

LEARNING PATH · BUILD A CULTURE OF EXPERIMENTATION

The Experimentation Culture Kit

Templates to start, grow and measure an experimentation programme

Start Charter, sponsor, champions, rituals, first 90 days	Grow Result one-pager, stall checklist, stakeholders, operating model	Measure Health scorecard, maturity self-assessment, training plan
---	---	---

14 TEMPLATES INSIDE

- 01

Big-tech habits checklist
- 02

Experimentation charter
- 03

Sponsor and champions map
- 04

Ritual agendas
- 05

First 90 days plan
- 06

Experiment result one-pager
- 07

Programme stall checklist
- 08

Stakeholder map and language card
- 09

Objection cards
- 10

Operating model chooser
- 11

RACI and guardrails
- 12

Programme health scorecard
- 13

Maturity self-assessment
- 14

Training plan by audience

BEFORE YOU START

How to use this kit

Fourteen practical templates, each taken from one article in our learning path *Build a culture of experimentation*.

Print what you need.

One page per template (rituals and maturity take two). Empty boxes, lines and cells are for your answers.

Fill it in together.

Most templates work best in a 30 to 60 minute session with the owner, the sponsor or the champions.

Read the article behind it.

Each page ends with the article that explains the template and its sources.

NO.	TEMPLATE	USE IT TO	EXPLAINED IN ARTICLE
01	Big-tech habits checklist	Borrow the habits of the companies that test at scale, at your own scale	1. Experimentation at scale
02	Experimentation charter	Agree purpose, scope, decision rights and rules on one page	2. Where to start
03	Sponsor and champions map	Name your sponsor, write the sponsor ask and score champion candidates	2. Where to start
04	Ritual agendas	Run the weekly stand-up, idea intake, show-and-tell, digest and quarterly review	2. Where to start
05	First 90 days plan	Plan the programme's first quarter in four phases with exit tests	2. Where to start
06	Experiment result one-pager	Write every result as a one-page story that leads with the decision	3. Building momentum
07	Programme stall checklist	Spot the six stall signals early and restart in 30 days	3. Building momentum
08	Stakeholder map and language card	Decide who hears what, how often, and in which words	4. Stakeholder communication
09	Objection cards	Prepare evidence-based answers to the eight objections you will hear	4. Stakeholder communication
10	Operating model chooser	Choose who builds, who decides and who sets the standards	5. Operating models
11	RACI and guardrails	Give every activity one accountable owner before you widen access	5. Operating models
12	Programme health scorecard	Track speed, trust, learning, reach and impact every quarter	6. Measuring the culture
13	Maturity self-assessment	Score twelve questions and place yourself on the five-level model	6. Measuring the culture
14	Training plan by audience	Train every audience for the decision it makes, and license test owners	7. Hiring and training

The seven articles of the learning path

- 1 **How Booking.com, Amazon, Microsoft, Google, Netflix and Airbnb Experiment at Scale (and What You Can Copy)**
<https://henkanpartners.ai/en/insights/experimentation-at-scale-big-tech>
- 2 **Where to Start: Finding Champions, Securing a Sponsor and Setting the Rituals of an Experimentation Programme**
<https://henkanpartners.ai/en/insights/start-experimentation-programme-champions-rituals>
- 3 **Building Momentum in an Experimentation Programme: Progress over Perfection**
<https://henkanpartners.ai/en/insights/experimentation-programme-momentum>
- 4 **How to Talk About Experimentation to Every Stakeholder: Executives, Analysts, Designers, Developers and More**
<https://henkanpartners.ai/en/insights/experimentation-stakeholder-communication>
- 5 **Who Owns Experimentation? Choosing an Operating Model: Centralised, Centre of Excellence, Hub-and-Spoke or Federated**
<https://henkanpartners.ai/en/insights/experimentation-operating-models>
- 6 **How to Measure an Experimentation Culture: Programme Metrics, a Learning Library and a Maturity Self-Assessment**
<https://henkanpartners.ai/en/insights/experimentation-programme-metrics-maturity>
- 7 **Hiring and Training an Experimentation Team: Roles, Skills and an Upskilling Plan**
<https://henkanpartners.ai/en/insights/hiring-training-experimentation-team>

TEMPLATE 01

Big-tech habits checklist

Sort the big-tech playbook into what to copy, what to adapt and what to leave alone, then plan the scaled-down version.

WHEN TO USE IT Before you design your programme, or whenever someone proposes copying a big-tech practice wholesale. The habits of the leaders travel to any company; their machinery does not.

1. Where do you stand? Tick what is already true for your programme

Copy as is

tick if in place

☐

Anyone may propose a test, with guardrails rather than managers saying no

☐

One written success metric and shared definitions

☐

A pre-launch checklist and a second reviewer on results

☐

A searchable library of every test

☐

Leaders who test their own ideas and accept losing

Adapt to your scale

tick if in place

☐

Test bolder changes

☐

Run a few tests in sequence

☐

One or two trained champions per team, like bar raisers or ambassadors

☐

A monthly class and one-hour onboarding

☐

Holdouts and long-term checks only on the biggest launches

Do not copy

tick if avoided

☐

Building your own platform

☐

Thousands of concurrent tests

☐

Large teams of data scientists

☐

Chasing their test counts

☐

Testing tiny tweaks without the traffic

2. Scale down the practices worth borrowing Mid-size versions from the article; add an owner and a date

BIG-TECH PRACTICE	MID-SIZE VERSION	EFFORT	IN PLACE?	OWNER	BY WHEN
Experiment council (Google)	A one-page test plan and checklist, reviewed by one trained colleague before launch	Low	☐ Yes ☐ Partly ☐ No		
Bar Raisers (Amazon), Ambassadors (Booking.com)	One trained champion per team who reviews plans 1-2 hours a week	Low	☐ Yes ☐ Partly ☐ No		
Monthly class (Microsoft)	A one-hour onboarding for anyone who wants to propose a test	Low	☐ Yes ☐ Partly ☐ No		
Central repository (Booking.com)	A shared library: hypothesis, screenshots, result, decision, tags	Low to medium	☐ Yes ☐ Partly ☐ No		
Real-time alerts (Booking.com)	Your tool's sample ratio mismatch and guardrail alerts, checked daily	Low	☐ Yes ☐ Partly ☐ No		
Metric review (LinkedIn)	One owner for the definition of each key metric in your analytics	Medium	☐ Yes ☐ Partly ☐ No		
In-house platform (all)	A commercial A/B testing tool, configured well	Medium	☐ Yes ☐ Partly ☐ No		

