CYBEHAVE · SPECIALIST

Running Behavioural Experiments

Course workbook

A/B tests, pilots, and evaluating whether an intervention worked.

Build Behaviour. Build Culture. Build Resilience.

This workbook belongs to

Organisation and role

The intervention I am about to test, in one line

cybehave.com/training · © 2026 CyBehave. All rights reserved. For use alongside the Running Behavioural Experiments video course.

BEFORE YOU START

How to use this workbook

This is the specialist course on evaluation: how to know whether a security intervention worked, with evidence that would survive a sceptic, a finance director or a statistician. Choice Architecture for Security module 7 is the light introduction; this course is the full method. The workbook follows the ten videos in order, holds every worksheet the videos refer to, and ends with the quizzes and their answers. By the end you will have a pre-registration page, a power worksheet, a threat table, an ethics test and a one-page report filled in for your own next test.

Five ways to get the most from it

1. **Bring a real intervention.** Something you are about to run or roll out. Every section has a "Your next test" box.
2. **Fill in the pre-registration page before module 4.** Claim, primary, guardrails, baseline, design, sample, stopping rule, prediction, analysis. Date it and do not change it.
3. **Find your comparison group.** The single most useful thing you can do this quarter is put the comparison group's line on the slide beside your own.
4. **Take the quizzes.** Scenarios, not definitions. Check the key and read the explanation for anything you missed.
5. **Do one thing per module.** Write the claim. Pick the rung. Do the arithmetic. Write the kill criteria. Walk the threats. Price it per unit.

The course at a glance

Module	Title	Tools in this workbook
0	Welcome	The evidence ladder; meet Leah.
1	The Question and the Outcome	Testable-claim card; outcome sheet; baseline window; the pre-registration page.
2	Designs	Counterfactual check; design chooser; unit-of-assignment sheet; stepped-wedge planner.
3	Sample Size, Power and How Long	Power worksheet; stopping rule; machinery checks.
4	Running a Pilot	Pilot plan; site chooser; implementation log; kill criteria; pilot report.
5	Analysing the Result	Difference-in-differences sheet; effect-size translator; null checklist.
6	Threats to Validity	The threat table, before and after.
7	Ethics, Fairness and Consent	The ethics test; what people are told; simulation rules; fairness check.
8	From Result to Decision	Decision rule; cost-per-unit sheet; maintenance dose; one-page report; experiment register; HEROES map.

Module	Title	Tools in this workbook
9	Close	Eight ideas; one experiment this quarter; glossary; further reading; quiz answers.

Three lines you will hear again

No comparison, no conclusion. · Write the prediction before the data. · The control group keeps every protection.

WORKED EXAMPLE

Meet Leah

Leah Brennan is Head of Security Awareness at Ashcombe Retail Group: 18,000 people, 240 stores in the UK and Ireland, a head office of 1,100, two distribution centres (Daventry and Larne) and an online operation. The CFO has asked her to justify the awareness budget, and her best answer so far is a before-and-after chart. The course follows three pieces of work in one year, and ends with the CFO's question: "What did we buy, per report?"

Where the course takes it
The store trial (modules 1, 2, 3, 5, 6, 8): a "Report it" till-point card plus a two-minute manager huddle, cluster-randomised across 240 stores (120 to 120, stratified by region and size), twelve weeks. Primary: real suspicious-message reports per 100 staff per month. Result: +1.9 per 100 staff per month (1.3 to 2.5); simulated reporting +9 points (6 to 12); guardrail held; about £28 per additional report, nearer £19 at scale.
The head-office pilot that misled (modules 2, 4, 6): a password manager rolled out to 1,100 people with no comparison. +29 points on the slide; a competitor breach and a CISO email in week three; Daventry, which got nothing, rose 14. Honest estimate about +17. The IT log showed the manager pre-installed on only 70% of laptops.
The stepped-wedge rollout (modules 2, 5, 8): a tailgating challenge at three sites that could not be randomised, switched on one site every four weeks in an order drawn by lot. 41 to 14 per 1,000 entries at each site; decay to 22 by week sixteen without a dose; held at 15 with a four-weekly handover line.

Use Leah as a mirror

Each time an Ashcombe card appears, ask: what is my equivalent? What is my comparison group? Which rung am I really on? What did my last intervention cost per unit of behaviour?

The evidence ladder: where was your last evaluation?

