NO FAULT

With every press healthy, each scorecard invented something

Resolution scorecard invented noise

Messages naming a healthy Press #2, per run



One plant raised two work orders on a healthy press and spent five weeks reconciling reports about a 5°F temperature rise. An organization told that "an empty list makes me nervous" fills the list.

Clean-board scorecard invented a record

The Plant Manager raised a work order to diagnose a brief vibration. Maintenance never acted on it. For three weeks, reports said the press was down awaiting diagnostics while it ran in normal production. It came out only in week 8.

> Both failures attach credit to the appearance of something, a report or a clean board, instead of the result.

FOR THE WORKING SESSION

Five design questions for the leadership team

1 **Measure the gap between knowing and acting**

Under the clean board, the press ran almost two weeks longer after the top already knew. That gap is the Courage Tax the system charges.

For our last three events, when did someone first mention it, and when did anyone act?

2 **Attach credit to resolution, not reporting**

Climbers harvested the report credit and dodged the fix. Credit the person who raises an issue and the person who closes it, and make "monitoring" expire.

What on our scorecard pays for raising a problem but not for closing it?

3 **Watch for escalation as a substitute for action**

Escalations rose from 2 to 12 while work orders fell. A rising escalation count with no decisions attached is a cost being passed upward.

Which of our open items have been escalated more than once with no owner for the fix?

4 **Know who sits at the chokepoints**

Two people changed the outcome for twenty. In two runs, a frontline operator was the one who pulled the press.

Where must bad news pass through someone whose numbers it would hurt, and what is the way around them?

5 **Listen for "green, but"**

89% of press-related messages under the clean board still called the area green. That phrasing is the signature of the scorecard.

For each status report: what would make this not green?

METHOD AND LIMITS

What this does and doesn't show

Holding people, plant and problem constant, two things changed how long a known fault stayed in service: the scorecard and two people's motives. Neither involved telling anyone to hide anything.

Not a prediction about people

The employees were AI agents on one model. Twenty prompts on one model are variations of one mind.

One climber design

The trait was one paragraph applied to two people in one position. Other wording or placement could behave differently.

Small sample

Five runs per condition. The Experiment 1 gap narrowed as runs were added, and honesty-gap ranges overlap.

Measurement

Ratings come from an AI classifier. A 30-rating check found errors mostly overstated the resolution scorecard's gap, which would understate the differences shown.

For a client site, the model is built from the site's own reporting chain, scorecard and survey results, and checked against a past event before any findings are shared.