Test Hypotheses

Welcome to *Test Hypotheses*!

The *Test Hypotheses* phase examines whether the selected intervention contributes to measurable changes in the target behaviour by addressing key behavioural barriers (identified during the *Define* and *Explore & Diagnose* phases).

The insights generated here guide whether an intervention should be adapted, scaled, or discontinued, and lay the foundation for the Scale phase, where evidence is translated into broader action.

Unlike previous chapters that anchor on specific steps and tools for applying behavioural science, this chapter — *Test Hypotheses* — is organised slightly differently.

Conducting rigorous evaluations to measure the effectiveness of behavioural interventions is complex and goes beyond the scope of this field guide. Still, this chapter offers essential guidance to help teams understand the value of impact evaluations, the challenges of demonstrating causality, and the key elements and decisions involved in designing and coordinating experiments. The goal of this chapter is to give UNICEF teams/partners a baseline understanding of testing intervention hypotheses to enable more efficient collaboration with evaluation specialists.

Given the wide range of evaluation approaches available, this chapter outlines the most commonly used methods within behavioural science. Many of these activities require advanced technical expertise. It is therefore recommended that teams consult evaluation specialists and use this chapter to understand what expertise may be needed, what to ask evaluation specialists, and which challenges may arise. For teams interested in directly conducting evaluations, additional manuals and practical resources are listed throughout the chapter as well as in the “Learn more” section at the end of this phase.

Why “Test Hypotheses”?

The “Why Test Hypotheses?” section will explore:

1. **The importance of evaluation**
2. **The causation challenge**
3. **The counterfactual framework:** understanding what would have happened otherwise
4. **How randomization creates the gold standard for counterfactuals**

This foundation sets the stage for choosing the right evaluation method and using evidence to strengthen our interventions.

1. The importance of evaluation

Imagine you are ill and a doctor offers you a new medication. When you ask about its effectiveness, the doctor replies: "We haven't tested it formally, but it probably works. Several patients who took it seemed to get better, and our team feels confident about it." Would you take it? Most people would refuse, and for good reason.

Yet when it comes to social programmes and behavioural interventions, we often do the opposite. We implement programmes based on good intentions, promising theoretical frameworks, and anecdotal stories of success, but without rigorous evidence of impact. So why do we hold social interventions to a lower standard than medicine, when both aim to improve human wellbeing?

When success stories mislead

International development has a long history of well-received programmes that captured attention, attracted funding, and seemed to promise breakthrough results — until their impact was rigorously evaluated. One of the most famous examples is microfinance.

Emerging in the 1980s, it was hailed as a transformative tool for poverty reduction. By offering small loans to individuals in low-income settings, microfinance aimed to foster entrepreneurship and economic growth. The model spread rapidly and gained widespread acclaim.

Over time, however, rigorous evaluations told a more complicated story. While microfinance improved access to credit, its impact on poverty reduction, economic mobility, and long-term well-being was less clear. Rigorous research studies highlighted increased debt burdens for borrowers, limited scalability, modest business outcomes, and little progress on addressing structural poverty[1 2 3]. Articles and books — such as "Big Money Backs Tiny Loans That Lead to Debt, Despair and Even Suicide" and More Than Good Intentions reflect the disillusionment that followed and the crucial role evaluation played in revealing what anecdotes could not.

Microfinance is not an isolated case. Other high-profile programmes — like PlayPump[4], One Laptop Per Child[5], and the Millennium Villages project[6] — generated early enthusiasm, but later fell short on their impact when carefully evaluated.

The assumption trap: Why good intentions aren't enough

Even with the best intentions, we're prone to making assumptions that can lead our interventions astray. This "assumption trap" operates at multiple levels:

- **Assuming to understand the problem.** Often, issues are diagnosed from individual perspectives, rather than a deep understanding of the lived experience of communities. What seems obvious from the outside may lack or miss crucial context and complexity.
- **Assuming to know what will work.** Based on personal expertise or past experiences, individuals may become convinced that certain approaches

1 John, B. (2024, November 14). Challenges and limitations of microfinance in achieving large-scale poverty reduction and job creation [Working paper].
2 Akbari, M., Nikijoo, I., Khodapanah, B., Foroudi, P., & Padash, H. (2025). Forty Years of Microfinance Research and Its Impact on Consumers: A Review and Research Agenda Using the ADO-TCM Framework. International Journal of Consumer Studies, 49(4), e70101.
3 Blanc, J. (2014). Microfinance, Debt and Over-Indebtedness: Juggling with Money, Isabelle Guérin, Solène Morvant-Roux et Magdalena Villarreal (dir.). Editions Routledge, Londres, Royaume-Uni, 2014, 316 pages. Revue internationale de l'économie sociale: recma, (334), 122-124.
4 UNICEF. (2007). An Evaluation of the PlayPump® Water System as an Appropriate Technology for Water, Sanitation and Hygiene Programmes https://www-tc.pbs.org/frontlineworld/stories/southernafrica904/flash/pdf/unicef_pp_report.pdf
5 Cristia, Julian and Ibarraran, Pablo and Cueto, Santiago and Santiago, Ana and Severin, Eugenio, Technology and Child Development: Evidence from the One Laptop Per Child Program (February 2012). IDB Working Paper No. IDB-WP-304, Available at SSRN: https://ssrn.com/abstract=2032444.
6 Mitchell, S., Gelman, A., Ross, R., Chen, J., Bari, S., Huynh, U. K., ... Sachs, J. D. (2018). The Millennium Villages Project: a retrospective, observational, endline evaluation. The Lancet Global Health, 6(5), e500–e513. https://doi.org/10.1016/S2214-109X(18)30065-2

will succeed without sufficient evidence. However, it's often possible that the factors that could impact the outcome haven't yet been observed.

- **Assuming implementation will go as planned.** It's common to underestimate practical challenges and overestimate how closely interventions will follow their design when applied in real-world settings.
- **Assuming positive anecdotes mean success.** When favourable feedback is shared or positive moments are observed, it's common to generalize these experiences, giving them more weight than they deserve in assessing overall impact.
- **Assuming correlation means causation.** When things improve after an intervention, it's natural to attribute the change to our work, even when other factors might be responsible.

These assumptions don't stem from carelessness or incompetence, they're a product of how human cognition works. As noted in previous modules, human minds seek patterns, prefer confirming evidence, and create coherent narratives — even when reality is messier. While these tendencies serve people well in many contexts, they can be misleading when evaluating complex social interventions.

Without systematic evaluation, these assumptions remain unchallenged. This may lead to investing in programmes that seem effective but don't actually create meaningful change, or worse, may cause unintended harm. Evaluation provides the structured process needed to move beyond assumptions and understand the true impact of the work.

The value proposition: Why evaluation is worth the investment

Evaluations aren't just academic exercises, they deliver concrete value:

- **Resource optimization:** In resource-constrained environments, evaluation helps direct limited funds toward interventions with proven impact.
- **Course correction:** Timely evaluation allows us to identify and address implementation problems before scaling, preventing the widespread adoption of ineffective approaches.
- **Stakeholder confidence:** Rigorous evaluation builds trust with donors, governments, and communities, facilitating partnerships and long-term support.
- **Scale and replication:** Well-evaluated programmes provide a blueprint for expansion, allowing successful approaches to benefit more communities.
- **Prevention of harm:** Evaluation can identify unintended negative consequences of well-intentioned programmes before they affect large populations.

2. The causation challenge

Beyond "before and after"

When we implement a programme and see improvements, it feels natural to assume our intervention made the difference. A vaccination campaign launches, and disease rates drop; a parent education programme begins, and school attendance rises. These connections seem obvious, but they might be misleading.

The fundamental challenge in evaluation is determining whether our intervention actually *caused* the changes we observe, or whether those changes occurred because of other factors. This is harder than it might appear at first glance.

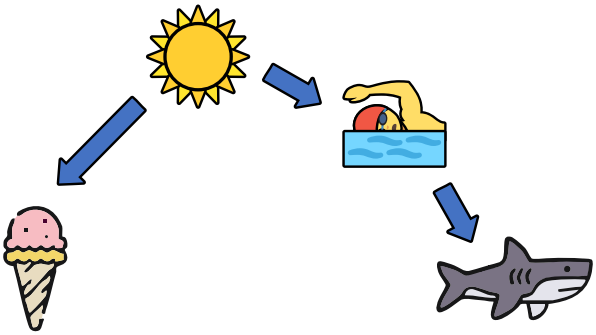
Correlation vs. causation

Correlation means two things happen together. Causation means one thing actually makes the other happen. This distinction is crucial for evaluating the impact of programmes or interventions.

Consider a classic example: ice cream and shark attacks. Data shows that when ice cream sales go up, so do shark attacks. Are ice cream sales causing shark attacks? Of course not. This is what statisticians call the "third variable problem" or "common cause confounding", when a hidden factor influences both variables at the same time. In this case, temperature is the hidden third variable that influences both outcomes independently. During summer months:

- Higher temperatures lead to increased ice cream consumption
- The same higher temperatures lead more people to swim in the ocean
- More swimmers in the water increases the likelihood of shark encounters

This can be illustrated with a simple causal diagram or Directed Acyclic Graph (DAG). The blue arrows represent causal influences. There is no arrow connecting ice cream sales and shark attacks because there's no direct causal relationship between them. They are correlated (they happen together) but not causally linked (one does not cause the other).



Why this matters for programmes: If we don't understand the difference between correlation and causation, we risk drawing the wrong conclusions — and designing or scaling interventions that are not actually responsible for the change we're seeing.

Consider a real-world development example. Imagine that a UNICEF nutrition programme is implemented in several communities. Soon after, children's growth indicators begin to improve. It might seem intuitive to attribute this improvement to the programme. But what else may be going on?

- ➔ Perhaps its harvest season, and food availability has naturally increased.

- Maybe another organization started providing clean water, reducing diarrhoeal disease.
- Or perhaps the government implemented an economic policy that increased household income at the same time.

Any of these factors could explain the improvement -or at least contribute to it. If we assume that our programme caused the change, when it was actually due to other factors, we may invest in or scale up interventions that don't actually work. Worse, we might overlook what really did drive the change, missing opportunities to replicate or strengthen more effective solutions.

Understanding correlation vs. causation helps us avoid these pitfalls. It pushes us to ask better questions, design smarter evaluations, and make more informed decisions.

Confounding factors: Why effects are hard to Isolate

When an intervention is implemented and outcomes are measured, many other factors beyond the programme can influence observed changes. These other unrelated factors are known as confounding variables. By failing to account for these variables, there is a risk of attributing effects to the intervention that were actually caused by something else.

This challenge is called endogeneity: a situation where the relationship between an intervention and its outcome is distorted because other variables are at play. Recognising this helps to understand why simple before and after comparisons can be misleading.

Below are some of the most common confounding factors, illustrated with examples drawn from typical UNICEF programme contexts:

Time-based confounders:

Changes that would have happened regardless of our intervention.