Rung	What it is	What it can show	My last evaluation stood here?
5 Randomised comparison	Groups made the same by chance	The effect, within an interval	

Rung	What it is	What it can show	My last evaluation stood here?
4 Comparison group	Two groups, two times, difference in differences	Most threats handled; selection open	
3 Before and after	One group, two times	Confounded by everything in between	
2 Activity	Sessions, completions, opens	That the machine ran	
1 Opinion	"People liked it"	Useful for design; evidence of nothing	

MODULE 1

The Question and the Outcome

Key ideas from the video

- A testable claim: who, does what, by how much, by when, with the comparison in it. "More than before" is not a comparison; "more than stores that did not get it" is. If it cannot fail, it is not a claim.
- One primary outcome, chosen before the data. Then two or three secondaries (supporting, never deciding), guardrails with limits (what the intervention must not break), and everything else labelled exploratory.
- Measure the behaviour directly from systems (reports from the tool, verified payees, manager weekly use, tailgating per 1,000 entries, shared logins). Proxies (quiz scores, completion, self-report, click rate alone, satisfaction) only when nothing direct exists, labelled with their distance. The Team Pulse is the observed-behaviour source, independent of the programme.
- A baseline window of three to six months, week by week, to see the noise; the same instrument throughout; a prediction with direction, size and date, sealed with someone who is not you.
- The one-page pre-registration: claim, primary, secondaries, guardrails, baseline, design, sample, stopping rule, prediction, analysis. Dated before launch, unchanged after (Test, Learn, Adapt, 2012).

Template: the testable claim

Who	Does what (the behaviour)	By how much, compared with whom	By when

Write it as one sentence underneath, with the comparison in it.

Template: outcomes

Type	Measure	Source system	Direct or proxy (distance)	Limit (guardrails)
Primary				
Secondary 1				
Secondary 2				
Guardrail 1				

Type	Measure	Source system	Direct or proxy (distance)	Limit (guardrails)
Guardrail 2				
Exploratory				

Template: baseline and prediction

Baseline window (from, to)	Primary by week or month (range seen)	Instrument frozen? (tool, definition, denominator)	Prediction: direction, size, date	Sealed with

The pre-registration page (fill in across modules 1 to 5, date it before launch)

Line	Entry
Claim	
Primary outcome and source	
Secondaries	
Guardrails and limits	
Baseline window and figures	
Design and unit of assignment (module 2)	
Sample: units per arm, people per unit, design effect, power (module 3)	
Duration, stop date, number of looks (module 3)	
Prediction, and who holds it	
Analysis: difference in differences, intention to treat, pre-registered subgroups (module 5)	
Decision rule: scale / adjust / stop (module 8)	
Dated and signed	

Leah writes the claim

First draft: "Improve phishing reporting in stores." Second: "Stores that get the card and the huddle will report real suspicious messages at least 50% more often than stores that do not, within twelve weeks." Primary: real reports per 100 staff per month, from the tool. Secondary: simulated-phish report rate. Guardrail: help-desk tickets from false alarms up no more than 25%. Baseline, six months by week: 0.6 real reports per store per month, bouncing 0.3 to 1.0; simulated 8%.

Prediction, sealed with the CFO's analyst: "By week twelve, treatment stores at least 1.2 real reports per 100 staff per month above control, and simulated reporting at least 3 points higher." Note underneath: "If neither moves, we stop printing cards. If reporting rises and help-desk load passes 25%, we redesign before scaling."

Your next test

Write your claim in one sentence, name the primary, and write the prediction with a number and a date.

Reflect & apply

Pause the video and take ten minutes with these. Write, do not just think.

1. Write your next intervention as a testable claim: who, does what, by how much, by when. Could a colleague tell from the sentence whether it came true?

2. What is the one primary outcome, and is it the behaviour itself or a proxy for it? If a proxy, how far is it from the behaviour?

3. Which guardrail could the intervention break: help-desk load, time, trust, another behaviour? What is the limit?

4. What is the baseline for the primary outcome today, over what period, from which system?

5. Write the prediction now, with a number and a date. Who will hold the envelope?

MODULE 2

Designs: From Before-and-After to Randomised

Key ideas from the video

- The effect is the difference between what happened and what would have happened anyway. History, regression to the mean, seasonality and maturation all move an outcome on their own, and before-and-after cannot separate any of them from the intervention because it has one group.
- Random assignment makes the groups the same on everything, measured and unmeasured, on average; selection is gone; the difference is the intervention plus chance of a known size. Assign by lot, never by who asked first. Stratify (pair by region and size, flip a coin within each pair).
- The unit of assignment: a person for a screen prompt nobody discusses; a team, store or site whenever people talk, share a manager or see each other. Cluster randomisation costs power and buys a clean comparison.
- When everyone must get it: the stepped wedge (Hussey and Hughes, 2007). All sites start as controls; each switches on in an order drawn by lot; the order is fixed and respected; the outcome is measured everywhere every period; a time effect is modelled.
- When you cannot randomise: a matched comparison group (same way, same weeks; cannot rule out what made the treated sites get chosen) or an interrupted time series (twelve or more points each side, a sharp start). Honest when the limits are stated; dishonest when presented as randomised. Name the rung.

