AU0010 / LEVEL 5

AI leadership – AI adoption, procurement and governance

Test before you buy

03

BLOCK

Benchmarking, user validation and decision-quality evidence evidence

180 minutes
+ 10-minute break

Skills England Apprenticeship Unit
Version 1.0 • 30-hour programme

OBJECTIVES

What you will be able to do

Workbook output: C – Testing, benchmarking & user validation

01

Design tests and benchmarks that answer a procurement decision.

02

Include user validation, accessibility and workflow consequences.

03

Interpret mixed evidence and challenge averages that hide high-consequence failures.

LO2, LO7 • K5, K7, K11, K13 • S3, S4, S10, S11

TODAY'S JOURNEY

Learn → interrogate → decide → document → apply

apply

What evidence would make a benchmark or user test decision-worthy?

LEARN

Get just enough input to make the decision.

INTERROGATE

Challenge the evidence, claims and gaps.

DECIDE

Make a conditional leadership judgement.

DOCUMENT

Record the decision and conditions in the workbook.

Use non-sensitive examples. The workbook is developmental, not the formal skills test.

VOTE → EXPLAIN

Opening metric challenge

“96% accurate.” Good enough?

Vote before seeing the detail. What would make this evidence decision-worthy?

TUTOR INPUT

What a useful benchmark must answer

Purpose

Which decision will the test support?

Baseline

What are we comparing against? against?

Representative data

Which cases, subgroups and edge cases are included?

Decision criteria

What threshold changes the recommendation?

TEST DESIGN CLINIC • 8 MIN

Design the benchmark before seeing results

01

Define the task and success measure.

02

Choose a baseline and hold-out data.

03

Include high-consequence cases.

04

Set stop/modify/continue thresholds.

Output: two thresholds and one failure condition

CURRICULUM READY

ROLE-BASED • 7 MIN

User validation is not just satisfaction

01

Adviser: what effort changes?

02

Supervisor: what do you need to monitor?

03

Accessibility reviewer: what must work without friction?

04

Customer lead: what outcome outcome must not worsen? worsen?

Output: one user-evidence requirement

CURRICULUM READY

FAILURE HUNT • 6 MIN

Where could the average hide failure?

01

Find a subgroup risk.

02

Find a workflow-recovery risk.

03

Find a human-oversight burden.

04

Find an accessibility or confidence-display issue.

Output: one failure that would change the decision

CURRICULUM READY

10-MINUTE BREAK

When we return: NORTHSTAR REVEAL 03

“96% accurate” — good enough?

Break is additional to the 180 eligible delivery minutes.

“96% accurate” — good enough?

Use only the information revealed in this block.

What evidence would make a benchmark or user test decision-worthy?

Do not read ahead. Future evidence is deliberately hidden.

Stage A headline

Vote first: Continue / Modify / Stop / More evidence required.

NexaRoute: “96% accurate on Northstar’s benchmark.”

Stage B — raw benchmark results

Route	Routine	Complex	Urgent	Overall
Manual	873/900 97.0%	202/220 91.8%	77/80 96.3%	1,152/1,200 96.0%
NexaRoute	891/900 99.0%	202/220 91.8%	59/80 73.8%	1,152/1,200 96.0%
HarbourPeak	864/900 96.0%	209/220 95.0%	76/80 95.0%	1,149/1,200 95.8%
SmartRoute	855/900 95.0%	198/220 90.0%	78/80 97.5%	1,131/1,200 94.3%

After review and user evidence

Evidence	NexaRoute	HarbourPeak	SmartRoute
After human review	98.7%; urgent 92.5%	99.0%; urgent 98.8%	98.0%; urgent 100%
Total handling time	3.1 min	3.6 min	4.1 min
User trial	19/24 positive	18/24 positive	15/24 positive
Observed issue	anchoring + keyboard issue	vague explanations	more correction effort

Decision challenge

Did the headline hide a high-consequence weakness?
Which option meets current progression gates?
What human review burden are you accepting?
What evidence would still block procurement?

CASE DECISION BOARD • 7 MIN

What would you do now?

01

Name the strongest evidence.

02

Name the most important unknown.

03

Make a conditional recommendation.

04

Record the decision in the workbook Northstar record. record.

Output: conditional decision + evidence gap

YOUR TURN

Workbook C – Testing, benchmarking & user validation

Apply the same decision logic to your organisation.

Worked example

Skim the Westbridge example to see the expected level of thinking.

Your organisation

Use anonymised, non-sensitive information. Unknown is valid.

Challenge & refine

Use peer questions to improve evidence, conditions and assumptions.

Decision record

Capture the Northstar decision separately from your workplace application.

The workbook is developmental learning. It is not the formal skills test.

INDIVIDUAL • 12 MIN

Build your C section

01

Complete the key output fields.

02

Mark assumptions clearly.

03

Avoid confidential names or customer details.

04

Leave uncertain areas as Unknown / follow-up.

Output: first draft of this workbook part

PAIR COACH • 6 MIN

Challenge one assumption

01

Explain one decision or rating.

02

Partner asks: what evidence supports it?

03

Partner asks: what would change it?

04

Revise, strengthen or mark as as unknown.

Output: one improved evidence statement

SMALL GROUP • 9 MIN

Evidence clinic

01

Each learner shares one decision.

02

Group identifies a missing evidence point.

03

Group names a risk or condition.

04

Learner records one refinement.

Output: evidence gap + condition

INDIVIDUAL • 8 MIN

Refine and record

01

Return to your workbook.

02

Update the decision, condition or evidence gap.

03

Record one follow-up action.

04

Prepare one sentence to share.

Output: refined workbook entry

DEBRIEF

Decision revisit

What evidence would make a benchmark or user test decision-worthy?

Before	Now	Still unknown	Action
What was your first instinct?	What evidence or condition changed your view?	What must be investigated before commitment?	What would you do next in your organisation?

RECAP

Block 03 principles

From testing evidence to acquisition route.

01

A headline metric is a starting point.

02

Test design determines decision quality.

03

Human review has a cost and a role.

04

Subgroups and consequences matter.

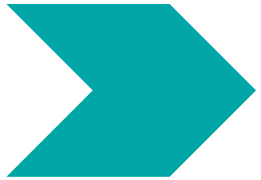
NEXT

Bridge to Block 04

Choose the acquisition route

Block 03

What evidence would make a make a benchmark or user user test decision-worthy? worthy?



Block 04

Choose the acquisition route route

Bring your workbook notes forward. Later decisions should change when better evidence appears.