- **Seasonal variations.** Nutrition indicators improve after a feeding programme begins, but the programme launched just before the harvest season, when food is naturally more available.
- **Pre-existing trends.** School enrolment rises after an education campaign, but data shows rates were already increasing steadily due to long-term economic development.

Selection confounders:

Differences between those who participate in a programme and those who don't.

- **Self-selection bias.** Families who join a parenting programme may already be more engaged in their children's development, making the programme appear more effective than it is.
- **Targeting bias.** A WASH programme targets communities with high rates of diarrhoeal disease. Even without the programme, these extreme rates might decline over time simply due to natural variation.

Environmental confounders:

External events or conditions occurring at the same time.

- **Concurrent programmes**. A child protection campaign launches as the government begins stricter enforcement of child labour laws. It becomes difficult to discern which initiative drove the observed changes.
- **Policy changes.** An early childhood nutrition programme rolls out just as a national food subsidy is introduced. Both could be influencing improved nutrition indicators.

Measurement confounders:

Changes in how we track or detect outcomes.

- **Improved monitoring.** After a new reporting system is introduced alongside an anti-trafficking initiative, the number of cases rises. This is not due to a rise in trafficking, but an improvement to the process of detection.

These examples illustrate why it's challenging to determine whether an intervention was the real cause of the observed changes. When multiple factors influence outcomes simultaneously, how can a programme's true impact be isolated?

3. The counterfactual framework: Understanding what would have happened otherwise

At the heart of causal inference lies a seemingly simple question: What would have happened if the intervention had not taken place?

This alternative scenario — where the programme didn't exist — is known as the counterfactual. It represents the benchmark against which the real-world outcome is compared to determine whether the programme truly made a difference.

Consider a child who receives a vaccine and doesn't contract the disease. Did the vaccine prevent illness, or would the child have remained healthy anyway? Both outcomes — seeing the child vaccinated and unvaccinated — can't be observed at once. This dilemma is what scholars call the "fundamental problem of causal inference." It simply isn't possible to observe both the actual and the counterfactual for the same individual.

Instead, the counterfactual is approximated by finding or creating a valid comparison group. This group is as similar as possible to the intervention group and experiences the same external conditions — such as changes in season, economic shifts, or policy reforms — but does not receive the intervention.

If both groups are exposed to the same context, any meaningful difference in outcomes between them can be attributed to the programme itself. This is the foundation of credible evaluation design. A carefully constructed counterfactual helps to move beyond assumptions and confidently answer an important question: Did the intervention make the difference, or would it have happened anyway?

Potential outcomes: A formal way to think about counterfactuals

To help with reasoning around casual impact, statisticians use what's called the potential outcomes framework. This provides a formal structure for thinking about the difference an intervention makes by creating multiple possible realities for each unit. A unit can be a person, household, school, or community, and for each unit, there are two potential outcomes:

- $\mathbf{Y}_1$: The outcome if the unit **receives** the treatment/intervention
- $\mathbf{Y}_0$: The outcome if the unit **does not receive** the treatment/intervention

The causal effect is the difference between these two potential outcomes: $\mathbf{Y}_1 - \mathbf{Y}_0$.

But here is the real challenge; we can only ever observe one of these outcomes for any individual. When a child receives a vaccine, we see what happens with vaccination (Y_1). However, it's impossible to know what would have happened to that same child without the vaccine (Y_0). This unobserved alternative, the counterfactual, remains forever unknown.

This is the fundamental problem of causal inference: the need to know both what happened and what would have happened otherwise, when only one reality can be observed.

Child A (Treatment Group) → Received intervention → Observed outcome: Y_1

Child B (Comparison Group) → Did NOT receive intervention → Observed outcome: Y_0

Causal effect = $Y_1 - Y_0$

The solution to counterfactuals: From individual to group counterfactuals

Since both outcomes cannot be observed for the same person (it's not yet possible to clone people or travel between parallel realities!), the focus can instead be shifted onto groups. By carefully constructing comparison groups that are very similar to one another, it's possible to approximate what would have happened to those in the group that received the intervention and those who did not.

Rather than asking, "Did this specific child gain weight because of the programme?" we instead ask

"On average, how much more weight do children gain when participating in the nutrition programme compared to when they don't participate?"

Now that the focus is on groups rather than individuals, it's time to explore how to create valid group-level counterfactuals. In other words, how to create a comparison group that might be similar across as many observable and unobservable factors/variables as possible.

The challenge of constructing valid counterfactuals

With an understanding of the need to compare groups and not individuals, the next concern is how to create a group that accurately represents what would have happened without the intervention.

Constructing this comparison group (or counterfactual) is one of the most important and challenging steps in causal evaluation. As noted earlier, confounding factors can easily distort conclusions. The quality of the causal inference depends entirely on how well the selected comparison group mirrors the intervention group in all ways except one: they did not receive the treatment or intervention.

This is where evaluation design comes in. Different approaches attempt to approximate what would have happened to our intervention group had they not received the treatment (the counterfactual). Each approach comes with its own trade-offs between rigour, feasibility, and risk of bias. Some designs offer stronger causal claims but require more control or resources; others are more flexible but introduce greater uncertainty.

- **Before-after comparison — Weak counterfactual**

  One common but flawed approach to evaluation is the before-after comparison. This method measures outcomes just before a programme begins and again afterward, attributing any change to the intervention. While simple and intuitive, this approach is highly vulnerable to confounding factors that can influence outcomes over time, independent of the programme itself. These include:

  - **Time-based confounders:** Seasonal variations or long-term trends (e.g., an agricultural training programme shows increased yields, but the evaluation period coincides with the natural growth season).
  - **Environmental confounders:** Simultaneous programmes or policy changes (e.g., a nutrition programme appears successful, but the government simultaneously introduced free school meals in the same area).
  - **Measurement confounders:** The act of measurement influences results (e.g., repeated surveys make households aware of "desired" behaviours like handwashing, prompting changes independent of the programme itself).

  All of these factors can create the illusion of impact, when in reality, the change might have occurred anyway.

  For example, a community health programme launched in April shows improved outcomes by August. The month of August, however, is also the start of the dry season, when waterborne illness naturally declines. Hence, the observed improvement may actually be unrelated to the intervention. The before–after approach essentially assumes that "the same group, earlier in time" can act as its own counterfactual. In dynamic, real-world settings, this assumption rarely holds — making this design a weak basis for drawing causal conclusions.

- **Non-equivalent comparison group — better, but flawed**

  A step up from before-after comparisons is a non-equivalent comparison group — a group that does not receive the intervention, but is observed during the same time period as the intervention group. This approach helps to address many time-based confounders since both groups are exposed to the same external conditions (e.g. seasons, policy shifts, or economic changes).

  However, this design is still vulnerable to selection confounders: differences between the groups that can affect outcomes independently of the intervention. These include:

- **Self-selection:** Individuals who choose to participate may already be more motivated, better resourced, or more health-conscious than those who don't.
- **Administrative selection:** Programmes are often intentionally delivered to areas with the greatest need or the highest potential for success, which can skew comparisons.
- **Baseline differences:** Even before the programme starts, comparison communities may differ in key ways, such as infrastructure, income, or demographics.

Researchers often try to match groups on observable characteristics, but this approach has its limits. Many important factors (e.g. attitudes, aspirations, resilience, or genetics) cannot be observed, but can still drive outcomes. These hidden differences make it difficult to confidently attribute changes to the intervention.

While stronger than before-after design, the non-equivalent comparison approach still falls short of producing high-confidence causal estimates unless additional methods (such as statistical adjustment or natural experiments) are applied carefully.

4. How randomization creates the gold standard for counterfactuals

The need for better counterfactuals.

As noted, both before–after comparisons and non-equivalent group designs are limited. They may help to observe change, but they struggle to isolate what caused that change — especially in the presence of confounding factors, both known and unknown.

Just like the example of ice cream and shark attacks, many real-world relationships are shaped by hidden variables. In programme evaluation, these hidden variables are often numerous, complex, and impossible to fully measure.

As such, it's necessary to find a way to create comparison groups that are balanced on characteristics that are both observed and unobserved.

This is where random sampling comes in. By first randomly sampling different groups (individuals, schools, or communities) from the target population and then randomly assigning which groups receive an intervention, this creates statistically equivalent groups. On average, these groups will be similar across all characteristics, both observed and unobserved, because they have been sampled from the same underlying population in a similar manner. This means that factors such as motivation, baseline health, income, unmeasured beliefs, or community norms will be randomly balanced.

When properly implemented, randomization ensures that the only difference between groups is whether or not they received the intervention. This approach makes it far more likely that any differences in outcomes are due to the programme itself, rather than external or pre-existing differences. For this reason, randomization is referred to as the "gold standard" in causal inference. When done well and with a large enough sample size, this approach provides the strongest possible confidence that the intervention caused the observed change.

Randomization as an approach to create counterfactuals

By randomly assigning which units (i.e. individuals, households, schools, or communities) receive an intervention, the influence of confounding factors is significantly reduced. All types of confounders noted earlier are now evenly distributed across groups by design. In practice, this means:

- Seasonal variations affect both groups equally
- Self-selection bias is eliminated, as participation is assigned, not chosen
- Pre-existing trends unfold similarly across both groups
- Concurrent programmes or policy changes impact both groups at the same time
- Measurement-related effects apply equally across groups

As a result, the only systematic difference between the groups is whether or not they receive the intervention. This allows for attributing any differences in outcomes directly to the intervention itself, rather than any other hidden or external influences.

To provide an example, imagine running a programme to encourage parents to send their children to school. There is a large group of eligible parents, but they are a diverse group, each shaped by different factors:

- Some are wealthy, others face financial hardship
- Some live close to school, others live far away
- Some completed higher education, others have little formal schooling
- Some have flexible jobs, others rigid schedules
- Some strongly value education, others are skeptical
- Some had good schooling experiences, others did not
- Some are highly motivated, others less so

By randomly assigning parents to either receive the intervention or be in the control group (i.e., don't receive the intervention), this ensures that all characteristics are distributed similarly between groups. This includes both observable characteristics, such as income and distance, and unobservable ones, such as beliefs and motivation.

Randomizing ensures that there is an equal distribution of these different traits across the two groups. In this example, any difference in school attendance rates after the intervention can be attributed to the intervention itself, rather than any pre-existing differences between the parents who did and didn't receive it.

The path from good intentions to real impact

Rigorous evaluation isn't just about academic credibility, it's about ensuring that programmes actually improve lives. As noted earlier, well-intentioned interventions can fail to deliver impact, waste precious resources, or even cause unintended harm when relying on assumptions rather than evidence. The difference between correlation and causation matters, as it determines whether solutions that truly work are scaled, or investment is made in programmes that simply happened to coincide with positive change. By constructing valid counterfactuals, ideally through randomization, understanding moves beyond what seems to work towards what actually works, for whom, and why.