Template: the counterfactual check

Threat to a before-and-after reading	Could it apply to my intervention? How?	What a comparison group would do about it
History (an event in the window)		
Regression to the mean (I intervened because the number was bad)		
Seasonality (the calendar)		
Maturation (learning, drift, familiarity)		

Template: the design chooser

Question	My answer	What it implies
Can some people or sites not get it, for a while?		Yes: randomise. No: stepped wedge (rolling out) or matched comparison (already out)
Do people in the two groups talk, share a manager or see each other?		Yes: unit is the team or site
How many units do I have?		Under 8: pilot with matched comparison. 20+: cluster trial

Question	My answer	What it implies
Is the outcome measured the same way everywhere, every week, already?		No: fix this first
The rung I will stand on, and will name in the report		

Template: unit of assignment and stratification

Unit (person / team / store / site)	Why this unit (how the treatment would leak at a smaller one)	Stratify by	How assigned (lot, coin, random column), by whom, on what date

Template: stepped-wedge planner

Site	Order (drawn by lot)	Switch-on date	Outcome measured every period from	Who holds the order against the keen manager
	1			
	2			
	3			
	4			

Leah picks a cluster trial, and sees what the pilot hid

Store trial: unit the store (staff talk and share a manager); 240 stores paired by region and size, a coin per pair; 120 and 120; twelve weeks. Head-office pilot: a password manager to all 1,100 at once, no comparison, +29 points on the slide; a competitor breach and a CISO email in week three; Daventry, which got nothing, +14 over the same weeks.

What she wrote: "Head office rose 29. A site with no rollout rose 14. The honest estimate is about 17, matched comparison added after the fact, rung four." The spring tailgating rollout at three sites became a stepped wedge: order by lot, one site every four weeks, at no extra cost.

Your next test

Name your counterfactual, your unit of assignment and your rung.

Reflect & apply

Pause the video and take ten minutes with these. Write, do not just think.

1. For your next intervention, what is the counterfactual: what would the outcome have done if you had done nothing? How do you know?

2. Could you randomise? If yes, at what unit: person, team, site? If no, what is the honest reason?

3. Which design on the ladder is the highest you could actually run, and what would it cost you in time or coverage?

4. If people in the two groups talk to each other, how would the intervention leak, and what unit of assignment stops it?

5. What rollout is already planned that could become a stepped wedge for free?

MODULE 3

Sample Size, Power and How Long to Run

Key ideas from the video

- Base rate: how often the outcome happens now. The rarer it is, the more people or time you need. Minimum detectable effect: the smallest change that would change a decision, not the change you hope for; it fixes sample, duration and cost. Power 80%: four tests in five find a real effect of that size.
- Rough rules: 8% to 11% needs about 1,400 per arm; 8% to 16% about 250 per arm; halving the effect quadruples the sample. Rare events: about 50 events per arm for a doubling. Design effect = 1 + (cluster size minus 1) times the intraclass correlation: 60 staff at 0.05 is about 4, so 240 stores of 60 behave like 3,600 people, not 14,400.
- How long: a full calendar cycle; measure from week three or report the novelty bump separately; run to the horizon you will scale for; a stop date written in the pre-registration.
- Peeking: every look is another chance for noise to cross the line; twelve weekly looks pass a test of nothing about one time in three. One look at the stop date, or pre-planned looks with stricter thresholds. Guardrails may be watched weekly.
- Tests of the machinery (Kohavi, Tang and Xu, 2020): the A/A check, the sample ratio, baseline balance.

The power worksheet

Line	Entry
Primary outcome and base rate, over what window	
Minimum detectable effect, and the decision it would change	
Unit of assignment; units per arm; people per unit	
Intraclass correlation assumed; design effect = 1 + (n − 1) × ICC	
Effective sample per arm (people per arm ÷ design effect); power at the chosen effect (calculator)	
Duration: cycle covered; novelty excluded; decay horizon	
Stop date; number of looks; who checks the sample ratio	

If the worksheet says the test cannot see the effect, change the test (longer, bigger, a better outcome, an honest pilot), never the effect.

Template: the stopping rule and the machinery

Item	My rule
Primary looked at on (date), and not before	
Pre-planned interim looks, if any, and their thresholds	

Item	My rule
Guardrails watched weekly by	
A/A check run on the baseline with a fake split: result	
Sample ratio expected / observed at the end	
Baseline balance on the primary (arm A / arm B)	

Leah does the arithmetic

Primary base 1.0 per 100 staff per month (0.6 per store of 60); secondary 8%. Minimum detectable effect: 3 points simulated; 50% on real. Unit store, 120 per arm, 60 staff; ICC assumed 0.05 (found 0.04); design effect about 4; effective sample about 1,800 per arm; power for 8% to 11% about 85%. Four months of baseline gave 290 real-report events per arm. Twelve weeks, three simulation rounds, weeks one and two reported as novelty, stop date fixed, guardrails weekly, primary once.