3. Your next three habits

HABIT 1

HABIT 2

HABIT 3

TEMPLATE 02

Experimentation charter

A one-page agreement on why you test, what is in scope, who decides what and which rules every test follows.

WHEN TO USE IT In weeks one and two of a new programme: the sponsor signs version 0 in week two. Update it at every quarterly review, so the rules stay when a sponsor or owner leaves.

PROGRAMME NAME	VERSION	DATE	NEXT REVIEW

SECTION	WHAT WE AGREE
1. Purpose Why we test, in one or two sentences	
2. Scope Which areas, channels and change types are in and out	
3. Sponsor and owner Names and roles	
4. Decision rights Who can propose, launch, stop and ship	
5. Success metric The primary metric tests are judged on	
6. Guardrails Metrics that must not degrade	
7. Trust standards Minimum quality every test meets	
8. Rituals Cadence of each ritual	
9. Idea intake How anyone submits an idea	
10. Sharing results Where results live and who sees them	
11. Boundaries Brand, legal, ethical and customer limits	
12. 90-day targets What success looks like by the first review	

SIGNED BY THE SPONSOR (NAME, DATE)	SIGNED BY THE PROGRAMME OWNER (NAME, DATE)

Section 4, decision rights, deserves the most care: who may launch a test without asking shapes your operating model (see template 10).

TEMPLATE 03

Sponsor and champions map

Put names on the starting coalition: one sponsor, one owner and four to six champions across functions.

WHEN TO USE IT Before the first test. Secure the sponsor first, then shortlist about ten champion candidates and invite the best four to six personally.

1. The coalition Write a name in each box; you do not need every function on day one

EXECUTIVE SPONSOR

Provides cover from above

PROGRAMME OWNER

The hub: backlog, rituals, QA, reporting

PRODUCT

DESIGN / UX

ENGINEERING

MARKETING / CRM

ANALYTICS

TRADING / MERCHANDISING

Champions carry the work sideways into their teams (about half a day a week each)

2. The sponsor ask Five traits to check, then four specific commitments to request

☐ Owns a roadmap you can test

☐ Cares about a metric testing can move

☐ Will show up

☐ Can be wrong in public

☐ Is staying

COMMITMENT	WHAT IT LOOKS LIKE IN PRACTICE	AGREED	NOTES
1. Name it publicly	An email or all-hands mention announcing the programme, its owner and its first area	☐	
2. Protect capacity	For example, one developer day a week and part of one analyst, ring-fenced for 90 days	☐	
3. Agree a decision rule	"Changes in the test area ship when a test supports them, or by a documented exception"	☐	
4. Protect losing tests	Opens the show-and-tell, celebrates a test that stopped a bad change	☐	

3. Champion shortlist Score each candidate 1 to 3 per criterion; invite those with 11 points or more

NAME	FUNCTION	CREDIBILITY	CURIOSITY	ACCESS TO A TEST SURFACE	MANAGER SUPPORT	COMFORT WITH BEING WRONG	TOTAL	INVITED
							/15	☐
							/15	☐
							/15	☐
							/15	☐
							/15	☐
							/15	☐

TEMPLATE 04

Ritual agendas

Five short rituals, each with a fixed cadence, a fixed agenda and one output. If a ritual has no output, cancel it.

WHEN TO USE IT Book all five in calendars for the next 90 days and paste the agenda into each invite. Timings add up to the slot length, so meetings end on time.

RITUAL	CADENCE	LENGTH	ATTENDEES	SINGLE OUTPUT	DAY AND TIME
Experiment stand-up	Weekly	30 min	Owner, analyst, developer, designer, champions with live tests	Updated test board and a named owner for every blocker	
Idea intake and prioritisation	Fortnightly	45 min	Owner, champions, product managers, researcher, analyst	Ranked backlog and the next two to three tests committed	
"What we learned" show-and-tell	Monthly	45 min	Open to all; sponsor opens	Learnings logged and one decision announced	
Results digest	Monthly	1 page	Whole company; sponsor forwards it	A written record people can search later	
Quarterly roadmap review	Quarterly	90 min	Sponsor, heads of product, marketing, tech; finance optional	Next quarter's areas, targets and capacity agreed	

1. Weekly experiment stand-up

Length 30 min Chair programme owner Board one card per test: idea, design, build, QA, live, analysing, decided

8 min	Live test health. Traffic split as expected (no sample ratio mismatch), guardrail metrics normal, any bugs reported.	☐
8 min	Launching this week. QA status against the launch QA checklist, sign-offs needed, planned start date.	☐
8 min	Ready to decide. Tests that reached their planned duration: who writes the result, when it goes to the decision owner.	☐
6 min	Blockers and asks. Each blocker gets a named owner and a date. Escalations to the sponsor are noted here.	☐

BLOCKER	OWNER	DATE	ESCALATE TO SPONSOR?
			☐ Yes
			☐ Yes

Anti-pattern: discussing results in depth. Park those for the show-and-tell or a separate analysis session.