This knowledge transforms how to design programmes, allocate resources, and ultimately serve communities. While rigorous evaluation may seem daunting, continuing interventions without knowing their true impact is far riskier.

Common concerns about randomization and responses

While randomization is one of the most rigorous approaches to address the challenge of causality and counterfactuals, it tends to generate concerns. These can be structured around six different categories:

1. Cost and resource

CONCERN: Randomized evaluations are expensive and resource intensive.

RESPONSE:

- While rigorous evaluations do require investment, the cost must be weighed against the value of reliable evidence
- Not all RCTs need to be large scale or expensive — small, focused studies can be cost-effective
- The cost of implementing ineffective programmes at scale is ultimately much higher than evaluation costs
- Existing data sources and clever designs can sometimes reduce costs substantially

2. Time constraints

CONCERN: Randomized evaluations take too long, and UNICEF needs to respond quickly.

RESPONSE:

- Rapid-cycle testing, a structured process of trying out small-scale interventions, measuring results quickly, and iterating based on feedback to refine solutions can provide initial insights in shorter timeframes
- Phased implementation allows for both immediate action and rigorous evaluation
- Time invested in evaluation prevents wasting years on ineffective approaches
- Some outcomes can be measured in the short term (e.g., adherence to medication such as ARVs) while others require tracking over longer periods (e.g., suppressed viral load)

3. Ethical concerns

CONCERN: It's unethical to withhold potentially beneficial programmes from control groups.

RESPONSE:

- When resources are limited, randomization is often the fairest allocation method
- There is no way to know whether programmes work without testing them, some may have no effect or even cause harm
- Phased implementation ensures all participants eventually receive the programme if proven effective
- The ethical imperative to ensure programmes truly help children justifies the process of rigorous testing

4. Political and stakeholder challenges

CONCERN: Government partners or communities may resist randomization.

RESPONSE:

- Framing matters — emphasize the benefits of this type of evaluation, including options for keeping it low-cost or time-bound
- Involve stakeholders early in the design process to address concerns and build ownership
- Create clear procedures for stopping trials if strong evidence of benefit emerges
- Explain that evaluation strengthens advocacy for successful programmes

5. Contextual relevance and generalizability

CONCERN: Results from one context won't apply to others where UNICEF works.

RESPONSE:

- Strategic site selection can improve generalizability
- Measuring implementation factors helps to understand what contextual elements matter
- Even localized evidence is better than no evidence
- Multiple evaluations across contexts can build a broader evidence base

6. Technical capacity

CONCERN: UNICEF staff may lack technical expertise to design and analyse randomized evaluations.

RESPONSE:

- Partnerships with academic institutions can supplement internal capacity
- Investment in staff training builds long-term organizational capability
- Simple randomized designs can be more accessible than complex quasi-experimental methods
- Evaluation specialists within UNICEF can provide technical support across programmes

Design considerations for impact evaluation

Having established why evaluation matters and how causation can be credibly identified, the focus now turns to designing evaluations that ask the right questions, measure the right outcomes, and generate evidence that directly informs decisions. The following section outlines key design considerations for conducting rigorous impact evaluations.

1. Defining research questions

Every evaluation begins with a simple challenge: clarifying the key questions that need to be answered. Begin by gathering key stakeholders and asking:

- What decisions depend on this evaluation?
- What would be done differently if X was known, versus Y?

Focus on 3-5 core questions that directly inform concrete actions. These typically fall into three categories:

- **Effectiveness**: Does the intervention work? How long does it take to work? How soon are effects observed?
- **Mechanism**: How does the intervention work?
- **Targeting**: For whom does it work best?

For each question, specify how the answer will guide decisions. If a vocational training programme increases employment by 10%, will it be scaled? What if it's only 5%? What if it works for men, but not women? Pre-defining these decision thresholds prevents any post-hoc rationalization and ensures that the evaluation generates actionable evidence. Remember to consider both positive and null scenarios – knowing something doesn't work is equally valuable for resource allocation.

Once the key research questions are defined, the next step is to determine how those questions will be answered. This requires translating broad objectives into precise, measurable outcomes and selecting the right indicators that capture real change.

2. Selecting primary outcomes and measurement strategy: Starting with precise questions

The success of the evaluation hinges on measuring the right things in the right way. This starts with transforming broad goals into precisely articulated evaluation questions. A vague question such as "Does this nutrition programme work?" leads to vague answers and unclear decisions. Instead, contrast this with a question such as "Does providing monthly nutritional counselling increase height-for-age z-scores by at least 0.2 standard deviations among children 6-24 months in rural communities within 12 months?"

In other words: "Does [intervention] lead to [specific measurable change, with a threshold if relevant] in [defined population] within [timeframe and context]?" This level of precision guides every subsequent measurement decision and ensures that results are both interpretable and actionable.

Once the evaluation questions have been stated, the next step is to identify the outcome measures that can help to respond to the questions.

3. Choosing what to measure

Choose outcomes that clearly show what the programme is trying to achieve. Focus on measures that directly indicate whether the intervention is making the intended difference. These should be specific enough to measure accurately and important enough to guide programme decisions.

Consider how close the outcomes are to the intervention. Proximal outcomes, those that occur soon after the intervention, are easier to change and measure, but may not capture the ultimate goal. Distal outcomes, those further down the results chain, show real impact, but they usually require larger samples and more time to detect.

For example, a school feeding programme might quickly increase attendance (proximal) but would require much larger data and more time to show improved learning (distal). Both types are useful: short-term outcomes show whether the programme is on track, while long-term ones show whether it's truly making a difference.

Once outcomes are clearly defined, it becomes essential to consider how they will be measured. Different data sources vary in accuracy, cost, and feasibility, and recognizing these trade-offs ensures that measurement decisions strengthen rather than compromise the evaluation's credibility.

Data collection methods and trade-offs

Distinct advantages and limitations of data sources:

Survey data provides flexibility to measure exactly what is needed but also introduces several challenges. Self-reported behaviours often suffer from social desirability bias. For example, parents often overreport vaccinations or underreport harsh discipline. Recall periods matter enormously; asking about events from last week yields different results in accuracy than last year. Survey fatigue can also compromise data quality in lengthy questionnaires. Marginalized groups may be more difficult to reach through phone surveys, and individuals in low-literacy communities may have difficulty understanding certain questions. Ensure cognitive interviews are carried out during piloting to ensure questions are culturally appropriate and understood as intended.

Administrative data, such as school records, clinic registers, or programme databases offers cost-effective, objective measurement but comes with constraints. Efforts are limited to what's already collected, which may not align perfectly with outcomes of interest. Data on certain ethnic groups or marginalized populations may be incomplete or missing. Data quality also varies wildly — some clinics maintain meticulous records while others barely function. Using administrative data often forces the unit of randomization to match the administrative levels of schools or clinics rather than individuals.

Direct observation and behavioural measures provide objective assessment but require careful implementation. Observers need extensive training to ensure consistency. It is important to approach communities with sensitivity, seek consent carefully, and avoid intruding on private or culturally sensitive spaces. Technology increasingly enables unobtrusive measurement (GPS tracking, sensor data) but may not be feasible in all contexts without requiring expensive equipment.

Biomarkers and anthropometric measures offer objectivity for health interventions but require specialized training and equipment. Can cold chains be maintained for blood samples? Will participants consent to invasive procedures? How will measurement errors be handled from different assessors, while ensuring that data collection is respectful and minimally burdensome for participants?

Identifying appropriate data sources is necessary but not sufficient alone; the timing and frequency of measurement are equally critical. Poorly timed data collection can obscure genuine effects or misrepresent programme performance.

4. Timing and frequency of measurement

Outcomes emerge on different timescales. Knowledge might change within weeks, behaviours over months, and health impacts across years. Measuring too early risks finding null effects for interventions that need time to work. Measuring too late risks missing effects that fade or become contaminated by other factors.

Consider multiple measurement points to understand effect dynamics. Does impact grow, plateau, or decay

over time? An initial boost that fades might suggest a need for reinforcement — fade-out effects are common with behaviour change interventions. Gradual change may suggest that effects build up over time and potentially from different processes. Ideally, try to budget for at least one follow-up beyond immediate post-intervention (e.g., six months later) to assess persistence.

Seasonal patterns can confound results if not carefully considered. For example, agricultural outcomes vary by harvest cycles, disease patterns follow seasonal trends, and school-based measures fluctuate with academic calendars. Time measurements to avoid conflating programme effects with seasonal variation, or ensure both treatment and control groups are measured simultaneously. Additionally, try to include data collection across these to understand how the effect interacts with these seasonal variations.

Timing indicates when change occurs, whereas mechanisms explain why. Measuring along the causal chain exposes how impact unfolds, where systems fail, and how programs can be refined.

5. Measuring mechanisms along the causal chain

Understanding why programs work (or don't) is equally as important as knowing whether they work. Mechanism measurement serves multiple purposes:

Theory validation: Do hypothesized pathways actually operate? A handwashing programme assumes: information -> knowledge -> attitude change -> behaviour change -> health improvement. Measuring each step validates or challenges these assumptions.

Failure diagnosis: When outcomes don't improve, mechanisms reveal where chains broke. Did teachers not receive training? Did they receive it but did not understand it? Did they understand but did not implement? Did they implement it but students weren't engaged? Each breakdown point suggests different solutions.

Programme refinement: Rather than abandoning "failed" programmes wholesale, mechanism data enables targeted fixes. If parents received nutrition information but lacked resources to buy diverse foods, adding vouchers might unlock impact.

Generalization: Programmes working through universal mechanisms, such as reminder effects, likely transfer across contexts better than those dependent on specific institutional features.

Don't just measure final outcomes, track intermediate steps. For a nutrition programme aimed at reducing child malnutrition, this could look like: caregiver knowledge (immediate), feeding practices (short-term), child dietary diversity (medium-term), and nutritional status (long-term). Each provides valuable information about how the programme is working.

Understanding mechanisms reveals how change happens, but to interpret those patterns accurately, it's important to ensure that the data truly reflects the populations to serve. Representativeness and inclusion in measurement are essential for generating evidence that captures diverse realities, not just those easiest to reach.

6. Ensuring representative and inclusive measurement

Who is measured matters as much as what is measured. School-based surveys systematically exclude out of school children, who are often the most vulnerable. Phone surveys exclude those without phones, and clinic data misses those not seeking care, or those unable to do so.

Build inclusive measurement strategies from the start, ensuring communities contribute to shaping indicators and data sources. Use multiple data sources to capture different populations. Oversample marginalized groups to ensure adequate representation. Translate instruments into local languages and pilot with diverse respondents. Train enumerators from communities being surveyed to improve rapport and understanding.

Consider whose perspective is being captured. Children, parents, teachers, and health workers may report differently about the same phenomenon. Even within

the same group (e.g., mothers), individuals' experiences may vary depending on their social identity, background, or position within the community. Triangulation across reporters can reveal important dynamics but requires clear protocols for handling discrepancies.