A/A check: 0.3 points, interval through zero. Sample ratio at week twelve: 120 and 119 (a store closed). Baseline balance 8.1% against 7.9%. The regional director asked for "a quick look" at week four; he was shown the guardrail chart and the sealed envelope, in that order.

Your next test

Fill the seven lines of the power worksheet for your test. If it does not reach 80%, write what you will change.

Reflect & apply

Pause the video and take ten minutes with these. Write, do not just think.

1. What is the base rate of your primary outcome, and what is the smallest change in it that would change a decision?

2. How many people or sites do you have, and is that enough to see that change? If not, what would you change: the effect you chase, the duration, or the outcome?

3. If you randomise by team or site, how alike are people within a unit? What does that do to your effective sample?

4. How long must the test run to cover a full cycle of the calendar and outlast the novelty? Write the stop date now.

5. Who will be tempted to look early, and what will you say to them?

MODULE 4

Running a Pilot

Key ideas from the video

- A pilot asks "can we deliver it?": fidelity, dose, cost, reaction, what breaks. An experiment asks "did it work?" A pilot that "worked" is a pilot whose outcome moved. Fold them together and you learn neither.
- Sites: typical, not keen; two or three, not one; two or three matched comparison sites measured alongside, which costs nothing and turns an anecdote into a rung-four estimate.
- Fidelity (delivered as designed?) and dose (how much arrived?), in a dated implementation log per site, checked weekly. A null at 40% fidelity is an untested intervention; a result at 100% fidelity will shrink at 60%, and you will know why.
- Kill criteria written before: harm (a guardrail breached), cost (over the allowance), delivery (fidelity under 60% at week four). Who decides, by when, on what evidence. The qualitative signal: five minutes with a manager and three staff at weeks two and six, verbatim.
- The pilot report designs the trial; it does not claim an effect. If it is used to justify scaling, somebody skipped a rung.

Template: pilot plan

The delivery question this pilot answers	Pilot sites (typical on what?)	Comparison sites	Duration	What the trial will later measure, measured now too

Template: implementation log (one per site, weekly)

Week	Delivered as designed? (fidelity)	How much arrived (dose: huddles held / planned, cards visible, installs / scope)	What changed, who changed it, why	What broke
1				
2				
3				
4				
5				
6				

Template: kill criteria

Kind	Criterion and limit	Evidence that would show it	Who decides	By when
Harm				
Cost				
Delivery				

Template: the pilot report (one page, no effect claimed)

Section	Entry
What was piloted, where, how long, against which comparisons	
Fidelity and dose per site, from the log	
Cost per site: money, minutes, goodwill	
What staff and managers said (verbatim) and what changed	
Outcome, treated and comparison, labelled indicative	
Kill criteria approached / not	
Decision: redesign / trial / stop; what the trial needs	

The head-office pilot in hindsight

Site: head office, keen, atypical, one site, no comparison planned. Fidelity from the IT log, found afterwards: pre-installed on 70% of laptops; 40% of the rest never self-installed; half the building got the designed intervention. The qualitative signal, week seven: "It asked me to import passwords from the browser and I didn't trust it." Kill criteria: none. Cost: 210 service-desk tickets, discovered on asking.

Salvaged: Daventry as comparison, +17 honest, rung four. The reconstructed log became the redesign: pre-install 100%, remove the import prompt. The new rule for every pilot: three typical sites, two comparisons, the log from day one, kill criteria signed by the sponsor.

Your next test

Write the delivery question your pilot answers, your three sites, and your two kill criteria.

Reflect & apply

Pause the video and take ten minutes with these. Write, do not just think.

1. What is the question your next pilot answers that only a pilot can: can this be delivered, at what cost, with what reaction?

2. Which sites would you have picked by instinct, and which would a sceptic say you should pick instead?

3. What does full fidelity look like for your intervention, and how will you know what dose each site actually received?

4. Write two kill criteria: one for harm and one for cost.

5. What will you measure during the pilot that the trial will measure later, so the two can be compared?

the indicative outcome, kill criteria approached, and the decision to redesign, trial or stop

MODULE 5

Analysing the Result

Key ideas from the video

- Difference in differences: treatment change minus comparison change. Report both changes and their difference; after-only lies when the arms started unequal.
- Effect size in plain units: percentage points, per hundred staff, number needed. Give the absolute number beside any relative rise. A p-value is not an effect size.
- The confidence interval is the range the data allow. Read the ends: does the pessimistic end still justify the decision? An interval through zero is "could be nothing, could be this much". Cluster trials need the interval at the cluster level.
- Intention to treat (as assigned) leads; per protocol alongside. Pre-registered heterogeneity guides targeting; subgroups found afterwards are next year's hypothesis. Regression to the mean and survivorship are both answered by the comparison group.
- A null: check fidelity, power, measurement, then belief. A well-run null saves the scale-up budget. Trials say it worked here, at this dose, this year; not why, and not elsewhere (Deaton and Cartwright, 2018). Borrow a statistician for stepped wedges, proper clustered intervals and board-facing decimals.

