MODULE 1 · SESSION 3

Run Usability Tests Before Release

How to find out whether real users can complete the job — before the cost of being wrong multiplies.

Theory session

Beginner

Running case: LumaCrate

ORIENTATION

Where this session sits

1.2 built the prototype; 1.3 finds out whether anyone else can use it

MODULE 1
Product Build and Validation

MODULE 2
MVP Economics and Measurement

MODULE 3
Sales Launch and Operations

SESSIONS IN THIS MODULE

1.1 Turn Product Requirements into Build Specs
From customer-informed needs to testable requirements

1.2 Shape a Prototype Around Critical Flows
Building only what answers the riskiest question

1.3 **Run Usability Tests Before Release** — YOU ARE HERE
Finding out whether real users can actually complete the job

SESSION OUTCOMES

What this session explains

Running a test that produces findings rather than reassurance

01

Usability is three measurable things

Effectiveness, efficiency and satisfaction — each observed differently.

02

Small samples find most problems

Why five users is usually the right number, and when it is not.

03

Task design decides what you learn

A leading task produces a passing result and no information.

04

Findings must be triaged

A list of every issue is not a plan; severity and frequency decide order.

DEFINITION

What usability means, precisely

The standard definition is deliberately narrow and measurable

WORKING DEFINITION

> Usability is the extent to which specified users can achieve specified goals in a specified context of use with effectiveness, efficiency and satisfaction.

1

Effectiveness

Did they complete the task, and was the result correct? Measured as a success rate, not an impression.

2

Efficiency

What did completion cost them — time, steps, errors, requests for help?

3

Satisfaction

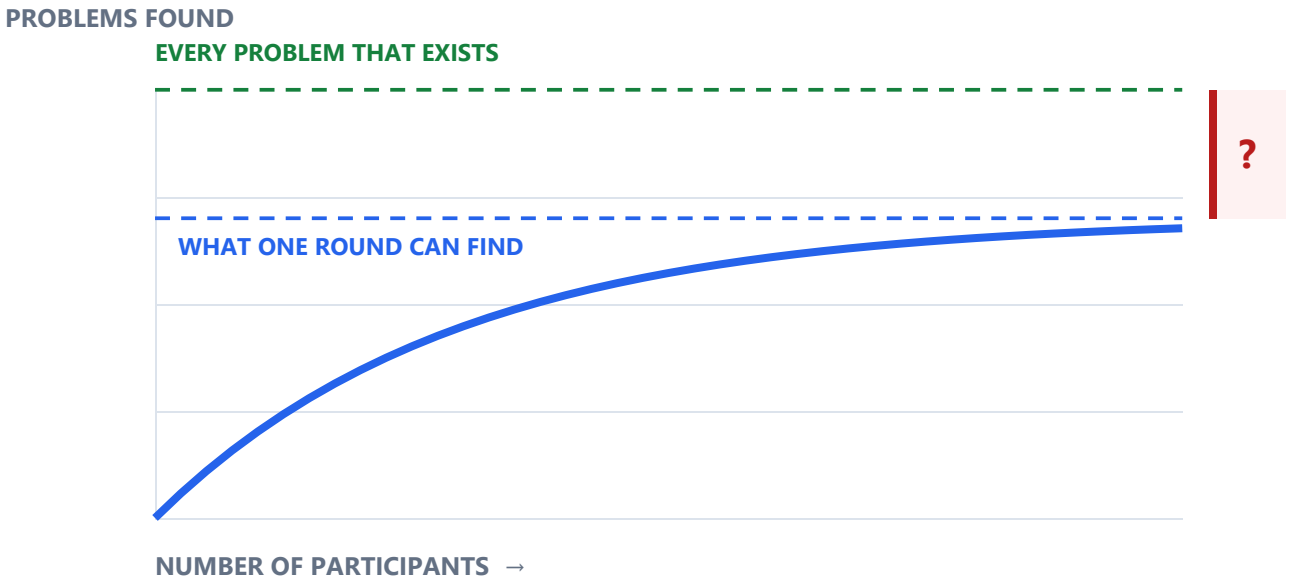
How acceptable was the experience? Measured with a consistent instrument so scores can be compared.