Even the most carefully designed instruments can fall short if they cannot be implemented effectively in routine real-world settings. Ensuring that measurement approaches are feasible, reliable, and field-ready is essential.

7. Practical measurement considerations

During the measurement process, keep these considerations in mind:

Pilot extensively. Never assume that instruments working elsewhere will also function in another context. Pilot with a small group of respondents mimicking actual field conditions. Test skip patterns, timing, translation, and comprehension. Debrief enumerators thoroughly, as they often spot problems that respondents won't mention.

Balance comprehensiveness with feasibility. Long instruments provide rich data but suffer from respondent fatigue, higher costs, and quality degradation. Most impacts can be detected with focused instruments. Reserve lengthy measurement for small-scale mechanism studies rather than large impact evaluations.

Plan for measurement error. All measurement contains error: anthropometric assessment varies between assessors, and test scores depend on testing conditions. Build in quality checks, such as standardization exercises for assessors, repeat measures on subsamples, and validation against external sources where possible.

Document everything. Create detailed protocols specifying exactly how each outcome is measured, coded, and constructed. Future researchers need to understand and potentially replicate chosen measures. Include survey instruments, training materials, and variable construction code in appendices.

Sound measurement practices ensure data quality, but its value depends on the study's ability to detect real effects. Adequate statistical power safeguards against false conclusions, ensuring that data not only describes what was observed, but reveals what truly worked.

8. Determining sample size and statistical power

Statistical power is essentially the evaluation's ability to detect a true effect when it exists, like having a radar sensitive enough to spot an approaching airplane. In contrast, an underpowered study is like searching for something with a dim flashlight, unable to find anything even when it's there. This becomes particularly important when determining how many participants are needed. If the sample is too small, there's a risk of concluding that the programme had no effect when in reality, it did (known as a "false negative"). However, gathering data from more participants than necessary wastes resources. The key factors affecting power include:

- The anticipated size of the programme's effects – bigger effects are easier to detect
- How much natural variation exists in the outcome measure – more variation requires larger samples
- The chosen unit of randomization – randomizing at a community level requires more effort than randomizing individuals

Build in realistic assumptions about attrition (10-20% is common), non-compliance (the treatment group not receiving the intervention), and contamination (the control group accessing the intervention). Each reduces an effective sample size; it's better to recruit 20% more participants than discover the study is underpowered after data collection.

9. Randomization architecture

When testing an intervention, choosing the level at which randomization will take place is critical. This means deciding whether to assign the intervention to individuals, groups, schools, communities, or another unit that makes sense for the project. The chosen level should match how the intervention is delivered in real life. For instance, if a new curriculum can only be introduced to an entire classroom rather than individual students, the classroom becomes the most practical unit for randomization.

The level of randomization also depends on what kind of information can be collected. If data can only be measured at a group level — such as household spending or school attendance — it makes sense to randomize at that same level.

TABLE 8. DIFFERENT TYPES OF UNITS OF RANDOMIZATION

UNIT OF RANDOMIZATION	ADVANTAGES	CONSIDERATIONS	WHEN TO USE	EXAMPLE
Individual-level randomization	Requires smaller sample size compared to other levels; high statistical efficiency.	Risk of spillover effects when individuals interact; logistical challenges in delivering different interventions within the same setting (e.g. community or classroom).	Suitable when interaction between individuals is minimal and the intervention can be easily targeted to specific individuals.	SMS vaccine reminders sent to caregivers randomly selected within a large urban district.
Household-level randomization	Captures household-level decision-making; aligns with how many behaviours and outcomes are shaped.	Spillovers are still possible when neighbours interact; clear definition of "household" is required; analysis may need to account for household size differences.	Appropriate when interventions affect or involve all members of a household (e.g. home visits, conditional cash transfers).	Conditional cash transfers provided to randomly assigned households with children under five.
Community or village-level randomization (cluster randomization)	Reduces risk of contamination or spillovers; often easier to manage politically and operationally.	Requires more clusters (communities) to reach statistical power; high variation between communities increases variance; implementation logistics more complex at scale.	Useful when individuals within communities are likely to influence each other or when interventions are delivered publicly (e.g. community mobilisation).	Community health worker-led immunisation campaigns tested across randomly assigned villages.

UNIT OF RANDOMIZATION	ADVANTAGES	CONSIDERATIONS	WHEN TO USE	EXAMPLE
Facility-level randomization (e.g. schools, clinics)	Practical for institutional delivery settings; aligns with existing organisational structures; suitable for staff-level or facility-wide interventions.	Facilities may vary in size, quality, or staff turnover; catchment area overlap can lead to spillovers; statistical power is limited by the number of facilities available.	Appropriate for evaluating interventions delivered through institutions, especially when individual-level targeting is impractical.	Interpersonal communication training provided to staff in randomly selected health clinics.

10. When randomization isn't possible

Sometimes randomization isn't feasible, due to political, ethical, or practical constraints. Stakeholders may view random assignment as unfair, programs may already be rolled out, or sample sizes may be too small for meaningful randomization. In these cases, quasi-experimental methods attempt to create valid comparisons using statistical techniques to approximate the counterfactual that randomization would have provided.

These approaches work by identifying and controlling for factors that influence both programme participation and outcomes, what is sometimes known as "closing backdoor paths." While these methods can provide valuable evidence, they require stronger assumptions about the data and context. They typically need larger samples, more extensive data collection, and more complex analysis than RCTs. Most importantly, they remain vulnerable to bias from unmeasured factors that randomization would have eliminated.

For more guidance on quasi-experimental approaches, see Appendix 1. Only pursue these methods with expert guidance, as their validity depends critically on context-specific assumptions that are often untestable.

Practical implementation checklist for impact evaluations

Evaluation protocol

Before developing an evaluation protocol, it is helpful to first outline any learning objectives, using a tool like the Learning Agenda. This tool helps define and organize the key questions to be answered in an impact evaluation, and its use is illustrated in Appendix 2 through the **case study on increasing childhood vaccination uptake in Lebanon**.

Once the Learning Agenda is in place, the evaluation protocol can then be built. The evaluation protocol provides a structured guide and plan on how the intervention will be evaluated. Developing an evaluation protocol transforms design decisions into a comprehensive technical document that guides implementation and analysis. Think of it as a contract with a future self, preventing selective analysis and ensuring scientific integrity. A strong protocol pre-specifies every analytical decision before seeing outcomes, protecting against conscious and unconscious bias toward finding positive results.

Essential protocol components include:

- Detailed description of the intervention — what exactly will be delivered, by whom, how often
- Theory of Change with clear causal pathways
- Evaluation questions mapped to specific hypotheses
- Power calculations with all assumptions made explicit
- Precise outcome definitions with exact survey questions (if the outcome measures are survey-based)
- Randomization procedures including stratification variables
- Analytical models with specific regression equations
- Covariate lists determined by theory, not data
- Subgroup analyses with clear rationale
- Procedures for handling attrition and non-compliance
- Robustness checks to test the sensitivity of findings
- Pre-specifying the analysis plan is particularly crucial for the following:
- Primary versus secondary outcomes, to prevent outcome switching
- Subgroup analyses, to avoid mining for significant effects
- Inclusion/exclusion criteria for analysis sample
- Handling of outliers and missing data
- Multiple testing adjustments

Any deviation from the protocol should be clearly labelled as exploratory in reports. For reference, see the case study in Lebanon, where the full evaluation protocol is available.

Consider registering the protocol in public repositories (AEA Social Science Registry, RIDIE, ClinicalTrials.gov) before data collection begins. Registration provides a timestamp, proving pre-specification and enabling the research community to track all studies, not just published successes. Include enough detail that another researcher could replicate the evaluation, while maintaining operational flexibility for necessary field adaptations that don't compromise the core design.

Implementation plan

This is the detailed operational roadmap that translates the evaluation design into real-world action. The plan takes the technical protocol and turns it into day-by-day implementation guidance, specifying who will do what, when, where, and with what resources – throughout the entire evaluation lifecycle. The plan includes the following:

- granular timelines with specific dates and milestones
- clear role assignments with individuals responsible for each task
- resource requirements (staff time, materials, transportation, technology)
- delivery tracking indicators
- contingency plans for common problems
- communication protocols between team members and partners

The Implementation Plan tool can be a simple and helpful template to develop a step-by-step guide on how the intervention will be carried out. An example of how this tool can be used is shown in Appendix 2.

Why it matters

The gap between a brilliant evaluation design and failed execution is usually poor implementation planning. Even rigorous research designs fail when tablets aren't charged, surveys aren't translated properly, or teams don't know who's responsible for participant recruitment. The implementation plan prevents the thousand small failures that can invalidate an otherwise well-designed evaluation. It ensures coordination among multiple actors (research team, implementing partners, government officials, community leaders) who may have different priorities and working styles.

Without clear operational planning, there's a risk of discovering too late that vaccination campaigns conflict with your data collection, key staff are unavailable during crucial periods, or materials weren't printed in time. The plan also provides a management tool for keeping complex operations on track and identifying problems before they cascade into evaluation failure.

Keep these key considerations in mind:

- **Plan around fixed constraints.** Work backwards from hard, fixed deadlines – such as agricultural seasons, school years, budget cycles, or religious holidays – to set realistic timelines, then add a 20-30% buffer time for inevitable delays. If data collection is estimated to take three weeks, plan for four.
- **Assign clear single-point accountability for each activity.** Avoid shared responsibility, which often means no responsibility. When "the team will conduct training" is noted, specify exactly who leads, who assists, and who is accountable if an activity doesn't occur.
- **Track delivery, not just outcomes.** Include simple delivery metrics that can be tracked weekly: the number of participants enrolled versus target, surveys completed per day, intervention sessions delivered as planned. These are different from outcome metrics and focus purely on whether activities occurred on schedule.
- **Create a comprehensive budget.** A detailed budget is one that includes often-forgotten costs like translation services, transport for supervisors doing quality checks, phone credit for follow-ups, refreshments for community meetings, and replacement materials for damaged items.
- **Specify data flow precisely.** Note any data-related considerations, such as how the data will be transferred from paper forms to digital databases, who checks the data and has access, how often, and where it will be stored.
- **Plan for common field problems with specific contingencies**. What if heavy rains prevent travel during the survey period? What if key staff get sick or quit? What if government priorities suddenly change and the partner agency is reassigned? What if participants are busy with harvest when follow-up was initially planned? Use the Implementation Risks and Mitigation tool to systematically document any anticipated risks and the strategies in place to address them. To see this tool in action, refer to Appendix 2 for the case study in Lebanon, which illustrates its practical application.

- **Provide practical tools.** Include templates and standard operating procedures as annexes, so field teams have practical tools, not just abstract plans — this means actual scripts for recruitment, step-by-step guides for data entry, and checklists for intervention delivery.

