Youth Impact

Iterative A/B Testing Toolkit

A jump-start guide to embedding rigorous, rapid, and regular learning in your organization

www.youth-impact.org

About this toolkit

Organizations need fast, affordable ways to optimize program cost-effectiveness on the path to scale. Iterative A/B testing is emerging as a practical, low-cost method for social and public sector organizations to learn, adapt, and improve rigorously.

At Youth Impact, we've been running A/B tests for nearly a decade—75 in total, now averaging one per program each school term—and we support partners to run their own. This toolkit codifies what we've learned: it explains how to get started, outlines best practices, and shares tools and tips developed through years of field-based testing.

The primary audience for this toolkit is anyone in the social or public sectors—NGOs, philanthropists, bilaterals, multilaterals, and regional or national governments—looking to integrate A/B testing into their M&E systems. These guidance and tools can be adapted to your specific context. While our experience centers on optimizing youth education programs, the toolkit applies broadly to any organization seeking to develop iterative learning systems. In the spirit of continuous improvement, we keep refining our A/B testing approach, and this toolkit will be updated over time.

If you're considering A/B testing but unsure if it's the right fit, see the organizational readiness 🔗 section of the toolkit. You can also take our self-assessment quiz 🔗 to gauge suitability for your organization. Finally, feel free to contact us directly by filling out this form 🔗 to help us determine how best to support you.

This version was last updated: December 2025

Authors

Noam Angrist
Co-founder, Youth Impact; Academic Director, What Works Hub for Global Education
Noam is an economist and researcher who has published in leading journals such as *Nature*. He is committed to translating evidence into scaled policy and practice, establishing research centers, funding flows, and implementing organizations in service of this goal. He has integrated evidence into high-profile policy efforts, including co-developing the World Bank Human Capital Index education pillar.

Amanda Beatty
Principal Researcher, Youth Impact
Amanda co-leads Youth Impact's A/B testing partnerships work, helping organizations build their experimentation capacity. She brings over 20 years of education research experience across Asia and Africa with institutions such as the World Bank, Mathematica, and MCC.

Claire Cullen
Head of Research and Innovation, Youth Impact
Claire leads Youth Impact's rapid experimentation and learning systems, advancing evidence-based education in over eight countries. She has worked with Oxford University, the World Bank, UNICEF, and DFAT.

Acknowledgements

We thank the Mulago Foundation, Jacobs Foundation, the Agency Fund, the What Works Hub for Global Education, and Paul Youn for their partnership in advancing this work. Our thanks also go to our A/B testing partners, whose collaboration, insights, and commitment to learning through experimentation have shaped and strengthened this toolkit. Thanks also to our Youth Impact colleagues past and present, in particular Moitshepi Matsheng, Konstantin Büchel, Colin Crossley, Tendekai Mukoyi, Karen Clune, Angelica Towne Amporo, and Elle Brooks.

Contents

List of tools

The following tools are referenced throughout this toolkit. Use this list to quickly navigate to where the tools are discussed, or to access and download each tool.

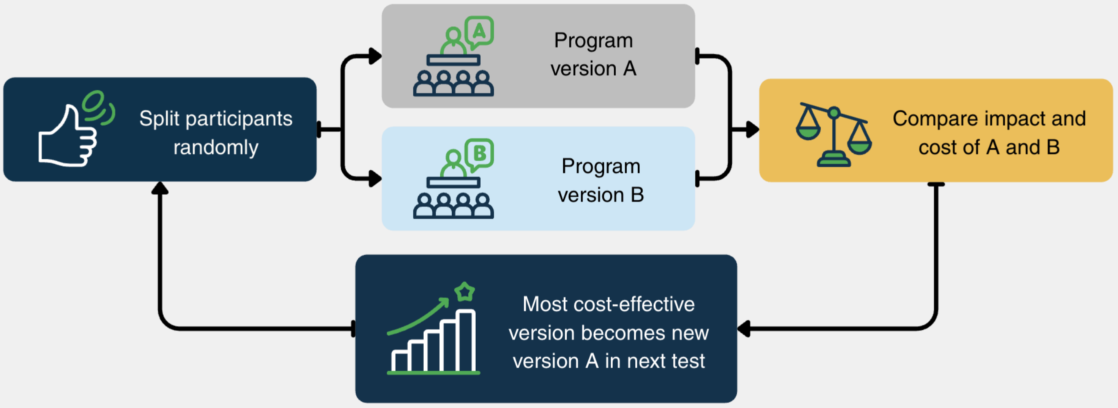
Introduction to iterative A/B testing

Iterative A/B testing is a nimble, rigorous methodology to optimize programs for greater cost-effectiveness and scalability.