2. Idea intake and prioritisation

Length 45 min, fortnightly Chair owner, with champions presenting their teams' ideas

Rule: no idea is discussed unless it was submitted in the intake form 48 hours before, with the problem, the evidence and the hypothesis.

10 min	New submissions. Each submitter has one minute: the problem, the evidence (analytics, research, survey, support tickets), the proposed change.	☐
15 min	Scoring. Score with one agreed framework.	☐
10 min	Feasibility and sizing. Developer and analyst flag build effort and whether traffic can support the test.	☐
10 min	Commit. Agree the next two or three tests to design, and retire ideas older than three months that never ranked.	☐

COMMITTED TEST	OWNER	TARGET LAUNCH

Anti-pattern: the most senior person's idea goes first. Score it like any other; that is the point.

3. Monthly “what we learned” show-and-tell

Length 45 min

Chair sponsor opens, owner hosts

Audience open invitation; record it

- 3 min

Sponsor opening. Why this matters, and one decision the sponsor made because of a test.

☐
- 5 min

Guess the winner. Show control and variant of the first test; the audience votes before the reveal.

☐
- 24 min

Three test stories. Told by the champion who ran it: question, what we changed, what happened, what we decided. At least one must be a loss or a flat result.

☐
- 5 min

What we will test next. Top of the backlog, and how people can submit ideas.

☐
- 8 min

Questions.

☐

TEST STORY (8 MIN EACH)	TOLD BY (CHAMPION)	RESULT	DECISION
		☐ Win ☐ Loss ☐ Flat	
		☐ Win ☐ Loss ☐ Flat	
		☐ Win ☐ Loss ☐ Flat	

Anti-pattern: a slide deck of uplift percentages. Tell stories about decisions.

4. Monthly results digest

Length one page

Owner programme owner writes it; sponsor forwards it with one line of their own

HEADLINE

The single most useful thing we learned this month, in one sentence

TESTS CONCLUDED

For each: the question, the result (win, loss, flat, inconclusive) and the decision taken

LEARNING OF THE MONTH

One insight about customers that holds beyond the single test

PROGRAMME PULSE

Tests launched, tests concluded, average days from idea to decision

COMING NEXT

Two or three tests about to launch, and who to contact

ASK

One thing you need: an idea for a page, a developer, a decision

5. Quarterly roadmap review

Length 90 min

Chair sponsor

Attendees heads of product, marketing and technology for the areas in scope, the owner, two champions; finance optional

- 15 min

Programme scorecard. Tests launched and concluded, time to decision, share of tests passing QA and trust checks, decisions taken.

☐
- 20 min

Learning themes. Three or four patterns from the quarter, with the decisions they changed.

☐
- 15 min

Value and cost. Decisions with estimated impact, bad changes avoided, and the capacity used. Keep estimates honest and ranged.

☐
- 25 min

Next quarter's areas and targets. Which areas to keep, which to add, what to stop.

☐
- 15 min

Capacity and asks. Developer, analyst and design time needed; any tool or training needs.

☐

AREAS AGREED FOR NEXT QUARTER	TARGETS AGREED	CAPACITY AGREED

TEMPLATE 05

First 90 days plan

Run the programme's first quarter in four phases. Each phase ends with an exit test, not a date.

WHEN TO USE IT At the start of a new programme. If the first test has not concluded cleanly by week six, do not move on to building rhythm; fix the build, QA or traffic problem first.

1

Set up
Weeks 1-2

DATES _____

☐ Sponsor ask agreed

☐ Owner named

☐ Four to six champions recruited

☐ First area chosen

☐ Tracking and testing tool checked

☐ Charter version 0

EXIT TEST

Sponsor has signed charter v0

☐ Passed on _____

2

First tests
Weeks 3-6

DATES _____

☐ Intake form live

☐ 15 to 20 ideas logged and scored

☐ Two or three tests designed, QA'd and launched

☐ Weekly stand-up running

EXIT TEST

First test concluded with no data-quality issues

☐ Passed on _____

3

Rhythm
Weeks 7-10

DATES _____

☐ First show-and-tell and first digest

☐ Two to four more tests

☐ Champion training

☐ Result write-up template in use

EXIT TEST

Rituals happen without the owner chasing attendance

☐ Passed on _____

4

Review and widen
Weeks 11-13

DATES _____

☐ Quarterly review

☐ Charter v1

☐ Programme scorecard

☐ Second area chosen

☐ Capacity plan for next quarter

EXIT TEST

Sponsor renews commitment and agrees the next area

☐ Passed on _____

Score your first area Traffic + Appetite + Politics (5 = least contested) + Test-ability, each 1-5

CANDIDATE AREA	TRAFFIC	APPETITE	POLITICS	TEST-ABILITY	TOTAL
					/20
					/20
					/20
					/20
					/20

Shortlist areas scoring 15 or more; pick the one whose owner will join the sponsor review.

FIRST-QUARTER MEASURE	TARGET	ACTUAL	OUR VIEW
Tests concluded cleanly			Measure the first quarter on three things: tests concluded cleanly, decisions taken because of a test, and rituals held on schedule. Promise a number of decisions and a working rhythm, never an uplift figure.
Decisions taken because of a test			
Rituals held on schedule			

TEMPLATE 06

Experiment result one-pager

One page per test, in the order a busy reader asks: what did we decide, why did we test, what changed, what happened, what did we learn and what happens next.

WHEN TO USE IT For every concluded test, including losses and flat results. Write the decision rule before launch, so the write-up takes minutes, not days. Store every one-pager in a searchable library.