Securing ethics approval

Ethics approval is the formal process of obtaining institutional review board (IRB) or ethics committee approval, which involves navigating submission requirements, timelines, and institutional procedures. This process is necessary to ensure the evaluation can legally and ethically proceed.

Beyond solely understanding ethical principles, this involves managing the practical bureaucratic process, including:

- identifying which IRB has jurisdiction (a university, ministry of health, UNICEF, or multiple)
- completing required training certifications for all team members
- preparing extensive documentation packages in specific formats
- responding to reviewer comments and requests for clarification
- maintaining compliance throughout the study, including amendments, adverse event reporting, and annual renewals

Why it matters

Ethics approval is legally required for research involving human subjects — proceeding without it can invalidate the entire evaluation, expose institutions to legal liability, and destroy community trust. Many journals won't publish results without proof of ethics approval, and donors increasingly require evidence of ethical clearance before releasing funds. The practical challenge is that ethics review can take 2-6 months, with multiple rounds of revision, and delays here cascade through the entire prospective timeline. A technically perfect evaluation is worthless if data collection cannot begin because ethics approval is pending. Moreover, maintaining ethics compliance throughout implementation requires systems for documenting protocol deviations, reporting adverse events, and ensuring all team members follow approved procedures.

Keep these key considerations in mind:

Start early. Begin the ethics process before finalizing all details — amendments can be submitted later for minor changes. Attaining initial approval starts the clock.

Identify the right authorities. Map out which ethics bodies have jurisdiction. Local university IRBs often require affiliation, national health ministry ethics committees may be needed for health research, multiple approvals may be required for multi-country studies, and some donors have their own ethics requirements.

Allow for realistic timelines. Budget significant time for the process: 2-3 weeks to prepare documents, 4-8 weeks for initial review (longer if a full board review is required), 2-3 weeks for responding to comments, 2-4 weeks for final approval, and potential additional time for local or national approvals.

Prepare a complete submission. Assemble a comprehensive documentation package, including the following: detailed protocol with background, objectives, methods; consent forms in all local languages with back-translations; survey instruments, even if still being refined; CVs and training certificates for all key personnel; data management plans with security measures; compensation structures with justification; risk mitigation and referral procedures.

Understand the level of review likely required, as this affects the timeline. An exempt review (minimal risk, specific categories) takes 2-3 weeks; an expedited review (minimal risk, not exempt) takes 4-6 weeks; a full board review (more than minimal risk, vulnerable populations) can take 2-3 months and only meets monthly.

Avoid common delays. Common reasons for delays include: incomplete applications missing required sections, consent forms using technical jargon or missing required elements, inadequate risk assessment or mitigation plans, compensation that appears coercive, unclear data protection procedures, and missing signatures or institutional approvals.

After approval, maintain compliance. This entails training all new team members on protocol, documenting and reporting any deviations, submitting amendments before

making changes, filing annual continuing review reports, and properly closing the study when it's complete.

Implementation monitoring and process evaluation

This is a comprehensive system for tracking how the intervention is actually delivered in the field, combining regular implementation monitoring routines with systematic process documentation. This involves structured check-ins (daily, weekly, or biweekly depending on intensity) using standardized tools to assess multiple dimensions of delivery, including:

- **Fidelity**: was the intervention delivered as designed?)
- **Reach**: what proportion of the target population received the intervention?
- **Dose**: the frequency and intensity of delivery
- **Quality**: how well was the intervention delivered?
- **Participant engagement**: (did participants understand and act on it?)
- **Contextual factors** affecting implementation

It includes real-time tracking of recruitment rates against targets, monitoring attrition patterns to maintain statistical power, documenting all adaptations made during delivery, and gathering feedback from implementers and participants regarding what's working and what isn't.

Why it matters

Many evaluations find no impact – not because the intervention doesn't work, but because it was never properly delivered. A lack of systematic monitoring poses a risk: for instance, after expensive endline data collection, it could be discovered that half of the treatment group never received the intervention, control groups accessed the treatment, or field staff modified the intervention beyond recognition. Process evaluation distinguishes between **theory failure** (the intervention genuinely doesn't work, even when delivered well) and **implementation failure** (it wasn't delivered properly, therefore its effectiveness cannot be judged).

This information is crucial for interpreting results. For example,if no impact was found, was it because the theory was wrong or because only 30% of participants received the full intervention? If positive effects were found, understanding what was actually delivered aids in replication efforts. Real-time monitoring allows for course correction while there's still time. For example, if recruitment is falling behind, efforts can be intensified before it threatens statistical power; if certain sites aren't delivering properly, additional support can be provided; if unexpected barriers emerge, solutions can be developed.

Process data also reveals critical insights for scaling decisions: which contexts facilitated smooth delivery, what adaptations were necessary, which implementation challenges are likely to persist at scale, and what level of quality is realistically achievable in routine conditions versus research settings.

Keep these key considerations in mind:

- **Establish monitoring routines that match implementation intensity but don't overwhelm field teams.** This can entail daily huddles for intensive interventions, weekly calls for standard programs, or monthly reviews for light-touch interventions.
- **Create simple tools**. Standardized monitoring sheets can capture essential information without creating an excessive burden of paperwork.
- **Track core implementation metrics separately from outcome data**. This can entail the per cent of planned sessions delivered, the average attendance or participation rates, per cent receiving full intervention dose, time between intervention components, or quality ratings from standardized observations.
- **Set up systems for data quality monitoring**. These can include high-frequency checks (automated daily or weekly data review for outliers, missing data, and suspicious patterns) along with back-checks (re-interviewing 10-20% of participants to verify data accuracy and catch any potential fraud).
- **Monitor assumptions affecting statistical power continuously**: Consider recruitment rates versus targets (is the project on track to reach the sample size?), overall attrition rates (what per cent was lost to follow-up?), differential attrition (is dropout higher in treatment or control?), compliance rates

(what per cent of the treatment group is actually receiving intervention?), and contamination (is the control group accessing the intervention?).

- **Document every adaptation using structured frameworks.** Specifically note what changed from protocol, why the change was necessary (barrier encountered, stakeholder request, feasibility issue), when did the change occur, who made the decision, was the change planned or reactive, and finally, if this affects the core intervention theory.
- **Create rapid feedback loops with implementers.** Use simple WhatsApp groups for real-time problem-solving, brief weekly check-in calls focusing on challenges and solutions, and monthly reviews of monitoring data to identify patterns.
- **Distinguish between core and adaptable components**. Differentiate between core components that must be maintained for the intervention theory to hold versus peripheral elements that can be adapted to context. Document both types, but treat them differently in analysis.
- **Gather participant feedback.** Build in participant feedback mechanisms through brief exit interviews after intervention sessions, periodic focus groups with participants, and suggestion boxes or hotlines for ongoing input.
- **Track contextual factors that might affect implementation or outcomes.** Note other concurrent programs or policies affecting the target population, seasonal factors (holidays, agricultural seasons, and weather), political or security situations, and health emergencies or other disruptions.
- **Maintain detailed logs.** These become crucial for interpreting results, informing scale-up decisions, and contributing to the broader evidence base on implementation challenges and solutions.

Cost-benefit analysis

This is a systematic calculation of all resources required to deliver the intervention and achieve measured impacts, producing standardized metrics that enable comparison across different interventions, delivery models, or investment options. This comprehensive accounting goes beyond simple programme budgets to capture the full economic cost of achieving outcomes, including:

- direct programme costs (staff, materials, operations)
- indirect costs often hidden in other budgets (supervision, administration, overhead)
- opportunity costs of resources used (volunteer time, government staff, participant time)
- startup versus running costs
- marginal costs of adding participants

The analysis produces metrics such as cost per child vaccinated, cost per percentage point increase in test scores, cost per year of life saved, or return on investment ratios that decision-makers can compare against benchmarks or alternative interventions. See the Cost-Benefit Analysis tool here to help identify, quantify, and compare intervention costs and benefits.

To see an example of how to apply this tool using the case study in Lebanon, refer to Appendix 2.

Why it matters

Even highly effective interventions may not be worth scaling if they're prohibitively expensive relative to alternatives. Cost-benefit analysis transforms evaluation from an academic exercise proving something works into actionable policy guidance, answering "Is this good value for money?" This is increasingly critical as donors demand evidence of both impact and cost-effectiveness, governments with limited budgets need to maximize outcomes per dollar spent, and scaling decisions require understanding how costs change with scale.

A programme achieving 10% improvement might seem successful, until realizing it costs five times more than an alternative achieving 8% improvement. Understanding cost structures also reveals opportunities for efficiency — perhaps 80% of impact could be achieved at 50% of cost by streamlining certain components, or fixed costs that seem high for a pilot would be negligible at scale. Without rigorous costing, programs may be abandoned as "too expensive" based on incomplete understanding, or scaled enthusiastically without recognizing unsustainable cost structures.

Keep these key considerations in mind:

- **Start early**. It's important to collect cost data from day one of implementation — retrospective cost reconstruction is unreliable and often impossible as receipts are lost, staff forget time allocations, and in-kind contributions go undocumented.
- **Capture all costs, not just those in the budget.** These costs can include staff time including preparation and training (even if paid by partners), volunteer time valued at local wage rates for equivalent work, government staff time even if not paid by project, participant costs (transportation, lost wages, childcare), donated materials or venues at market value, and organizational overhead reasonably attributed to the project.
- **Distinguish between different cost categories, as they behave differently at scale**. These include fixed costs (training development, initial setup) that don't increase with participants, variable costs (per participant materials, incentives) that scale linearly, and step costs (supervision, new sites) that jump at thresholds.
- **Track costs from multiple perspectives, as different stakeholders care about different numbers**. Consider the implementer perspective (what does it cost this team to run?), government perspective (what would it cost to integrate into existing systems?), societal perspective (including all costs, regardless of who pays), and participant perspective (what does participation cost beneficiaries?).
- **Calculate multiple cost-effectiveness metrics to enable different comparisons**. This can include the cost per participant reached/enrolled/completing; cost per unit of primary outcome achieved; cost per standardized effect size for academic comparisons; incremental cost-effectiveness ratios if comparing variants.
- **Compare results to relevant benchmarks**. Take note of similar interventions in the chosen context, the government's revealed willingness to pay for similar outcomes, international standards (such as WHO thresholds for health interventions), and alternative ways to achieve similar goals.
- **Include sensitivity analyses showing how cost-effectiveness changes under different assumptions**. Consider if effects persist for two years versus one year, if the intervention is implemented by government versus NGO salaries, at different scales (100, 1,000, 10,000 participants), or with different compliance or attrition rates.
- **Document cost drivers transparently. What makes this intervention expensive or cheap?** Could specific components be modified to reduce costs without sacrificing effectiveness? What economies or diseconomies of scale are likely? What hidden costs might emerge in routine implementation versus research conditions?
- **Present findings in accessible ways for policy audiences**. This can include simple cost per outcome metrics, not complex economic models, visual comparisons to familiar interventions, clear statements about confidence intervals, and practical implications for budget planning.
- **Interpret cost-effectiveness wisely.** Remember that cheapest isn't always best — sometimes higher-cost interventions are worthwhile if impacts are proportionally larger or if they reach populations that cheaper alternatives miss.