Template: difference in differences

	Baseline	End	Change
Treatment arm			
Comparison arm			
Difference in differences (treatment change minus comparison change)			

Template: effect-size translator

Form	Your figure
Percentage points	
Events per 100 staff per month	
Number needed (staff exposed per one additional unit of behaviour)	
Relative rise (always beside the absolute)	
95% interval (cluster level if clustered), and what the pessimistic end would mean for the decision	
Intention to treat / per protocol	
Pre-registered subgroups / exploratory (labelled)	

Template: reading a null

Check	Finding
Fidelity: was it delivered, at what dose?	
Power: could this test have seen the effect I specified?	
Measurement: did the outcome capture the behaviour, both arms, same way, throughout?	
Belief: delivered, powered, measured, and still nothing. Decision: stop, and write up why	

The store trial reports

Week twelve, primary looked at once. Simulated reporting: treatment 8% to 19%, control 8% to 10%; difference in differences +9, interval 6 to 12 at store level; prediction at least 3, met. Real reports per 100 staff per month: 1.0 to 3.2 against 1.0 to 1.3; +1.9; prediction at least 1.2, met; one extra real report per 53 staff per month. Guardrail: help-desk +14%, inside 25%. Contamination: 11 control stores; intention to treat +9, per protocol +10; led with +9.

Heterogeneity pre-registered by size: over 80 staff +12, under 30 +5. By region, not pre-registered: Ireland +13, Scotland +6, labelled exploratory. Weeks one and two separately: +14, the novelty bump.

Your next test

Fill the difference-in-differences box for your last evaluation. If you have no comparison row, write what you can honestly claim.

Reflect & apply

Pause the video and take ten minutes with these. Write, do not just think.

1. For your last intervention, what was the change in the comparison group over the same period? If you do not know, what is the honest effect you can claim?

2. Write your result as an effect size in plain terms: percentage points, per hundred staff, or number needed. Then add the interval.

3. Who in your treatment group did not actually receive the intervention, and are they in your analysis? They should be.

4. Which subgroup result are you most tempted to headline, and was it pre-registered?

5. If the result is a null, what are the three explanations you will check before concluding the intervention does nothing?

MODULE 6

Threats to Validity

Key ideas from the video

- The checklist (Campbell and Stanley, 1963; Shadish, Cook and Campbell, 2002): history, instrumentation, selection, contamination, attrition, Hawthorne and novelty, decay, reactivity.
- Randomisation handles history, selection, similar attrition and Hawthorne when both arms are observed. It does not handle instrumentation, contamination or decay; those are yours to design out and log.
- Contamination shrinks the effect towards zero (report the estimate as a floor); instrumentation puts a step in the line (one instrument, frozen); attrition moves the line with nobody changing (leavers against stayers at baseline, per arm); the Hawthorne effect is real but poorly sized (McCambridge, Witton and Elbourne, 2014), so both arms observed and the bump reported separately.
- Decay is not a threat to the twelve-week estimate; it is a threat to the claim at fifty-two. Measure at the scale horizon; set the maintenance dose.
- When a threat landed: report the threat, the direction it pushes the estimate, and the bound. Never pretend it did not happen.

The threat table (fill in before launch, and again at the end)

Threat	Could it happen here? How?	What in the design handles it	Did it land? What did it do to the estimate? What we report
History		Concurrent comparison; event log	
Instrumentation		One instrument, frozen; parallel period if it must change	
Selection		Random assignment; baseline balance	
Contamination		Unit people do not cross; brief; log; intention to treat	
Attrition		Per arm; leavers against stayers at baseline	
Hawthorne and novelty		Both arms observed; from week three; report the bump	
Decay		Measure at the scale horizon; maintenance dose	
Reactivity		An outcome people cannot game; direct measure beside any proxy	

Template: event log (keep it open for the whole window)

Date	Event (incident, news, email, reorganisation, vendor change, new manager)	Which arms it touched	Which measures it could move

Leah walks the checklist

Landed: history (the breach and the CISO email; unhandled in the pilot, absorbed by the trial); instrumentation (vendor template change between rounds two and three; both arms; real-report primary unaffected; reported); contamination (11 control stores; logged the day she heard; +9 as a floor); attrition (9% turnover, 8.8 against 9.3; leavers 7.6% baseline reporting against stayers 8.1%; not differential; reported).

Did not land: selection (randomised, balanced), Hawthorne (same simulation schedule both arms; week-one bump +14 reported separately), reactivity (real reports cannot be gamed by clicking faster). Open: decay. Tailgating at Daventry decayed from 14 to 22 per 1,000 by week sixteen. A week-26 look on the store trial, pre-registered now.