TEST NAME / ID	PAGE OR JOURNEY	OWNER	LAUNCH → DECISION DATES

WRITE BEFORE LAUNCH
If the variant wins, we will
If it loses, we will

HEADLINE The decision, in one sentence

WHY WE TESTED Problem, evidence, hypothesis

WHAT WE CHANGED Screenshot of control and variant

CONTROL

VARIANT

WHAT HAPPENED Primary metric, interval, guardrails, SRM

METRIC	NAME	CONTROL	VARIANT	EFFECT AND INTERVAL (RANGE, NOT A POINT)
Primary				
Guardrail				
Guardrail				
Guardrail				

☐ QA passed ☐ SRM check passed ☐ No data incidents | ☐ Win ☐ Loss ☐ Flat ☐ Inconclusive

WHAT WE LEARNED One insight about customers that holds beyond this test

NEXT Ship, iterate or stop, with an owner and a date

☐ Ship ☐ Iterate ☐ Stop OWNER DATE

Five rules: lead with the decision · show the change · report a range, not just a number · name one thing you learned about customers · make it findable

TEMPLATE 07

Programme stall checklist

Programmes rarely announce that they are stalling. They drift. Check the six signals monthly and restart deliberately.

WHEN TO USE IT At every monthly review, and whenever the test count drops, the backlog fills with cosmetic ideas or the sponsor stops asking for the digest.

1. Six stall signals Tick every signal you see today

SEEN	SIGNAL	HOW TO SPOT IT	LIKELY CAUSE	FIRST FIX
☐	Velocity falls	Tests started per month down for two months in a row	Empty backlog, a bottleneck in build or QA	Map cycle time; refill the backlog to five briefed ideas
☐	Only cosmetic tests	Most tests change copy, colour or layout on one element	No research input; fear of bigger changes	Reserve slots for iterations and big bets; run a research sprint
☐	Results not used	Winners wait more than a month to ship	No owner for implementation; developer capacity	Name an owner and date in every one-pager
☐	No decisions changed	Tests confirm plans already made	Testing used as validation, not decision making	Test before the decision is taken; agree decision rules up front
☐	Silence	Digest unread; sponsor no longer asks	Results not translated into business language	Rewrite results as one-pagers; restore a monthly leadership slot
☐	Untrusted data	SRMs, broken tracking or disputed results recur	No QA or automated checks	Pause new launches for one week; fix checks; rerun the last disputed test

2. Diagnose with numbers

TESTS STARTED PER MONTH, LAST SIX MONTHS

M-6M-5M-4M-3M-2M-1

CYCLE TIME (DAYS, IDEA TO DECISION), LAST FIVE TESTS

TEST 1TEST 2TEST 3TEST 4TEST 5

RESULTS THAT WERE NEVER IMPLEMENTED

3. Restart in 30 days

WHEN	STEP	OWNER	DONE
Week 1	Diagnose. Count tests started per month for the last six months, map cycle time for the last five tests, and list results that were never implemented.		☐
Week 1	Fix trust. Check SRM and tracking on recent tests. If anything is broken, fix it before launching more.		☐
Week 2	Pick one important question. Ask the sponsor which decision they most want evidence on this quarter, and design a test around it.		☐
Weeks 2-3	Restore the rhythm. Reinstate one weekly review of 30 minutes and one monthly results slot with leadership. Fill the backlog with five briefed ideas.		☐
Week 4	Tell the story. Publish a one-pager for every result, including the losses, and a short note on what changed in the programme.		☐

Most stalled programmes do not lack ideas or tools. They lack a fixed weekly rhythm and someone whose job is to keep it. Restore the rhythm first.

TEMPLATE 08

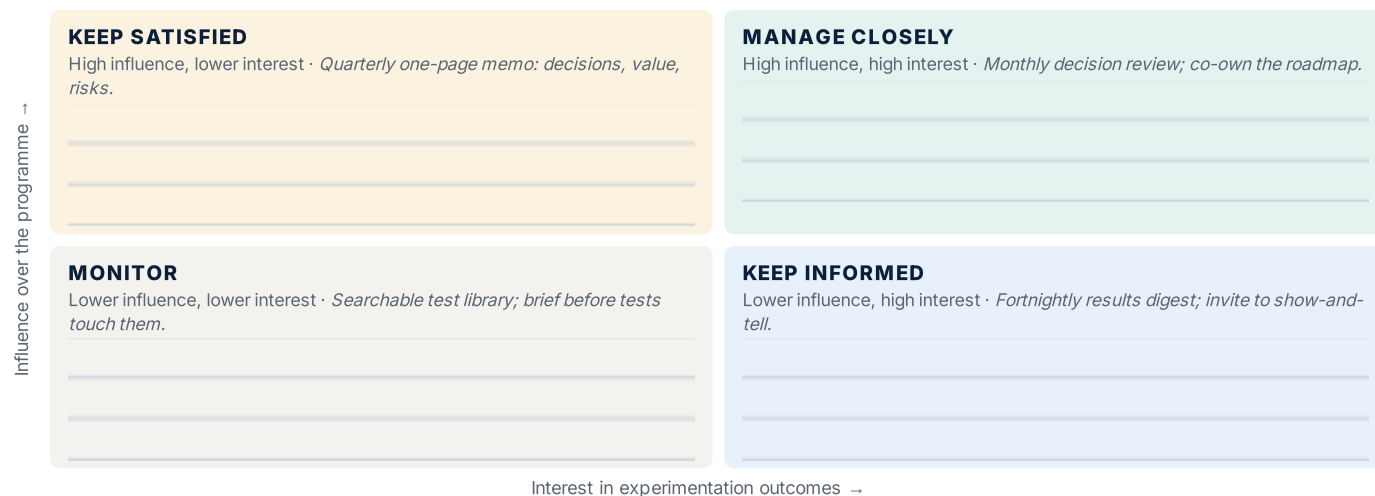
Stakeholder map and language card

Decide who needs to hear what and how often, then tell one story in each audience's dialect. Change the framing, never the result.

