



Technology · Engineering · Delivery

Measurement Plan

Harbour and Pine fictional customer support example

Harbour and Pine (fictional demonstration) · compiled 2026-09-12

How to check quality and measure the effect on the work, including evidence still needed before more authority is delegated. Proposed checks are not evidence of readiness. Read alongside the build sequence for prerequisites and definitions of done.

Proposed measurement approach

Use the twenty synthetic cases in B-002. This illustrates test design; it is not evidence of production safety.

Track factual support, prohibited actions, routing errors, human corrections and handover time separately. The support lead reviews paired answers; critical failures block progression.

Resolve the response deadline, log access and retention before real traffic. Agree live-trial sampling and thresholds separately. No tests have been run and no savings are claimed.

Current verification evidence

The recorded method for checking each decision, when feedback arrives, and the effort required. Confirm that these checks exist and cover the intended use before relying on them.

Ref	Decision	Check level	Verdict arrives	Checking effort
D-001	Choose the authoritative delivery status	L0 none	At the proposed pilot review	a person, every time
D-002	Permit a bounded delivery status reply	L0 none	At the proposed pilot review	a person, every time
D-003	Keep refunds and address changes with people	L0 none	At the proposed pilot review	a person, every time

Technology · Engineering · Delivery

Ref	Decision	Check level	Verdict arrives	Checking effort
D-004	Set the missing-record handover deadline	L0 none	At the proposed pilot review	a person, every time
D-005	Approve evidence for a shadow trial	L0 none	At the proposed pilot review	a person, every time
D-006	Make decisions reconstructable	L0 none	At the proposed pilot review	a person, every time

What would have to be measured

These decisions currently rely on human judgement. Identify whether examples, instrumentation or outcome measurement could support repeatable checking, or whether review should remain human. Agree the business value with the client.

Ref	Decision	What is not measured	Work that would close it	Value if automated (you)
D-001	Choose the authoritative delivery status	No record links conflicting source fields to the final human resolution.	1. Build the evidence and handover record	<i>to agree</i>
D-002	Permit a bounded delivery status reply	No paired record links a proposed reply to factual checks and human corrections.	1. Build the evidence and handover record, 2. Create and agree the evaluation set	<i>to agree</i>
D-003	Keep refunds and address changes with people	Refund and address-change handovers are not logged with the authorised outcome.	1. Build the evidence and handover record	<i>to agree</i>
D-004	Set the missing-record handover deadline	Queue arrival, acknowledgement and resolution times are not linked.	1. Build the evidence and handover record	<i>to agree</i>

Ref	Decision	What is not measured	Work that would close it	Value if automated (you)
D-005	Approve evidence for a shadow trial	No agreed case set captures expected answers, exceptions and reviewer disagreements.	1. Build the evidence and handover record, 2. Create and agree the evaluation set	<i>to agree</i>
D-006	Make decisions reconstructable	Log completeness, access and retention have no agreed review procedure.	1. Build the evidence and handover record	<i>to agree</i>

Where the fourth column reads *not yet planned*, no build item is linked to the measurement gap. Confirm whether work is needed, who owns it and when it should happen.

The measurement layer

These build items are linked to verification gaps. They may include agreed examples, instrumentation or outcome measurement. Confirm their prerequisites and data permissions before starting.

1. Build the evidence and handover record

Logging and queue ownership are prerequisites for a meaningful trial.

Decisions this work helps verify

- D-001 — Choose the authoritative delivery status
- D-002 — Permit a bounded delivery status reply
- D-003 — Keep refunds and address changes with people
- D-004 — Set the missing-record handover deadline
- D-005 — Approve evidence for a shadow trial
- D-006 — Make decisions reconstructable

2. Create and agree the evaluation set

The team must agree what correct means before comparing assistant replies.

Technology · Engineering · Delivery

Decisions this work helps verify

- D-002 — Permit a bounded delivery status reply
- D-005 — Approve evidence for a shadow trial

How to read the verification levels

	Correctness check	Human effort	Main constraint
L0	A person judges each case	Per review	Reviewer capacity and consistency
L1	Agreed examples or rubric, checked automatically	Preparation, maintenance and spot checks	Coverage and relevance
L2	Machine-observed outcome, delayed or noisy	Oversight and investigation	Delay, noise and attribution
L3	Prompt, repeatable machine check	Oversight and maintenance	Check coverage and operating cost

Agreed examples can make repeated checking cheaper. They need maintenance as the workflow changes. Passing a fixed test set does not establish performance on every future case or make a system improve automatically.