Your next test

Fill the first three columns of the threat table for your test now, before launch.

Reflect & apply

Pause the video and take ten minutes with these. Write, do not just think.

1. Walk the eight threats against your last evaluation. Which three most plausibly explain the result instead of the intervention?

2. What changed in how the outcome was measured during the window: tool, template, definition, denominator?

3. How would your intervention leak from treated to control units, and did it?

4. Who left during the window, and were they different from those who stayed on the outcome?

5. When will you measure again, to see whether the effect decayed, and what will you do if it did?

MODULE 7

Ethics, Fairness and Consent

Key ideas from the video

- The first rule: the control group keeps every protection it has today. You test an addition (a card, a prompt, a message, a huddle, a default, timing, framing, dose), never the removal of a security control. If a control-group colleague were harmed, could you say they had everything they would have had anyway?
- Tell people before that the programme tests what it does, that some things are tried in some places first, that nothing in a test counts against anyone, and who to ask. Allocation and start date may stay quiet, with the reason written down and time-limited. Debrief everyone after; the control arm gets the intervention if it worked.
- Individual signed consent is rarely right at work; what replaces it is transparency about the practice, a real opt-out, consultation with representatives, and a design you could publish.
- Phishing simulations (NCSC): clicking is not a failure; punishing it teaches people to hide mistakes; the outcome is the report rate; the landing page thanks reporters; no names to managers; no humiliating lures.
- Data: aggregate to the unit of assignment; minimise, name who sees it, delete on a date; DPIA when high risk (individual tracking, profiling, decisions about a person); nothing below five people reported. Fairness: who does the design burden or leave out (shift, site, contractor, language)? Would you publish the allocation list?

The ethics test (a design passes all eight or does not run)

Question	Yes / no, and the evidence
Does the control group keep every protection it has today?	
Is the thing being tested an addition, not a removal?	
Have people been told that the programme tests what it does, and who to ask?	
Is any non-disclosure written down with its reason, and time-limited to the test?	
Can nothing anyone does in the test be held against them, and does every manager know that?	
Is the data aggregated, minimised, and deleted on a date?	
Has someone outside the team looked for the group the design treats worse?	
Will everyone, control arm included, hear the result and get the intervention if it worked?	

Template: what people are told

Before (by whom, when, the four points)	What stays quiet, why, until when	After: the debrief, to whom, on what date; control-arm rollout date

Template: simulation rules and fairness check

Item	My rule
Outcome measured (report rate, never click rate alone)	
The landing page text for a person who clicks	
Where click data lives by name, and who may never see it	
Lures excluded	
Group the design could burden or leave out (shift, site, contractor, language, reading level, device)	
Reviewed by (HR, DPO, staff forum, works council) and what they asked for	

Leah takes the trial through HR, the staff forum and the DPO

Control stores kept everything; only the card and the huddle were new. Told before by the retail director: "We test what we do. Some stores will see new things first. Nothing in a test counts against anyone." Allocation not announced, reason written down. Simulations: outcome the report rate; click data never left the vendor console by name; landing page rewritten to thank reporters; a regional manager's request for "the click list by name" refused, with the NCSC guidance attached. Data: store-level counts; deleted twelve months after the report; the DPO: no DPIA needed at this level, noted on file.

What the reviewers added: a one-minute handover version of the huddle for night-shift replenishment teams, in both arms; the card in Irish and English for the Irish stores; the control-arm debrief date fixed before launch (week thirteen, if the result cleared the prediction; it did).

Your next test

Run your design through the eight questions. Write the one you cannot yet answer yes to, and what would change it.

Reflect & apply

Pause the video and take ten minutes with these. Write, do not just think.

1. In your next test, what does the control group get? List every protection they keep. Is anything being withheld that reduces their safety?

2. What will people be told before the test, and what would be lost if they were told everything?

3. If the test involves simulated phishing, what happens to a person who clicks? Write the sentence they will read.

4. Which group could be treated less fairly by the design: a site, a shift, a contractor group, a language? What would you change?