This method involves (1) making a tweak or variation to an existing program; (2) randomly allocating individuals or groups into the status-quo version of the program (A) or the tweaked version (B); (3) comparing changes in outcomes and associated costs as a result of the tweak; and (4) deciding which program version to implement given cost-effectiveness results.

Figure 1 below shows this process. A central feature of A/B testing is that it allows for continuous, iterative learning and program improvement where the results of one test are fed into planning for the next test.

Figure 1: A/B testing involves randomizing individuals or groups to version A or B, comparing the relative impact and costs of A and B, and choosing the most cost-effective version of the program


Split participants randomly

Program version A

Program version B

Compare impact and cost of A and B

Most cost-effective version becomes new version A in next test

Cost-reducing vs. effectiveness-enhancing tests

A/B tests aim to address both sides of the cost-effectiveness equation – reducing cost and enhancing effectiveness. Both types of tests are essential for program optimization. Cost-reducing tests ensure scalability and sustainability, while effectiveness-enhancing tests maximize program impact. Organizations often cycle between both types of tests.

Cost-reducing
These tests aim to reduce cost while maintaining program impact. They typically *remove or simplify* a program component, for example reducing staff time or materials. Lowering these costs, while preserving impact, can result in large efficiency gains. Cost-reducing tests can make programs substantially more scalable, both for implementing organizations and by facilitating government adoption.

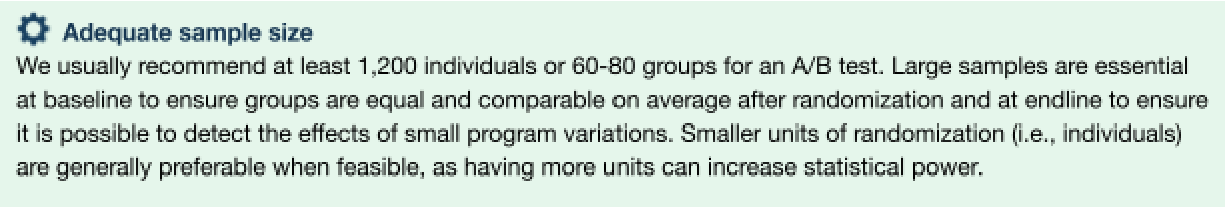
Effectiveness-enhancing
These tests aim to increase program impact at minimal additional cost. They typically add a program component. Effectiveness-enhancing tests aim to improve effectiveness to enable greater impact over time and at a larger scale. These tests also aim to improve efficiency, since enhancements have low marginal cost, while measurably enhancing impact.

Principles of A/B testing: Rigorous, rapid, regular

A/B testing is useful to implementers operating in the social and public sectors because they are rigorous, rapid, and regular. We summarize these principles in Table 1 and describe them in detail below. You can find more in our note: Angrist et al., 2024🔗.

Rigorous

Like randomized controlled trials (RCTs), A/B tests are rigorous since they use randomization to generate causal evidence. However, while RCTs typically ask "does the program work?" by comparing a treatment group to a no-program control, A/B tests ask "how can the program work more effectively, cheaply, and scalably?" by comparing different optimized versions of the same program. The focus of A/B testing is optimization rather than generating proof of concept. A/B testing also ensures all participants receive some level of programming, even if different variations, which is often necessary for programs operating at scale. Randomization and having a sufficient sample size make for a rigorous test.