ISO 9241-11:2018. Ergonomics of human-system interaction — Usability: Definitions and concepts.

CONCEPT 1 · SAMPLE SIZE

Why five users finds most of it

Problem discovery saturates quickly, then flattens



Around five participants surfaces the large majority of problems in a single round; further participants mostly re-find the same ones.

THE IMPLICATION

Test more often, not bigger

Three rounds of five, with fixes between, finds far more than one round of fifteen.

WHEN FIVE IS NOT ENOUGH

Distinct user groups

Each genuinely different group needs its own participants. ASSMANN used twenty-three across a broad range.

CAUTION

Counting ≠ measuring

Five users find problems. They cannot establish a reliable success rate — that needs a larger sample.

Nielsen, J. & Landauer, T. (1993). A mathematical model of the finding of usability problems. Proceedings of INTERCHI '93, 206–213.

CONCEPT 2 · PURPOSE

Formative and summative testing

One improves the design; the other judges it. They are run differently.

FORMATIVE

Testing to improve

WHEN

During design, repeatedly, while change is still cheap.

OUTPUT

A ranked list of problems and the design changes they imply.

SAMPLE

Small — around five per user group, per round.

SUMMATIVE

Testing to judge

WHEN

At a release gate, or to compare against a benchmark or a competitor.

OUTPUT

Metrics: success rate, time on task, error rate, satisfaction score.

SAMPLE

Larger, because the numbers must be stable enough to act on.

WHICH ONE YOU NEED

Most pre-release testing should be formative. Running a summative test on a design nobody has yet improved measures something you already know is unfinished.

CONCEPT 3 · TASKS

Writing a task that can actually fail

The wording of the task decides whether the test can produce a finding

Rule	Weak version	Stronger version
Give a goal, not steps	“Tap Settings, then Devices, then Add.”	“You have just unboxed the camera. Set it up so you can see your hallway.”
Avoid the product's own words	“Create a new zone.”	“Group the two upstairs sensors so they act together.”
Give a realistic reason	“Explore the app.”	“You are going away for the weekend and want to be alerted if a door opens.”
Make success observable	“See if it works.”	“Stop when you can see the live view on your phone.”

THE TEST OF A GOOD TASK

If you cannot state in advance what counts as completing the task, you will not be able to say afterwards whether it was completed.

CONCEPT 4 · RUNNING THE SESSION

Three rules that keep the data clean

Each one prevents the moderator from destroying the finding

1

Do not help

The urge to rescue a struggling participant is strong and ruins the observation. Silence is the instrument. Wait.

2

Ask them to think aloud

“Tell me what you are looking for.” It reveals the mental model — which is where the design problem usually lives.

3

Record behaviour, not opinion

What they did and where they hesitated is data. What they say they would do later is not.

Ericsson, K. & Simon, H. (1984). Protocol Analysis: Verbal Reports as Data. Cambridge, MA: MIT Press.

CONCEPT 5 · RECRUITMENT

Who you test with decides what you find

ASSMANN's choice of participants is the whole lesson of this case

CONVENIENT

People close to the team

WHO THEY ARE

Colleagues, friends, early enthusiasts, other technical people.

WHAT HAPPENS

They recover from confusion using knowledge your real users do not have.

RESULT

A clean-looking test and a product that fails in the field.

REPRESENTATIVE

People like the intended user

WHO THEY ARE

Matching the real range of ability, motivation and context — including the unenthusiastic.

WHAT HAPPENS

They stop where real users stop, for the same reasons.

ASSMANN

Including testers with low to moderate interest in technology is what exposed the connection failure.

THE UNCOMFORTABLE RULE

The people most likely to reveal a problem are the ones least invested in the product succeeding.