5. Who outside your team has seen the design: HR, the DPO, a staff forum, the works council? What did they ask for?

MODULE 8

From Result to Decision

Key ideas from the video

- Three decisions, by a rule written before the result: scale (cleared the prediction, pessimistic end still worth it, guardrails held, fidelity realistic); adjust (effect present but a guardrail close, fidelity low, or a subgroup flat: change the thing the log points at and test the change); stop (a well-run null or an effect too small at the pessimistic end: stop spending and write up why).
- Cost per unit of behaviour: everything the intervention cost in the window, divided by the additional units of behaviour from the difference in differences. £4,800 and 170 additional real reports is about £28 each, nearer £19 at scale. Put the value of a unit beside it (the mean cost of an unreported incident).
- Decay is normal; unreported decay is the problem. Measure at the scale horizon; set and test the maintenance dose; budget the dose, not just the launch.
- The one-page report: claim and result in one sentence; the comparison and its rung; secondaries and guardrails; threats that landed; cost per unit and the decision; a footer naming the pre-registration date, who held the prediction, who ran the analysis, when the data are deleted.
- The evaluation habit: an experiment register including nulls; a standing control; every rollout a stepped wedge; one set of templates; a borrowed statistician an afternoon a quarter. In HEROES: the intervention record and its APEASE pass hold the design; Update Progress holds the pilot report, result and decision; the behaviour's direct measure and the Team Pulse are the outcomes; the board pack carries the behaviour line. HEROES does not randomise or run the statistics.

Template: the decision rule (write before launch)

Decision	Condition on the primary	Condition on guardrails and fidelity	What happens next
Scale			
Adjust			
Stop			

Template: cost per unit of behaviour

Line	Figure
Numerator: materials + staff time at loaded rate + your time + tickets generated (the window)	
Denominator: additional units of behaviour from the difference in differences (not the raw change)	
Cost per unit in the window	
Cost per unit at scale (fixed costs already spent)	

Line	Figure
Value of one unit beside it (e.g. mean cost of an unreported incident), and its source	

Template: decay and the maintenance dose

Measure again at (week)	Effect then	Maintenance dose (the smallest repeat that holds it)	Who delivers it, on what rhythm	Dose budgeted?

The one-page experiment report

Section	Your line
1 The claim and the result, one sentence: behaviour, effect, interval, comparison	
2 The comparison: design, unit, assignment, per arm, baseline balance, the rung named	
3 Secondaries and guardrails	
4 Threats that landed, two lines, with direction and bound	
5 Cost per unit and the decision; the dose; the next look booked	
Footer: pre-registered on; prediction held by; analysis by; data deleted on	

Template: the experiment register

Test	Claim	Design and rung	Result (effect, interval)	Decision	Date

Template: HEROES map

Course piece	Where it lives in HEROES	Set up?
The claim, the behaviour, the design, the APEASE pass	The intervention record (SHIELD designer: behaviour, COM-B diagnosis, functions, techniques, APEASE)	

Course piece	Where it lives in HEROES	Set up?
Pilot report, result, decision	Update Progress on the intervention	
System outcome	The priority behaviour's direct measure	
Observed-behaviour outcome, independent of the programme	The quarterly Team Pulse (Behaviour Index)	
The behaviour line for the board	Board pack; your one page behind it	
Randomisation and statistics	Yours: a spreadsheet, a calculator, a borrowed statistician	

Leah answers the CFO

Store trial: +1.9 real reports per 100 staff per month (1.3 to 2.5); +9 simulated (6 to 12); guardrail held; £28 per additional report, nearer £19 at scale; decision: scale, all 240 stores in week thirteen. Password manager: +17, rung four; decision: adjust, pre-install 100%, no import prompt, stepped wedge across the distribution centres and online. Tailgating: 41 to 14 per 1,000; decay to 22 by week sixteen; held at 15 with a four-weekly handover line; decision: scale, dose budgeted. Week-26 look on the store trial: +7 simulated, +1.6 real; decayed, holding, dose confirmed.

"Per additional real report, about £19 at scale. The mean unreported phishing incident last year cost us £2,400 to contain. The pilot number I nearly gave you in March was 29 points; the real one was 17. Here is the register." The CFO: "Which line is the control group?" He moved the awareness budget out of the discretionary column.

Your next test

Write your decision rule, and the first line of your one-page report as it would read today.

Reflect & apply

Pause the video and take ten minutes with these. Write, do not just think.

1. For your last intervention, what did it cost per unit of behaviour: per report, per verified payment, per enrolled account? If you cannot say, what is missing?

2. What is the maintenance dose, and who delivers it after the launch energy fades?

3. Write the first line of your one-page experiment report. Does it name the behaviour, the effect, the interval and the comparison?

4. Where does the design and the result of your next test live in HEROES, and who will read it there?

5. What is the standing control: the team or site that gets the baseline every time, so every future test has a comparison waiting?

COURSE SUMMARY

Close

Eight ideas to carry forward

1. **A claim that could be wrong.** Who, does what, by how much, by when. One primary outcome, guardrails, a baseline window, a sealed prediction.
2. **No comparison, no conclusion.** Randomise where you can; cluster where people talk; step where you must roll out; match where nothing else is possible. Name the rung.
3. **Size it before you start.** Base rate, smallest effect worth finding, the cluster penalty, a full cycle, a stop date. Look once.
4. **A pilot answers "can we deliver it".** Typical sites, a comparison alongside, the implementation log, kill criteria signed before.
5. **Treatment change minus comparison change.** Effect size in plain units, with an interval; as assigned; subgroups labelled; a null checked four ways.
6. **Walk the eight threats, twice.** History, instrumentation, selection, contamination, attrition, Hawthorne, decay, reactivity. Report what landed.
7. **The control group keeps every protection.** Test the nudge, never the control. Tell people, debrief everyone, blame nobody, aggregate the data.
8. **Decide by the rule you wrote.** Scale, adjust or stop; cost per unit of behaviour; the maintenance dose; one page, behaviour first.