Rigorous testing strengthens the link between design and decision-making. By grounding conclusions in evidence rather than assumptions, impact evaluations enable UNICEF and partners to start identifying which interventions are effective, which require adaptation, and which should be discontinued. Through systematic measurement and transparent analysis, behavioural insights and design hypotheses are translated into credible, actionable evidence that informs decisions about scale, policy, and resource allocation.

As programmes move into the Scale phase of the DEPTHS process, generated evidence from Test Hypotheses is applied. The findings from rigorous testing guide how interventions are refined, integrated into systems, and expanded responsibly. This ensures that intervention scale-up and replication decisions are grounded in demonstrated impact, rather than assumption.

Appendix

APPENDIX 1:

Quasi-experimental alternatives when randomization is not feasible

This appendix provides additional technical guidance for practitioners designing impact evaluations under real-world, routine conditions. While the main chapter highlights randomized designs as the most reliable method for establishing causality, practical realities sometimes limit their use. In such cases, evaluators may need to consider alternative approaches to randomization that still aim to generate credible, evidence-based insights.

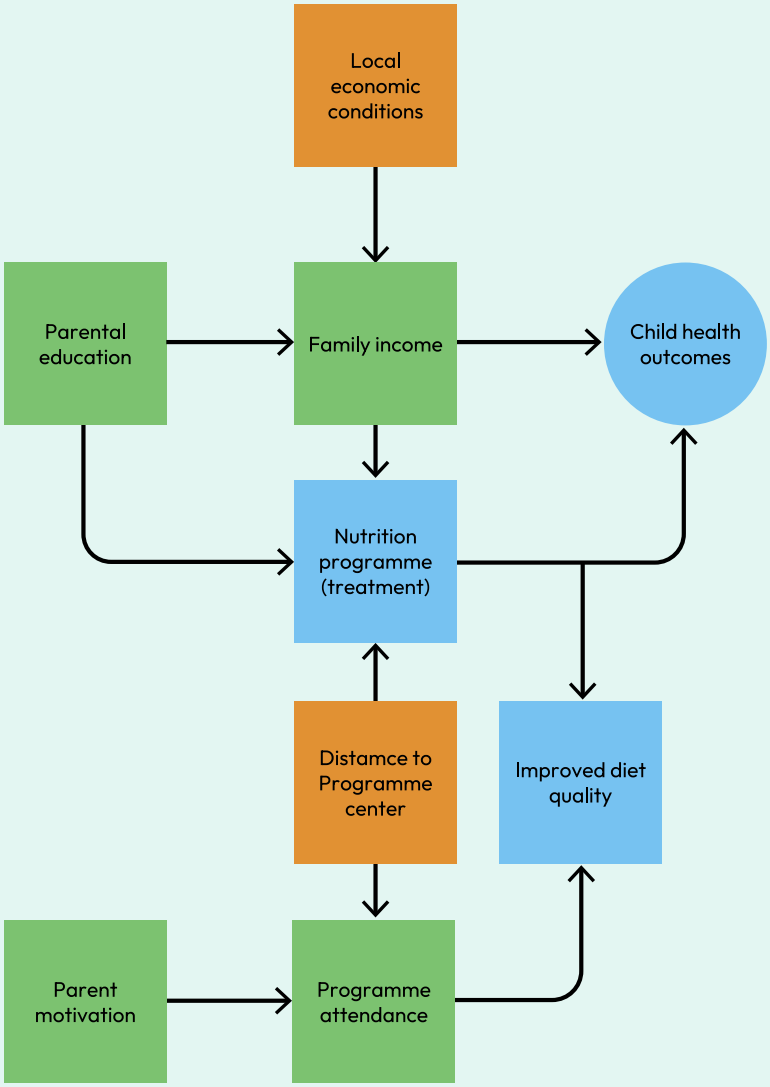
The priority in evaluation design should be to randomize, as it remains the most reliable method for establishing causality. Randomization eliminates systematic differences between groups, allowing for greater confidence that observed effects are attributable to the intervention itself. However, there are circumstances in which randomization is not possible due to political, ethical, logistical, or practical constraints. In such cases, alternative approaches may be considered.

While these alternatives can still yield valuable insights, they introduce both greater operational complexity and increased statistical demands. More importantly, such methods also carry a heightened risk of bias. Engaging experts in causal inference and evaluation design is strongly recommended when non-randomized designs are pursued.

1. **Closing backdoor paths: A DAG perspective**. When randomization is not possible, it becomes essential to identify and control for confounding factors — an approach known in causal inference as "closing backdoor paths." **Directed acyclic graphs** (DAGs) provide a useful framework for understanding this process.

   Consider a DAG illustrating the evaluation of a community nutrition programme targeting children (right). In such a diagram:

   - Arrows represent direct causal effects
   - Variables (nodes) denote factors that influence outcomes or programme participation
   - Paths between variables represent potential causal or non-causal associations



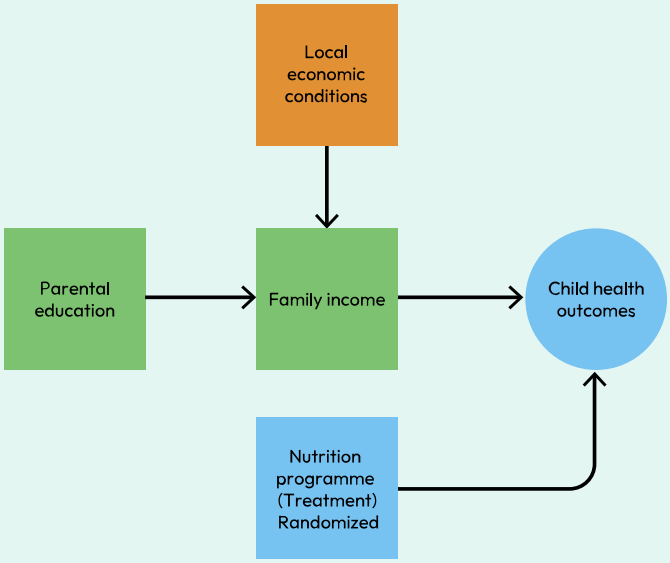
2. **Identifying backdoor paths.** A "backdoor path" refers to any pathway between the treatment (nutrition programme participation) and the outcome (child health outcomes) that does not follow the direct causal direction. These paths introduce associations that can bias causal estimates.

   In the example DAG, several backdoor paths may exist:

   - **Programme → Family Income → Child Health:** Families with higher income may be both more likely to participate in the programme and to have healthier children, regardless of the programme.
   - **Programme ← Parental Education → Child Health:** More educated parents may enrol more frequently and also provide better care.
   - **Programme ← Parental Education → Family Income → Child Health**: Parental education influences income, which in turn affects both participation and health outcomes.

3. **The goal: Closing all backdoor paths.** To isolate the causal effect of the nutrition programme, all backdoor paths must be closed. A path is considered closed when:
   - **A variable on the path is controlled for (conditioned on)**. For example: Controlling for family income closes the path Programme → Family Income → Child Health.
   - **The path contains a collider**. A collider is a variable influenced by two or more variables. For instance, "Programme Attendance" may be influenced by both "Distance to Programme Centre" and "Parent Motivation." This path is naturally closed unless the collider is incorrectly conditioned on, which would re-open the path.
   - **The path includes a mediator that is intentionally left uncontrolled**. For example, "Improved Diet Quality" lies on the causal chain between programme participation and child health. If the total effect is of interest, this variable should not be controlled for.

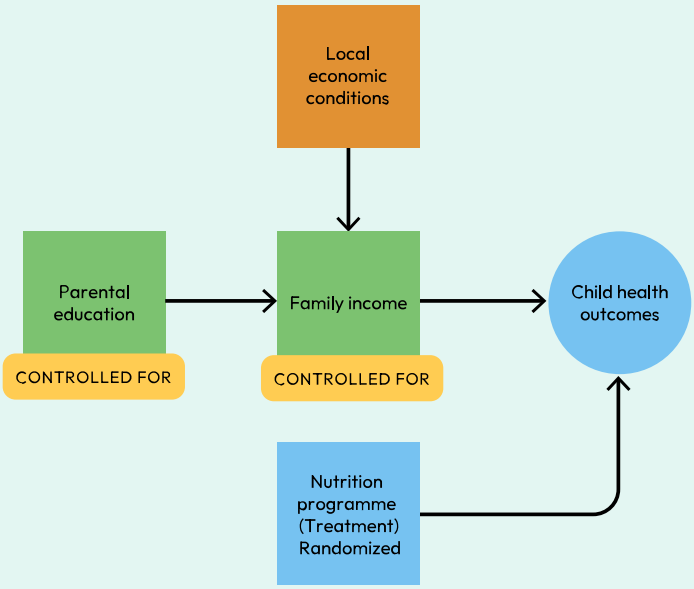
   When randomization is applied, all arrows pointing into the "Programme Participation" node are effectively cut. This renders participation independent of all confounders and closes all backdoor paths at once. This is the core advantage of randomization — it eliminates the need to identify or measure every potential source of bias.



4. **Closing backdoor paths without randomization.** In the absence of randomization, statistical methods can close backdoor paths. Two common strategies include:

   a. **Controlling for observed confounders.** Measure and adjust for variables such as parental education and family income through methods like regression adjustment or matching. This approach requires:

      - Identification of all relevant confounders
      - Accurate measurement of those confounders
      - Correct modelling of their relationships with treatment and outcomes



   b. **Using instrumental variables.** This method relies on a variable that influences programme participation but is not directly related to the outcome. For example, "Distance to Programme Centre" may determine participation without affecting child health directly. This variation can be used to estimate causal effects even in the presence of unobserved confounding.

   The opportunities and limitations of each method depend on the context and the availability of data. Different situations will offer different leverage points for closing backdoor paths, and choosing the appropriate approach requires careful consideration of both design and data constraints.

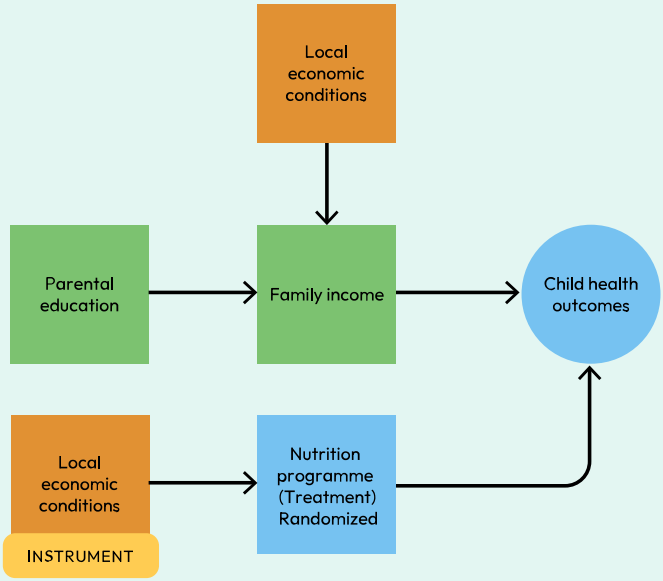


Table 3 in Step 1 of this chapter includes a description of different approaches when randomization is not possible.