CONCEPT 6 · SCOPE

Feature testing and end-to-end testing

The ASSMANN failure lived in the join between two components, where neither team was looking

FEATURE TESTING

One component at a time

WHAT IT CHECKS

That each part behaves as specified, in isolation.

STRENGTH

Fast, repeatable, easy to automate and to assign.

BLIND SPOT

Every handover between components — which is where integration failures live.

END-TO-END TESTING

The whole journey

WHAT IT CHECKS

That a real person can get from nothing to value, crossing every boundary.

STRENGTH

Finds failures that no component owner is responsible for.

ASSMANN

The camera would not connect to the app — a failure that belonged to neither side alone.

WHERE DEFECTS HIDE

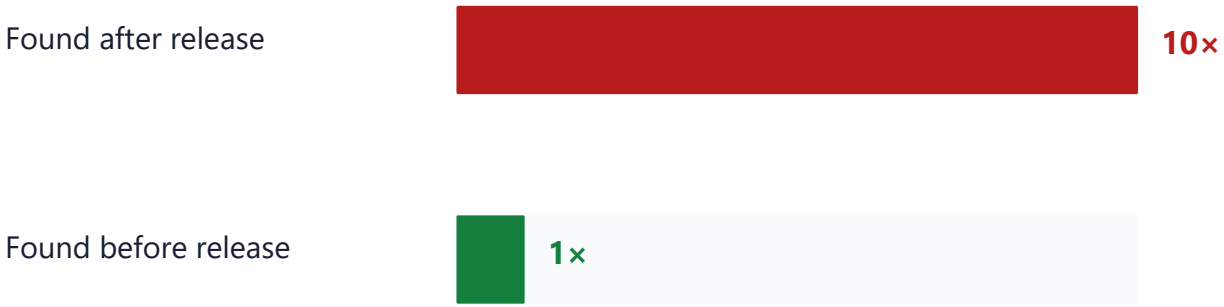
Every product with more than one component has a seam. Pre-launch testing that never crosses it is testing the parts you were already confident about.

CONCEPT 7 · TIMING

What it costs to find a defect late

The multiplier is why pre-launch testing is an economic decision, not a quality ritual

RELATIVE COST TO FIX, BY WHEN IT IS FOUND



The ASSMANN case reports roughly a tenfold difference — consistent with long-standing findings in software engineering.

Cost-escalation findings originate with Boehm, B. (1981), Software Engineering Economics, and are widely replicated.

WHY IT MULTIPLIES

The work repeats

A late fix means re-testing, re-releasing, supporting confused customers and sometimes recalling units.

THE OTHER COST

Customers leave

Reported in the case: 64% of consumers would switch over usability issues. That loss is not recovered by fixing the bug.

THE IMPLICATION

Test before you can afford not to

The cheapest round of testing is the one that happens before anyone has shipped.

CONCEPT 8 · TRIAGE

From a list of issues to a plan

Severity combines how badly it blocks the user with how often it happens

CRITICAL

The user cannot complete the task at all, and there is no workaround. Fix before release, without exception.

SERIOUS

Completion is possible but costly — confusion, errors, or help required. Fix before release if at all possible.

MINOR

Irritating, recovered from quickly. Schedule it; do not let it displace the rows above.

WHAT EACH FINDING MUST CARRY

Field	Why it is needed
What was observed	The behaviour, not your interpretation of it. “Three of five stopped at the pairing screen.”
How many, out of how many	Frequency separates a quirk from a pattern.
Severity and why	Blocks completion, or merely slows it.
Suspected cause	Marked as a hypothesis, so the fix can be wrong without the finding being wrong.

CONCEPT 9 · FAILURE MODES

Three tests that produce reassurance instead of findings

Each one returns a clean result and leaves the problem in place

The guided tour

Looks like The moderator explains as the participant goes.

Why it fails Nobody can get lost, so nothing can be learned.