WHEN TO USE IT When you set up the programme and each time a new stakeholder joins. Place every team on the grid with real names, then adapt the language card to your company's words.

1. Map stakeholders by influence and interest

Write names in each quadrant; the channel is in italics



2. Three rules for every audience

Decision first, data second

Open with what was decided or what needs deciding.

Give a range, not a point

"Between +0.4% and +3.8% orders, most likely around +2%" is honest and still useful.

Translate into their unit

The test does not change; the unit does.

3. Language card

Phrases are examples to adapt, not scripts; add your own words in the last column

Audience	Lead with	Unit to use	Say	Avoid	Our words
CEO / exec committee	The decision and the risk avoided	EUR range, strategic priority	"Before we commit, let's test the riskiest part."	Test counts, p-values, "you were wrong"	
Finance	Incremental margin and how it will be verified	Margin range, payback, holdout	"A holdout will confirm the real figure."	Summed uplifts, annualised single numbers	
Product managers	The roadmap question answered	Their outcome metric	"Which assumption would hurt most if wrong?"	"Your feature lost."	
Designers / UX	What we learned about users	Behaviour plus the why	"Bring us your boldest concept."	"The data says your design is worse."	
Developers	Stability and clean-up plan	Milliseconds, errors, tickets	"Speed is a guardrail metric."	"It's just a small script."	
Data analysts	Method and validity checks	CI, power, SRM	"Written down before launch."	"Directionally positive."	
Merchandising / trading	Calendar fit and category learning	Sales, sell-through	"This test pauses during the sale."	"Ignore the weekly dip."	
Brand / marketing	Guardrails they set	Brand and retention metrics	"Nothing goes live that you haven't seen."	"Conversion is all that matters."	
Legal / compliance	Predictable review path	Test-type risk level	"Pre-approved list, 48-hour path for the rest."	"Everyone else does it."	
Customer service	What customers will see	Contacts per 1,000 orders	"Here is what it looks like from Monday."	"It's only a test."	

From: [How to Talk About Experimentation to Every Stakeholder: Executives, Analysts, Designers, Developers and More](https://henkanpartners.ai/en/insights/experimentation-stakeholder-communication)
henkanpartners.ai/en/insights/experimentation-stakeholder-communication

TEMPLATE 09

Objection cards

Eight objections come back in every company. Each tells you which scorecard the speaker is protecting; each needs evidence, not enthusiasm.

WHEN TO USE IT Before a sponsor meeting, a quarterly review or any conversation with a sceptical stakeholder. Write one piece of evidence from your own company on each card, then cut them out.

1 "We don't have enough traffic"

ANSWER Test bigger changes on fewer, busier pages
SHOW Sample-size calculation per page
SAY "Our product pages can detect a 3% change in four weeks; our FAQ pages cannot, so we'll use research there."
AVOID "Let's run it anyway and see."
 OUR EVIDENCE (FROM OUR OWN COMPANY)

2 "It will hurt the brand"

ANSWER Brand sets the guardrails before we build
SHOW Brand-approved variants; brand metrics
SAY "What would make a variant unacceptable to you? We'll write it into the brief."
AVOID "It's only for two weeks."
 OUR EVIDENCE (FROM OUR OWN COMPANY)

3 "The boss has already decided"

ANSWER Test to size and de-risk, not to overrule
SHOW Rollout plan with a holdout
SAY "Let's roll this out as a test, so we can show the board what it delivered."
AVOID "We need to test whether you're right."
 OUR EVIDENCE (FROM OUR OWN COMPANY)

4 "Testing is too slow"

ANSWER Slower than guessing, faster than a failed launch
SHOW Time-to-decision; cost of a bad release
SAY "Four weeks to know, versus six months to find out after launch."
AVOID "Good science takes time."
 OUR EVIDENCE (FROM OUR OWN COMPANY)

5 "The test lost, so it was useless"

ANSWER A loss stopped a bad change from shipping
SHOW Losses avoided; learning rate
SAY "This test stopped a change that would have cut orders; that's the programme doing its job."
AVOID "It was inconclusive, but directionally positive."
 OUR EVIDENCE (FROM OUR OWN COMPANY)

6 "It's a risk for SEO"

ANSWER Google publishes rules for safe testing
SHOW Canonicals, 302s, short test duration
SAY "We follow Google's published A/B testing guidance; here is the checklist."
AVOID "Google won't notice."
 OUR EVIDENCE (FROM OUR OWN COMPANY)

7 "It's just button colours"

ANSWER We test decisions, and small can be big
SHOW This quarter's tests by type
SAY "Here are this quarter's tests by type; most change what customers decide, not what they see."
AVOID "Every detail matters."
 OUR EVIDENCE (FROM OUR OWN COMPANY)

8 "AI will do it for us"

ANSWER AI speeds building; it does not replace evidence
SHOW QA, SRM checks, human sign-off
SAY "AI lets us test more ideas for the same budget; we still decide which ideas and whether the result holds."
AVOID "AI can't do what we do."
 OUR EVIDENCE (FROM OUR OWN COMPANY)

The answer is never "you're wrong"; it is "here is how we handle that risk".

TEMPLATE 10

Operating model chooser

Choose who generates ideas, who builds and runs tests, who decides what ships and who sets the standards.

WHEN TO USE IT When you set up the programme, and once a year with the programme metrics in front of you. The right model depends on volume and maturity, not ambition.