The lines underneath

No comparison, no conclusion. · Write the prediction before the data. · The control group keeps every protection.

One experiment this quarter

Take the next thing you were going to roll out anyway. Randomise the order or hold back a comparison. Write the prediction. Look once. Put the comparison group's line on the slide. Write it here, with a date.

Glossary

Term	Meaning
Counterfactual	What the outcome would have done if you had done nothing. Every design is an attempt to estimate it.
Before-and-after	One group, two times. Confounded by history, regression to the mean, seasonality and maturation.
Comparison group	A group that experienced everything except the intervention, measured the same way at the same time.
Random assignment	Allocating units to arms by chance, so the arms are the same on everything on average and selection is removed.
Unit of assignment	What is randomised: a person, a team, a store, a site. Choose the unit people do not cross.
Cluster trial / design effect / ICC	Randomising groups rather than people; the penalty for people within a group resembling each other; the intraclass correlation that sets it.
Stratification	Pairing or grouping units on known factors before randomising within each pair, so the arms are balanced on them.
Stepped wedge	All units start as controls and switch on in a randomised order until all are treated; each is compared with itself and with units not yet switched.
Matched comparison / interrupted time series	Non-randomised designs: similar untreated units measured alongside; or many time points before and after a sharp switch.
Primary / secondary / guardrail / exploratory	The one deciding outcome; supporting outcomes; what must not break, with a limit; everything looked at afterwards.
Pre-registration	The one page written and dated before launch: claim, outcomes, baseline, design, sample, stopping rule, prediction, analysis, decision rule.
Base rate / minimum detectable effect / power	How often the outcome happens now; the smallest change worth finding; the chance the test sees that change if real.
Peeking / stopping rule	Looking at the primary repeatedly, which inflates false positives; the pre-planned look(s) that prevent it.
A/A check / sample ratio mismatch	Running the pipeline on two untreated groups to check it reads zero; a mismatch between assigned and observed counts that signals a broken pipeline.
Fidelity / dose	Whether the intervention was delivered as designed; how much of it actually arrived.
Kill criteria	Conditions written in advance that end a pilot: harm, cost, delivery.
Difference in differences	Treatment change minus comparison change.
Effect size / confidence interval	How big, in plain units; the range of effects the data allow.
Intention to treat / per protocol	Analysing units as assigned; analysing only those who actually received it, shown alongside.
Heterogeneity / multiple comparisons	Different effects in different subgroups; the inflation of false findings from slicing many ways.
Regression to the mean / survivorship	Extreme values moving towards average on their own; a population changed by who left.

Term	Meaning
Threats to validity	History, instrumentation, selection, contamination, attrition, Hawthorne and novelty, decay, reactivity.
Maintenance dose	The smallest repeat that holds an effect after the launch energy fades.
Cost per unit of behaviour	Everything the intervention cost, divided by the additional units of behaviour it produced.
Standing control / experiment register	Units that always receive the baseline so a comparison is waiting; the record of every test, nulls included.

Further reading

- Haynes, L., Service, O., Goldacre, B. and Torgerson, D. (2012). Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. Cabinet Office Behavioural Insights Team. Free on gov.uk.
- Gerber, A. S. and Green, D. P. (2012). Field Experiments: Design, Analysis, and Interpretation. Norton.
- Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin. Campbell, D. T. and Stanley, J. C. (1963). Experimental and Quasi-Experimental Designs for Research.
- Hussey, M. A. and Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. Contemporary Clinical Trials, 28(2).
- Donner, A. and Klar, N. (2000). Design and Analysis of Cluster Randomization Trials in Health Research. Arnold.
- Kohavi, R., Tang, D. and Xu, Y. (2020). Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge University Press.
- McCambridge, J., Witton, J. and Elbourne, D. R. (2014). Systematic review of the Hawthorne effect. Journal of Clinical Epidemiology, 67(3).
- Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. Social Science and Medicine, 210.
- National Cyber Security Centre. Phishing attacks: defending your organisation. ncsc.gov.uk.
- Michie, S., Atkins, L. and West, R. (2014). The Behaviour Change Wheel: A Guide to Designing Interventions (APEASE).
- CyBehave Academy: Choice Architecture for Security, module 7; How to measure a Security Champions programme, cybehave.com/measuring-champions.

Notes