Fix Set the task, then stop talking entirely.

The friendly sample

Looks like Testing with colleagues or enthusiasts.

Why it fails They bring knowledge your users lack and goodwill your users will not.

Fix Recruit for the real range, including the uninterested.

The unactioned report

Looks like A thorough document, circulated and filed.

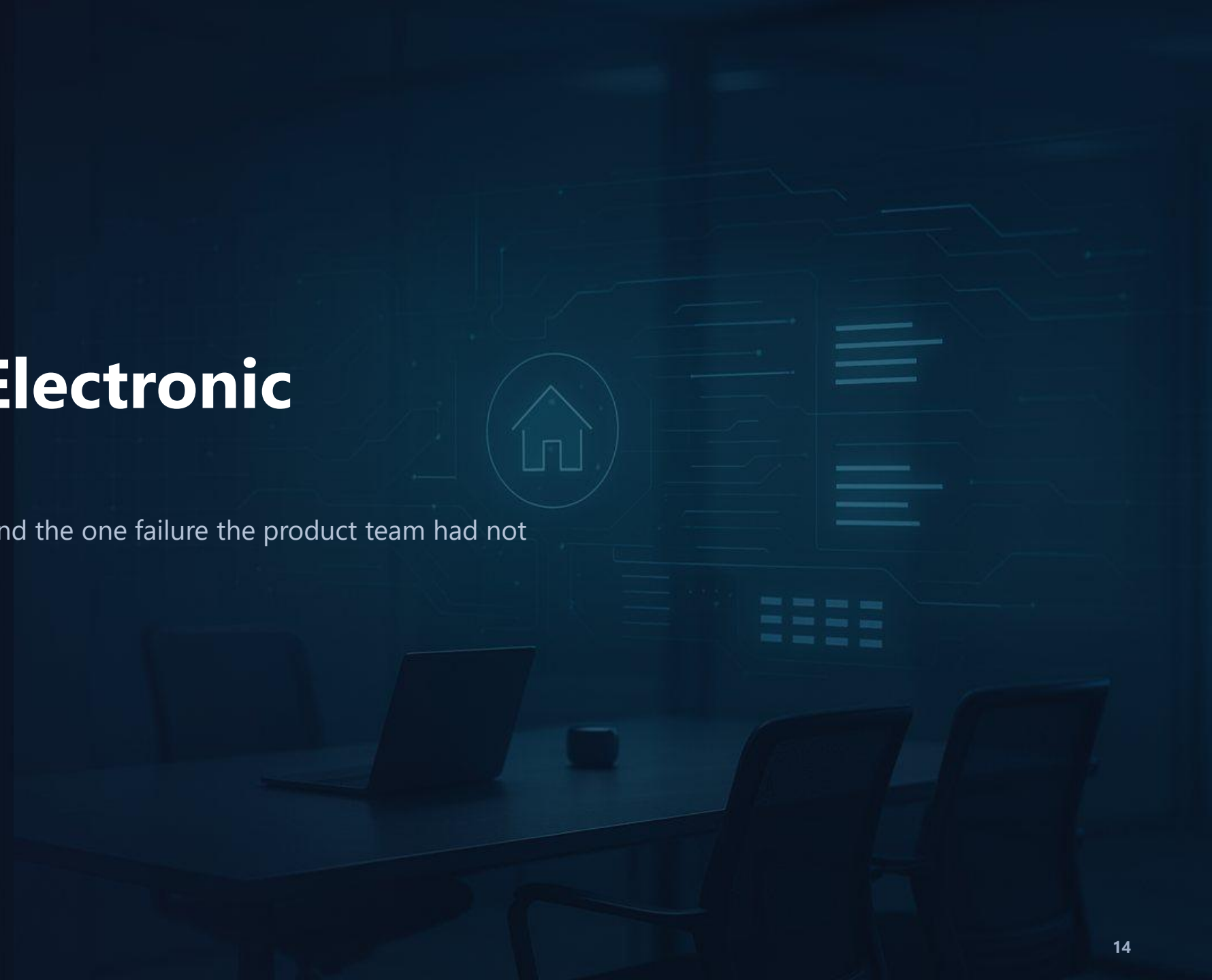
Why it fails Findings with no owner and no severity are not decisions.