1. Tick every statement that is true today The model with the most ticks is your starting point

Centralised	Centre of excellence	Hub-and-spoke	Federated
One team ideates, builds, runs and analyses every test for the business	A central team owns tools, methods and quality, runs complex tests and coaches business teams that start to test	A hub sets standards; analysts, developers or 'A/B test owners' are embedded in business units, with a line back to the hub	Product teams own the whole test lifecycle on a shared self-serve platform; a small central team runs the platform and standards
WHEN IT FITS	WHEN IT FITS	WHEN IT FITS	WHEN IT FITS
☐ You run fewer than about 50 tests a year, on one website or app ☐ Testing is new, and you need consistent quality to earn trust ☐ Only one or two teams have both the traffic and the appetite to test ☐ You work with an agency or fractional team that brings the specialist skills	☐ Demand for tests clearly exceeds what one team can build ☐ Two to five business teams have the traffic and the will to test, but not the skills ☐ You run roughly 50 to a few hundred tests a year ☐ You need to keep one quality bar while capacity grows	☐ Several business units each have their own roadmap and enough traffic ☐ Units want testing capacity they control, but you cannot afford inconsistent methods ☐ You run a few hundred tests a year or more ☐ You can place at least one specialist in each unit	☐ Many independent product teams ship to the same platform every week ☐ A trusted platform, shared metric definitions and a searchable test library are in place ☐ Training and in-person support are funded centrally ☐ You run 1,000 or more experiments a year (the 'fly' phase)
BUILDS AND RUNS Central team	BUILDS AND RUNS Central team, some business teams	BUILDS AND RUNS Embedded specialists	BUILDS AND RUNS Product teams
DECIDES WHAT SHIPS Central team with stakeholders	DECIDES WHAT SHIPS Business owner, centre reviews	DECIDES WHAT SHIPS Business unit lead	DECIDES WHAT SHIPS Product team
SETS STANDARDS Central team	SETS STANDARDS Centre	SETS STANDARDS Hub	SETS STANDARDS Platform and enablement team
TICKS	TICKS	TICKS	TICKS

2. Signals to move, or to step back

Signals you have outgrown your current model	Signals to step back
☐ Queue time. Good ideas wait more than a quarter for build capacity. ☐ Shadow testing. Teams run tests outside the programme. ☐ Narrow coverage. Most tests come from one team or one page type. ☐ Specialist overload. Central analysts spend more time on set-up than on analysis and coaching. ☐ Demand for autonomy. Business units ask to own their roadmap, not just submit ideas.	☐ Rising SRM failures, broken tracking or tests launched without QA. ☐ Different teams reporting different numbers for the same metric. ☐ Wins declared on secondary metrics, or tests stopped as soon as they look good. ☐ Results not recorded anywhere a newcomer could find them.

3. Our decision

ACCOUNTABLE OWNER (ONE NAME)

OUR MODEL TODAY

NEXT MODEL, AND THE SIGNAL THAT WILL TRIGGER THE MOVE

TEMPLATE 11

RACI and guardrails

Every model needs the same seven roles; what changes is who holds them. Give each activity exactly one accountable owner.

WHEN TO USE IT

When you write the mandate for your operating model, and before you give more people the right to test. Write R, A, C or I in each cell.

R

Responsible: does the work

A

Accountable: owns the outcome, one per activity

C

Consulted

I

Informed

ACTIVITY	EXECUTIVE SPONSOR	EXPERIMENTATION LEAD	ANALYST / DATA SCIENTIST	DEVELOPER	DESIGNER / UX RESEARCHER	PRODUCT OWNER	CHAMPIONS	ONE A? (CHECK)
Programme strategy and targets								
Test idea intake and prioritisation								
Hypothesis and test design								
Build and QA								
Launch and monitoring (SRM, guardrails)								
Analysis and readout								
Ship or stop decision								
Test library and sharing								
Training and certification								
Tooling and budget								

In the article's reference version (centre of excellence or hub-and-spoke), the experimentation lead owns the process and the product owner owns the ship or stop decision. In a centralised model the lead takes most R cells; in a federated model the product owner does.

Six guardrails to have in place before democratising

Install the missing ones first, starting with the test library and metric definitions

IN PLACE	GUARDRAIL	WHAT IT IS	OWNER	DATE
☐	1. Training and certification	A short course and a practical test before someone can launch alone		
☐	2. Peer review	A second person checks the hypothesis, metrics and analysis plan before launch and the readout before a decision		
☐	3. QA checklist	A standard pre-launch and post-launch check of variations, tracking and traffic split		
☐	4. Metric definitions	One shared definition of each decision metric, computed the same way for every test		
☐	5. Guardrail metrics	Metrics every test must not harm, with harm thresholds agreed in advance		
☐	6. A central test library	One searchable record of every test: hypothesis, design, result, decision		

TEMPLATE 12

Programme health scorecard

One page, five families of metrics. Pair every speed metric with a trust metric, and never set a win-rate target.

WHEN TO USE IT Monthly for the programme team, quarterly with the sponsor. Targets are set per organisation; published reference points are context, not targets.