APPENDIX 2:

The case of increasing childhood vaccination uptake in Lebanon

This phase also uses the case study introduced in previous phases, aimed at increasing childhood vaccination in low resource settings in Lebanon, in order to illustrate tools in the DEPTHS process.

Context of the case study:

Through the Define, Explore & Diagnose, and Prototype Designs phases, the Lebanon team identified a target behaviour to change (caregivers returning with their child for the next scheduled vaccination) and its main behavioural barriers (cognitive overload from competing tasks, emotional stress related to clinic visits, and social perceptions shaped by community views). They also tested a prototype of a potential solution: a paper-based appointment card intended to make the next vaccination visit salient and easier to remember.

The intervention was tested through a randomized controlled trial involving 12,332 un- or under-vaccinated children across 6,160 households. Households were randomly assigned to either a treatment group that received the appointment card during outreach visits or a control group that did not. The primary outcome was defined at the household level: whether at least one eligible child in the household received the next required vaccine within 21 days of becoming due.

This binary, household-level measure was selected to better reflect caregiver decision-making, recognizing that the marginal cost of vaccinating additional children in the same household is typically low. The analysis used logistic regression to estimate the intention-to-treat effect, with robust standard errors clustered at the outreach worker level. Baseline covariates included household size, nationality, and previous immunisation behaviour.

The trial found that households receiving the appointment card were significantly more likely to return for vaccination, with an increase of 6.7 percentage points compared to the control group. These results demonstrate that a behaviourally informed, low-cost intervention can measurably improve childhood immunisation uptake when targeted at key decision points. The findings have informed ongoing policy discussions and demonstrate the value of structured hypothesis testing in behavioural public health programming.

Per the field guide narrative, this appendix includes a completed example to illustrate its practical application and to support teams in using the referenced tools with ease and consistency. These tools include:

- Learning Agenda
- Implementation Plan
- Implementation Risk Mitigation
- Cost Benefit Analysis

Application of the Learning Agenda

This Learning Agenda was not developed by the original project team. It is a recreated example based on real project data and context.

The following illustrates how to apply the **Learning Agenda** introduced in the field guide, using the case study in Lebanon. This tool offers how to frame and prioritize key learning questions that guide both evaluation design and interpretation of results. The example highlights how a clear Learning Agenda — focused on what the team needs to learn, why it matters, and how findings will be used — ensures that evaluation is purposeful, actionable, and adaptive.

Learning Agenda

D E P TEST HYPOTHESES S

DEPTHS TOOLKIT

Learning Agenda

Use this worksheet to define the key elements of your evaluation plan using the PICOS framework.

Intervention: _____

Problem *What is the problem that the intervention will address?*

PICOS

Population *Who will the intervention be addressed at?*

Intervention *What is the specific intervention that will be implemented and evaluated?*

Comparison *What is the counterfactual group?*

[ ] Randomized Control Group

[ ] Non-Random Control Group (Quasi-experimental methods)

[ ] No Counterfactual (i.e., no control group)

Outcome *What are the main outcome measures that will be used to know if the intervention has been effective?*

Study and evaluation type *What type of evaluation approach will be used and what type of study?*

[ ] Impact

[ ] Process

Key lessons

How will the results be interpreted and used to improve the intervention and the outcome statement?

If the results are positive:

If the results are null or it is inconclusive:

If the results are negative (the intervention backfires):

The Lebanon team first developed a simple Learning Agenda to clarify what they hoped to learn in the evaluation and why. They identified a specific problem: caregivers in Lebanese and Syrian communities were missing routine vaccinations for their children, often because they forgot or misunderstood when to return. The intervention, a simple, behaviourally designed postcard, was meant to address this by serving as a physical reminder.

The primary learning question they posed was: "Does providing a personalized appointment reminder card increase timely childhood vaccination uptake among caregivers in low-income Lebanese and Syrian communities?"

Secondary questions included how the intervention influenced caregivers' intention to return and their recall of the follow-up date. Anticipated interpretations of results were also mapped.

For example, if uptake increased, the intervention could be scaled. If results were mixed or null, design tweaks and further testing might be needed. If backfiring occurred, the team would investigate unintended effects like mistrust or misinterpretation.

Application of the Implementation Plan

This Implementation Plan was not developed by the original project team. It is a recreated example based on real project data and context.

The following illustrates how to apply the Implementation Plan introduced in the narrative of the field guide, using the case study in Lebanon. This tool demonstrates how to translate evaluation design into structured operational planning — defining key activities, roles, timelines, and monitoring indicators. The example provides a practical reference for developing implementation plans that promote coordination, transparency, and accountability throughout the evaluation process.

Developing an Implementation Plan

To operationalize the evaluation, the Lebanon team also developed a clear Implementation Plan outlining the practical steps, responsibilities, and timeline for rollout. This plan detailed each phase, from finalizing the postcard design and conducting a small pre-test, to collecting baseline data and assigning participants to treatment or control groups. Each stage was mapped with corresponding leads (e.g. the UNICEF project team, J-PAL MENA, the Ministry of Public Health), a defined timeline, and specific indicators for tracking progress, such as enrolment rates, distribution coverage, and data quality checks. This plan served not only as a coordination tool but also as a foundation for transparency and accountability throughout implementation.

Equipped with the Evaluation Plan and Implementation Plan, the team was ready to submit to an Institutional Review Board for an ethics review and approval before they implemented and evaluated the intervention.

D E P TEST HYPOTHESES S

DEPTHS TOOLKIT

Implementation Plan

Intervention: Appointment reminder card

Use this worksheet to break down your intervention into actionable steps. For each priority area, define the key activities, assign responsibilities, set a timeline, and identify what resources and monitoring indicators will be needed to track progress.

Priority area *[Insert Priority]*	**Key activities** *[List 1–2 actions]*	**Responsible** *[Lead & supporters]*	**Timeline** *[MM/YY]*	**Resources and budget** *[Inputs needed]*	**Monitoring indicators** *[e.g., % of activities done]*
Pilot conducted	Finalise postcard design and conduct small pretest in 1–2 clinics	AIA programme team, MoPH design unit, UNICEF project team	03/23	Design consultant, transport	Postcard design finalised and feedback integrated
Baseline data collection	Collect baseline data on caregiver return rates and clinic performance	Evaluation team (J-PAL MENA), MoPH, UNICEF project team	03–04/23	Enumerator time, travel, tablets	% baseline surveys completed, data quality checks
Recruitment conducted	Identify and enrol eligible caregivers during routine clinic visits	Health workers, clinic admin staff, UNICEF project team	05–07/23	Staff orientation, clinic rosters	# of caregivers enrolled; consent rate
Assignment conducted	Randomly assign caregivers to postcard (treatment) or standard care (control)	Evaluation team, AIA implementation support, UNICEF project team	05–07/23	Randomisation protocol, data forms	Randomisation completed, balance checks verified
Intervention begins	Begin distribution of postcards to treatment group caregivers post-vaccination	Clinic staff, supervised by AIA field team, UNICEF project team	05/23	Printed postcards, tracking forms	% of eligible caregivers receiving postcard
Implementation check	Conduct midline supervision calls and field visits to assess fidelity and reach	AIA field team, MoPH district focal points, UNICEF project team	06/23	Phone/data, site visit costs	% of clinics reporting on routine, deviations noted
Intervention ends	Conclude distribution phase and stop enrolment of new participants	Clinic staff, evaluation team notified, UNICEF project team	07/23	N/A	Cut-off date enforced across all pilot sites
Data collected	Administer follow-up caregiver surveys; extract clinic return records	J-PAL enumerators, clinic M&E staff, UNICEF project team	08–09/23	Survey tools, transportation, incentives	% follow-up surveys completed, records extracted
Data cleaning and analysis	Clean datasets, conduct statistical analysis on primary and secondary outcomes	Evaluation analysts (J-PAL MENA), UNICEF project team	10–11/23	Analyst time, software	Final analysis plan completed; findings validated

Application of the Risk Mitigation Tool

The following illustrates how to apply the **Implementation Risks and Mitigation Tool** introduced in the main field guide, using the case study in Lebanon. This tool demonstrates how potential operational, contextual, and behavioural risks can be systematically identified and managed throughout the evaluation process. The example provides a practical reference for teams seeking to anticipate and address implementation challenges before they threaten evaluation integrity or programme success.

Navigating risks during rollout

The team used the Implementation Risks and Mitigation tool dynamically, not only at the planning stage but as part of active implementation. Five key risks emerged:

RISK	LIKELIHOOD	IMPACT	MITIGATION
Postcards not arriving on time	Medium	High	The UNICEF team worked with a courier service to stagger delivery schedules and provide buffer stock.
Nurses not delivering postcards with the right explanation	High	Medium	Voice note training was implemented and laminated cheat sheets were distributed.
Caregivers misplacing the postcard	Medium	Medium	The use of clear plastic sleeves and reminders to stick the postcard near the calendar at home.
Data logs not being consistently updated	Medium	High	A quick log reminder was integrated at the end of each shift via WhatsApp prompts.
Contextual disruptions (e.g. transport strike, weather)	Low to Medium	High	Local focal points were assigned for flexible reallocation of delivery routes.

Each risk was assigned to either the UNICEF team or a MoPH supervisor for follow-up, with deadlines linked to the implementation timeline.

Application of the Cost-Benefit Analysis

The following illustrates how to apply the Cost-Benefit Analysis Tool introduced in the narrative of the field guide, using the case study in Lebanon. This tool demonstrates how programme costs and benefits can be systematically identified, quantified, and compared to assess value for money. The example provides a practical reference for teams seeking to estimate the economic efficiency of behavioural interventions and to inform decisions about scaling, adaptation, or resource allocation.

Cost-Benefit Analysis

The primary goal of the intervention was to increase timely vaccination coverage among caregivers in low-resource urban clinics in Lebanon. In the absence of the intervention, data suggested that many caregivers would delay return visits, risking missed or incomplete immunization. Offering a simple behavioural prompt, the postcard aimed to change that trajectory.

Costs were relatively modest. While the study did not report financial data directly, a reasonable estimate placed the cost per printed postcard at under $0.20, including design

D E P TEST HYPOTHESES S

DEPTHS TOOLKIT

Cost-Benefit Analysis

Intervention: Appointment reminder card

Use this worksheet to weigh the costs and benefits of your intervention and decide whether it's worth scaling, adapting, or stopping.