Fix Triage, assign, and re-test the fixes.

CASE STUDY

ASSMANN Electronic

A pre-launch crowd test that found the one failure the product team had not thought to look for.



CASE STUDY · ASSMANN ELECTRONIC

The failure nobody was looking for

Broad recruitment and end-to-end scope are what made it visible

HOW THE TEST WAS RUN

End to end, not feature by feature

SCOPE

An End-to-End / BugAbility test covering the whole journey, not isolated features.

PARTICIPANTS

Twenty-three crowd-testers, deliberately including people with low to moderate interest in technology.

TASK

Search for functional and usability issues across the camera-and-app experience.

WHAT IT FOUND

The unexpected failure

THE DEFECT

Some testers could not connect the camera to the app at all.

WHY IT MATTERED

The failure was unexpected for the team — precisely the kind that scripted internal checks miss.

THE COMMERCIAL STAKE

Reported that 64% of consumers would switch over usability issues; late fixes cost around 10× early ones.

READING THE CASE

The value was not that a bug existed. It was that the bug sat outside what the team expected, and only a broad, unscripted, end-to-end test could surface it.

Drawn from the session case pack.

APPLICATION

Bringing the theory to LumaCrate

One round of testing, designed to be able to fail

Decision	What you choose	LumaCrate example
Purpose	Formative or summative.	Formative — we still intend to change the design.
Participants	The real range, not the convenient one.	Five people who have never used a labelling system, recruited outside the company.
Scope	End to end, crossing every seam.	From opening the box to the first successfully shelved item.
Tasks	Goals with observable completion.	“Set this up so you can find a specific item later.” Stop when they locate it.
Measures	Effectiveness, efficiency, satisfaction.	Completion rate, time to first shelf, number of requests for help.
Triage	Severity, frequency, owner, re-test.	Anything blocking completion for two or more participants blocks release.

WHAT COMES NEXT

Module 2 takes the validated product and asks what the smallest version worth shipping actually is.

SUMMARY

What to carry out of this session

1. **Usability is measurable, not a matter of taste.**
   Effectiveness, efficiency and satisfaction each have their own observation.

2. **Test small and often rather than large and once.**
   Around five users per group finds most problems; the value is in iterating between rounds.

3. **The task wording decides what you can learn.**
   Give a goal with observable completion, in the user's words, not the product's.

4. **Recruit the people least invested in your success.**
   ASSMANN's connection failure surfaced because low-interest testers were included.

5. **Cross the seams.**
   Integration points belong to no component owner, which is exactly why defects survive there until launch.

FURTHER READING

Sources for this session

Primary references behind each concept

DEFINITION	ISO 9241-11:2018. *Ergonomics of human-system interaction — Usability: Definitions and concepts.*
SAMPLE SIZE	Nielsen, J. & Landauer, T. (1993). A mathematical model of the finding of usability problems. *Proceedings of INTERCHI '93*, 206–213.
METHOD	Nielsen, J. (1993). *Usability Engineering.* San Diego: Academic Press.
PRACTICE	Krug, S. (2009). *Rocket Surgery Made Easy.* Berkeley: New Riders.
THINK-ALOUD	Ericsson, K. & Simon, H. (1984). *Protocol Analysis: Verbal Reports as Data.* Cambridge, MA: MIT Press.
MEASUREMENT	Sauro, J. & Lewis, J. (2016). *Quantifying the User Experience* (2nd ed.). Cambridge, MA: Morgan Kaufmann.
COST OF DEFECTS	Boehm, B. (1981). *Software Engineering Economics.* Englewood Cliffs: Prentice-Hall.