QUARTER

PROGRAMME

PREPARED BY

REVIEWED WITH SPONSOR ON

FAMILY	METRIC AND DEFINITION	LAST Q	THIS Q	TARGET	STATUS
Speed	Tests launched per month Experiments that started receiving traffic in the month				
	Median time to decision Median days from launch to a recorded decision (ship, stop or iterate)				
	Idea-to-launch cycle time Median days from an idea being approved in the backlog to launch				
Trust	Valid-test rate Completed tests that passed QA and the SRM check with no unresolved data incident				
	Pre-registration rate Tests with hypothesis, primary metric, guardrails and sample size written before launch				
Learning	Learning rate Valid tests with a decision-ready answer: a win, a detected regression, or a well-powered neutral				
	Win rate (context only, no target) Completed tests where the primary metric improved significantly			n/a	
	Library logging rate Completed tests with a full library record within 14 days of the decision				
Reach	Active teams and experimenters Teams and people that launched at least one test in the quarter				
	Share of key decisions tested Decisions on a pre-agreed list that were tested or had a recorded reason not to test				
	Library usage Searches or views per month, or share of new ideas that cite past tests				
Impact	Implementation rate Winning variants live in production within 60 days of the decision				
	Attributed value Annualised value of implemented winners, discounted for decay and uncertainty, agreed with finance				
	Losses avoided Count (and estimated value) of changes stopped because a test showed harm				

On target Improving Needs action · circle one per row

ONE LINE OF COMMENTARY PER FAMILY

Speed

Trust

Learning

Reach

Impact

TOP THREE ACTIONS FOR NEXT QUARTER

1.

2.

3.

TEMPLATE 13

Maturity self-assessment

Twelve questions, two per dimension, place your programme on the five-level Henkan & Partners maturity model.

WHEN TO USE IT Take a baseline now and re-run it every six months. Answer for what actually happened in the last six months, not for what is planned. Ask two or three people from different teams to answer separately, then compare.

SCORE EACH QUESTION

- ① Never or not in place ② Rarely or just started ③ Sometimes or partly in place ④ Mostly in place
 ⑤ Always, fully in place and used

LEADERSHIP AND DECISIONS

Q1 When a significant change to the site, app or offer is planned, is it tested, or is a reason for not testing written down? ① ② ③ ④ ⑤

Q2 Do senior leaders put their own ideas to a test and act on results that go against them? ① ② ③ ④ ⑤

OWNERSHIP AND PEOPLE

Q3 Is there a named programme owner with protected time, and a sponsor who reviews the programme at least once a quarter? ① ② ③ ④ ⑤

Q4 Do teams beyond the core team launch tests themselves, within shared guardrails? (1 = one person or an agency only; 5 = most product and marketing teams) ① ② ③ ④ ⑤

PROCESS AND QUALITY

Q5 Does every test have a written hypothesis, a primary metric, guardrail metrics and a sample size before launch? ① ② ③ ④ ⑤

Q6 Are QA and sample ratio mismatch checks run on every test, with failures recorded rather than quietly re-run? ① ② ③ ④ ⑤

METRICS AND MEASUREMENT

Q7 Do you track programme metrics (speed, trust, learning, reach, impact) and review them on a fixed cadence? ① ② ③ ④ ⑤

Q8 Is the value of the programme estimated with a documented, conservative method that finance has agreed? ① ② ③ ④ ⑤

KNOWLEDGE AND MEMORY

Q9 Is every test, including losers and flat results, stored in one searchable library using a standard record? ① ② ③ ④ ⑤

Q10 Do new test ideas start with a search of past results, and do you analyse your test history as a whole at least once a year? ① ② ③ ④ ⑤

PLATFORM AND DATA

Q11 Do you trust the data behind your tests: analytics audited, consent effects understood, metric definitions shared? ① ② ③ ④ ⑤

Q12 Can tests be set up, monitored and analysed without bespoke engineering each time, and can people (or AI assistants) query results in a governed way? ① ② ③ ④ ⑤

Scoring grid

Dimension score = average of its two questions; overall score = average of the six dimension scores

DIMENSION	QUESTION SCORE		QUESTION SCORE		DIMENSION SCORE	DIMENSION LEVEL
Leadership and decisions	Q1		Q2			
Ownership and people	Q3		Q4			
Process and quality	Q5		Q6			
Metrics and measurement	Q7		Q8			
Knowledge and memory	Q9		Q10			
Platform and data	Q11		Q12			
Overall score (average of the six dimension scores)						

LEVEL 1 Ad hoc 1.0–1.79	LEVEL 2 Repeatable 1.8–2.59	LEVEL 3 Trusted 2.6–3.39	LEVEL 4 Scaled 3.4–4.19	LEVEL 5 Evidence by default 4.2–5.0
---	---	--	---	---

Weakest-link rule. Your overall level cannot be more than one level above your weakest dimension. A programme with Level 4 reach but Level 1 data quality is a Level 2 programme. Improve the lowest-scoring dimension first, then the one your sponsor cares about most.

OUR LEVEL (AFTER THE WEAKEST-LINK RULE)	WEAKEST DIMENSION	PRIORITY FOR NEXT QUARTER

The five levels, dimension by dimension

Shade the cell that best describes you in each row

DIMENSION	1. AD HOC	2. REPEATABLE	3. TRUSTED	4. SCALED	5. EVIDENCE BY DEFAULT
Leadership and decisions	Decisions are made by opinion; tests confirm ideas already chosen.	A sponsor backs the programme; some decisions wait for test results.	Leaders accept results that go against them; a written rule says which changes must be tested.	The share of key decisions tested is tracked and reviewed by leadership.	Evidence is expected for every key decision; leaders put their own ideas to a test in public.
Ownership and people	No owner; tests are run by whoever has time, or by an agency on request.	A named owner with protected time, supported by a small core team or partner.	The core team sets standards, trains others and reviews every test.	Several teams launch tests, with a champion in each; the central team acts as a centre of excellence.	Most product and marketing teams test on their own; the central team runs the platform and quality.
Process and quality	No written hypothesis; QA is ad hoc; tests are stopped when the result looks good.	Hypothesis template, prioritised backlog and a QA checklist before launch.	Sample size and guardrails set before launch; automated sample ratio mismatch (SRM) checks; peer review of analysis.	Standards are built into tools and templates; the valid-test rate is tracked and stays high.	The process is improved from meta-analysis of past tests; quality issues are rare and visible.
Metrics and measurement	Only the win or loss of individual tests is reported.	Test count and win rate are tracked.	A scorecard covering speed, trust and learning is reviewed monthly; metric definitions and a primary success metric are shared.	A full scorecard including reach and impact is reviewed quarterly with the sponsor; value is estimated with finance.	Programme metrics are part of business reviews; holdout groups measure cumulative impact; the success metric feeds team goals.
Knowledge and memory	Results live in slide decks, inboxes and people's heads.	A shared spreadsheet or page lists tests and outcomes.	A searchable library holds a standard record for every test, losers and flat results included, with consistent tags.	New ideas start with a library search; a meta-analysis runs at least once a year; a results digest goes out regularly.	The library is the system of record, linked to decisions and implementation; AI assistants can query it under governance.
Platform and data	A testing tool is used occasionally; analytics has not been audited.	One testing tool, connected to analytics, with basic tracking checks.	Audited analytics, known consent effects, a shared metric catalogue and alerting on broken tests.	Client-side and server-side testing, self-service set-up and integrations with analytics and customer data.	The platform supports many concurrent tests, checks for interactions and stops harmful tests automatically.