Define the Purpose

The project aimed to improve on-time childhood vaccination through a simple, low-cost reminder postcard. Caregivers—especially low-literacy and refugee families—benefited by receiving a clear next-visit reminder. Without the intervention, drop-out rates and missed vaccinations remained high.

List All Costs

- Direct: Card design, printing, training, and distribution (~$0.20/card).
- Indirect: Staff time, supervision, and monitoring.
- Opportunity: Minimal, as delivery occurred during existing visits.
- Total: Estimated <$10,000 for pilot phase.

List All Benefits

- Direct: More children vaccinated on time.
- Indirect: Reduced dropouts, fewer missed appointments, better caregiver engagement.
- Equity: Stronger impact among Syrian and low-literacy groups.

Assign Values

- Costs valued via procurement/staffing rates.
- Benefits: 7 percentage point increase in timely returns, ~350 additional children vaccinated.
- Proxy value per timely vaccination: ~$50–$150 (based on WHO estimates).

Compare Costs and Benefits

- Cost per additional timely vaccination: ~$28
- Estimated BCR: ~3:1
- Strong return on investment.

Test Assumptions

Sensitivity checks showed results held under:

- Lower effect sizes (3–5%)
- Higher postcard costs
- Even in worst-case scenarios, benefits exceeded costs.

Make a Judgment

The intervention is cost-effective, scalable, and equitable. Recommended for scale-up in similar low-resource settings, with continued feedback loops to optimise delivery.

and distribution. Assuming a modest pilot scope (e.g. <10,000 caregivers), the total cost likely remained under $10,000, including materials, supervision, and staff time.

On the benefits side, the study demonstrated that the postcard increased return rates by seven percentage points. Public health literature suggests that each additional timely vaccination contributes to long-term benefits in disease prevention, reduced child mortality, and lower healthcare costs. Using WHO estimates, a conservative valuation of each timely vaccination could be $50–$150 in societal benefits. When applied across the intervention group, this translates into significant aggregate gains, likely yielding a benefit-cost ratio between 3:1 and 5:1.

To test robustness, the team considered a range of assumptions. Even under pessimistic scenarios (e.g. higher costs or weaker effects), the intervention still appeared cost-effective, largely due to its low unit cost and scalable design.

Importantly, the team also considered equity and inclusion. The intervention disproportionately benefited marginalised caregivers, especially Syrian families, highlighting its potential as a low-cost strategy for reducing health access disparities. The reminder postcard was not only cost-effective but also equity-enhancing, a key consideration for future scale-up.

Learn more

This field guide offers practical tools, frameworks, and worksheets to help teams apply behavioural science to real-world development challenges. However, no guide can cover everything. Behavioural science sits at the intersection of multiple disciplines, ranging from human-centred design and implementation science, to ethics, measurement, and evaluation. That's why we've included this section – for those who are curious to dig deeper, sharpen their ethical practice, strengthen implementation design, or explore how to select better outcome measures. The resources below offer curated starting points for a self-paced learning journey.

"I want more detailed step-by-step guidance on how to conduct experiments."

There are multiple manuals, resources, and courses on conducting experiments for social programmes. Some helpful free resources are the J-PAL's website on Introduction to randomized evaluations and the World Bank's Impact Evaluation in Practice Guide.

"I want to improve how I approach ethics in applied behavioural science."

Ethics is foundational to any research or behavioural project involving people. Whether the task at hand is writing consent forms, evaluating risk, or navigating power dynamics, these resources provide accessible and practical support:

- UNICEF's Ethics Toolkit for Applied Behavioural Science Projects helps teams to reflect on ethical risks early and integrate safeguards throughout implementation.
- UNICEF's Procedure on Ethical Standards in Research, Evaluation, Data Collection, and Analysis outlines the organization's official protocols and expectations.
- Informed Consent Checklist (J-PAL) is an annotated template with guidance on what to include in participant consent forms.
- UNICEF Consent Templates (see page 41) include editable samples for participants, caregivers, and gatekeepers.

"I need to take an ethics course for IRB."

There are multiple training sessions available, with some organizations even offering their own internal ethics training with a certificate allowed by different IRBs. For those seeking external training, explore the following resources:

- HHS Human Research Protection Training (US-based, free certification, ~5–6 hours)
- Tri-Council Policy Course on Research Ethics (Canada-based, free certification, ~4 hours)

"I want to explore how to design and measure implementation more effectively."

Understanding what exactly was accomplished, along with how, is essential to recognizing whether a behavioural intervention worked. The Implementation Outcome Repository offers guidance and examples for measuring constructs like feasibility, fidelity, and acceptability.

"I want to improve the way I select or adapt outcome measures."

A strong outcome measure doesn't just test effectiveness — it captures the right behaviour in the right way. If planning to test behavioural change or proxy outcomes, the Psychometric Properties of Implementation Measures reviews validity and reliability of commonly used tools in implementation science.

"I want to assess the quality and rigour of evaluation reports and studies."

If responsible for reviewing, commissioning, or interpreting studies, it's important to understand not only what a report says, but how trustworthy its findings are. These tools and articles help to assess study quality, whether reviewing an impact evaluation, implementation report, or academic article.

Assessing overall design and reporting rigour

- Gates' DAC Assessment Tool (DAT) helps to assess whether a study was well-designed, well-analysed, and clearly communicated. Originally built for clinical trials, it's applicable across sectors.
- Publishing Quantitative Papers with Rigor and Transparency article offers digestible guidance on transparency and robustness for teams writing up results.

Reviewing systematic reviews

The Evidence Project Risk of Bias Tool evaluates rigor in both randomized and non-randomized studies. It's particularly useful when reading systematic reviews or mixed-methods syntheses.

Evaluating qualitative research rigor

- Indicators of Rigor in Qualitative Research explains how to judge the credibility, transferability, and dependability of qualitative studies.
- Information Power in Qualitative Sampling offers a helpful alternative to the idea of "saturation" for justifying sample sizes in interviews.

Resources

1. Akbari, M., Nikijoo, I., Khodapanah, B., Foroudi, P., & Padash, H. (2025). Forty Years of Microfinance Research and Its Impact on Consumers: A Review and Research Agenda Using the ADO-TCM Framework. International Journal of Consumer Studies, 49(4), e70101.
2. Behavioural Insights Team. How to Run Simple Behavioural Insight Projects. 2022. https://www.bi.team/wp-content/uploads/2022/11/BIT-Handbook-How-to-run-simple-BI-projects.pdf.
3. Blanc, J. (2014). Microfinance, Debt and Over-Indebtedness: Juggling with Money, Isabelle Guérin, Solène Morvant-Roux et Magdalena Villarreal (dir.). Editions Routledge, Londres, Royaume-Uni, 2014, 316 pages. Revue internationale de l'économie sociale: recma, (334), 122-124.
4. Bloomberg. "Big Money Backs Tiny Loans That Lead to Debt, Despair and Even Suicide." Bloomberg.com, May 3, 2022. https://www.bloomberg.com/graphics/2022-microfinance-banks-profit-off-developing-world/.
5. Clemens, Michael A., and Gabriel Demombynes. "When Does Rigorous Impact Evaluation Make a Difference? The Case of the Millennium Villages." Journal of Development Effectiveness 3, no. 3 (2011): 305–339. https://doi.org/10.1080/19439342.2011.587017.
6. Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín. "Technology and Child Development: Evidence from the One Laptop per Child Program." American Economic Journal: Applied Economics 9, no. 3 (2017): 295–320. https://doi.org/10.1257/app.20150385.
7. Evaluation Hub. "Run Evaluations." https://www.bitevaluationhub.com/run-evaluations.
8. John, B. (2024, November 14). Challenges and limitations of microfinance in achieving large-scale poverty reduction and job creation [Working paper].
9. J-PAL. "Design and Iterate the Implementation Strategy." https://www.povertyactionlab.org/resource/design-and-iterate-implementation-strategy.
10. J-PAL. "Ethical Conduct of Randomized Evaluations." https://www.povertyactionlab.org/resource/ethical-conduct-randomized-evaluations.
11. J-PAL. "Impact Evaluation Methods Table." https://www.povertyactionlab.org/sites/default/files/research-resources/impact-evaluation-methods-table.pdf.
12. J-PAL. "Power Calculations Exercise." https://www.povertyactionlab.org/sites/default/files/Exercise-PowerCalcs_0.pdf.
13. J-PAL. "Questionnaire Piloting." https://www.povertyactionlab.org/resource/questionnaire-piloting.
14. J-PAL. "Data Security Procedures for Researchers." https://www.povertyactionlab.org/resource/data-security-procedures-researchers.
15. J-PAL and IPA. Implementing Impact Evaluations: Case Study. 2023. https://poverty-action.org/sites/default/files/2023-03/Case-Study-Implementing-Impact-Evaluations.pdf.

16. Karlan, Dean, and Jacob Appel. More Than Good Intentions: Improving the Ways the World's Poor Borrow, Save, Farm, Learn, and Stay Healthy. Penguin, 2011.

17. NYU Office of Research. "IRB Decision Tree." https://www.nyu.edu/content/dam/nyu/research/documents/IRB/IRBDecisionTree.pdf.

18. Shelly, Sarah, et al. "Improving Communication with Participants in Behavioural Trials." 2023.

19. UNICEF. An Evaluation of the PlayPump® Water System as an Appropriate Technology for Water, Sanitation and Hygiene Programmes. 2007. http://www-tc.pbs.org/frontlineworld/stories/southernafrica904/flash/pdf/unicef_pp_report.pdf.

20. UNICEF. Ethical Considerations When Applying Behavioural Science to Programmes with Children. Innocenti, 2021. https://www.unicef.org/innocenti/media/5186/file/UNICEF-Ethical-Considerations-Behavioural-Science-Children-2021.pdf.

21. UNICEF. "UNICEF Procedure for Ethical Standards in Research, Evaluation, Data Collection and Analysis." https://www.unicef.org/evaluation/documents/unicef-procedure-ethical-standards-research-evaluation-data-collection-and-analysis.

22. UK What Works Evaluation Hub. "Pilot Impact Studies." https://evaluationhub.eif.org.uk/pilot-impact-studies/.

23. University of California Santa Barbara Library. "Data Evaluation Checklist." https://www.library.ucsb.edu/sites/default/files/attachments/data-curation/resources/DataEvaluationChecklist.pdf.

24. World Health Organization. Monitoring the Building Blocks of Health Systems: A Handbook of Indicators and Their Measurement Strategies. 2010. https://iris.who.int/bitstream/handle/10665/44708/9789241502320_eng.pdf.

25. Australian Institute of Family Studies. Process Evaluation. 2025. https://aifs.gov.au/sites/default/files/2025-03/2502%20EES%20process%20evaluation_1.pdf.

26. Tableau. "What Is Data Cleaning?" https://www.tableau.com/learn/articles/what-is-data-cleaning.

27. BetterEvaluation. "Data Cleaning." https://www.betterevaluation.org/methods-approaches/methods/data-cleaning.