TEMPLATE 14

Training plan by audience

Train every audience for the decision it makes, not for the tool, and tie training to a “licence to test”.

WHEN TO USE IT When you launch the programme's training path, and each quarter as new teams start to test. Teach with your own data and make your experts teach.

1. Training by audience Who attends, when, and who teaches

AUDIENCE	FORMAT	WHAT THEY LEARN	WHO (NAMES)	DATE	TRAINER	DONE
Executives and sponsors	1 hour	How to read a result and its interval, why most tests lose, what a guardrail is, and the three questions to ask before accepting a win				☐
Product owners and managers	Half a day, then two coached tests	Turning evidence into a hypothesis, choosing one primary metric, reading results and deciding				☐
Designers and UX writers	Half a day	Designing variations that isolate one change, accessibility and brand guardrails, and when a change should simply ship without a test				☐
Marketing, trading and CRM	2 hours	Why a campaign launched mid-test can ruin it, how to read a result, and how to submit ideas with evidence				☐
Developers	1 day, hands-on	Tool set-up, stable selectors, flicker and performance, QA across devices, and the sample ratio check; server-side testing for those who need it				☐
Analysts	2 to 3 days, then peer review	Power and sample size, fixed-horizon versus sequential testing, variance reduction (CUPED), multiple comparisons and false positive risk				☐

2. A four-level licence to test Record who holds each level

LEVEL	WHO	HOW TO REACH IT	WHAT IT ALLOWS	PEOPLE AT THIS LEVEL
1. Reader	Anyone in product, marketing, design, leadership	1-2 hour introduction; quiz	Read results and the test repository; submit ideas with evidence	
2. Test owner	Product owners, champions, marketers who run tests	Foundations course, then 3 tests reviewed by the central team	Launch standard tests (front-end, one primary metric, guardrails) on their area	
3. Advanced owner	Senior owners, developers, analysts	Advanced modules (server-side, sequential testing, multi-metric) and 10+ tests with clean reviews	Launch complex tests; review level-2 owners' tests	
4. Reviewer / methods expert	Analysts, statisticians, the central team	Peer-nominated; teaches at least one module a year	Approve non-standard designs; audit random samples; maintain standards	

3. Keep quality visible

Review a random sample.
Pick experiments at random each month; it keeps quality visible without making every test wait for approval.

SAMPLE SIZE PER MONTH

Track the pass rate.
The share of tests that pass QA and sample ratio checks is a programme health metric (template 12).

PASS RATE THIS QUARTER

Add AI literacy.
What not to put into AI tools, how to check AI output, and which uses your legal team has approved.

MODULE OWNER

HENKAN & PARTNERS

About Henkan & Partners

Henkan & Partners is a boutique CRO and experimentation consultancy. We help e-commerce teams build and run experimentation programmes: securing the sponsor, writing the charter, setting up rituals, QA and reporting, and running fractional experimentation teams while you build your own.

Paris

Casablanca

India

Keep learning: the seven articles behind this kit

- 1 **How Booking.com, Amazon, Microsoft, Google, Netflix and Airbnb Experiment at Scale (and What You Can Copy)**
henkanpartners.ai/en/insights/experimentation-at-scale-big-tech
- 2 **Where to Start: Finding Champions, Securing a Sponsor and Setting the Rituals of an Experimentation Programme**
henkanpartners.ai/en/insights/start-experimentation-programme-champions-rituals
- 3 **Building Momentum in an Experimentation Programme: Progress over Perfection**
henkanpartners.ai/en/insights/experimentation-programme-momentum
- 4 **How to Talk About Experimentation to Every Stakeholder: Executives, Analysts, Designers, Developers and More**
henkanpartners.ai/en/insights/experimentation-stakeholder-communication
- 5 **Who Owns Experimentation? Choosing an Operating Model: Centralised, Centre of Excellence, Hub-and-Spoke or Federated**
henkanpartners.ai/en/insights/experimentation-operating-models
- 6 **How to Measure an Experimentation Culture: Programme Metrics, a Learning Library and a Maturity Self-Assessment**
henkanpartners.ai/en/insights/experimentation-programme-metrics-maturity
- 7 **Hiring and Training an Experimentation Team: Roles, Skills and an Upskilling Plan**
henkanpartners.ai/en/insights/hiring-training-experimentation-team

TALK TO US

henkanpartners.ai/en/contact

If you want help choosing your first area, running the first rituals, or an outside view on why a programme has stalled.

The Experimentation Culture Kit · October 2026. Frameworks and templates are Henkan & Partners frameworks; quoted practices are attributed to their sources in the articles. You are welcome to use and adapt these templates inside your organisation; please keep the source line on each page.