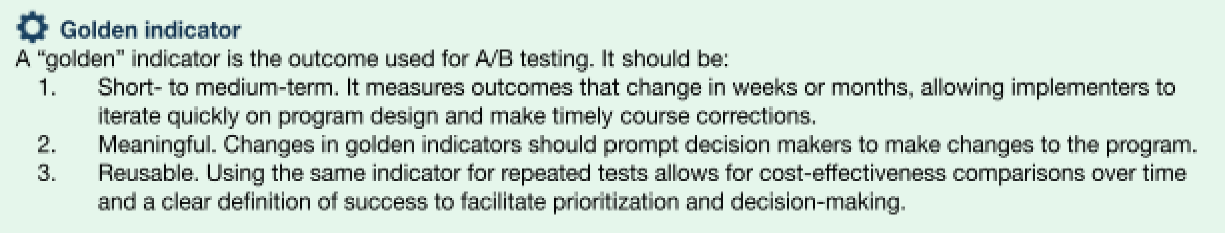


Adequate sample size

We usually recommend at least 1,200 individuals or 60-80 groups for an A/B test. Large samples are essential at baseline to ensure groups are equal and comparable on average after randomization and at endline to ensure it is possible to detect the effects of small program variations. Smaller units of randomization (i.e., individuals) are generally preferable when feasible, as having more units can increase statistical power.

Rapid

A/B testing produces results quickly, helping organizational decision makers implement program improvements within real-time implementation cycles rather than waiting years for evaluation findings. The rapid nature of A/B testing makes it particularly valuable for organizations scaling programs, where small improvements in cost-effectiveness can have large cumulative impacts. By testing variations quickly, organizations can identify and adopt improvements during the scaling process, rather than discovering optimization opportunities after programs have already been implemented at full scale and too late to make programmatic changes. A “golden indicator” facilitates A/B testing because it is measured quickly and is consequential enough to inform decisions.



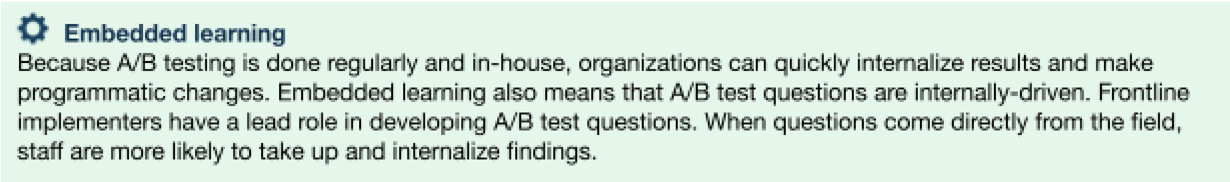
Golden indicator

A “golden” indicator is the outcome used for A/B testing. It should be:

1. Short- to medium-term. It measures outcomes that change in weeks or months, allowing implementers to iterate quickly on program design and make timely course corrections.
2. Meaningful. Changes in golden indicators should prompt decision makers to make changes to the program.
3. Reusable. Using the same indicator for repeated tests allows for cost-effectiveness comparisons over time and a clear definition of success to facilitate prioritization and decision-making.

Regular

Regular testing requires robust data systems that collect high-frequency outcome data on regular cycles—ideally monthly, termly, or quarterly. This frequency allows organizations to track changes over time and conduct multiple tests per year, enabling continuity and cumulative learning. Earlier tests inform future tests, compounding learning over time. Regular testing also enables a shift away from "single slice" learning as is often the case with one-off evaluations towards ongoing iteration and improvement. As you test and adapt more, you can strengthen learning muscle across your organization. Innovation often requires a "fail fast" approach and rarely is every A/B test transformative. A/B testing is most powerful as a process, with small effects accumulating into large returns over time. We find it often takes five to ten tries to identify a breakthrough innovation.



Embedded learning

Because A/B testing is done regularly and in-house, organizations can quickly internalize results and make programmatic changes. Embedded learning also means that A/B test questions are internally-driven. Frontline implementers have a lead role in developing A/B test questions. When questions come directly from the field, staff are more likely to take up and internalize findings.

Table 1: Like RCTs, A/B tests are rigorous; but unlike typical RCTs, they are rapid and regular

Typical A/B test attribute	Typical RCT attribute
Rigorous	
🟢 Randomized—results capture causal impacts. Multiple groups receive the same program with a tweak to test "how the program can work more effectively, cheaply, and scalably."	🟢 Randomized—results capture causal impacts. Often the main comparison is a no-program control group to test the overall question "does the program work?"
Rapid	
🟢 Results reported in weeks or months using short- and mid-term outcomes to inform real-time decisions.	🔴 Results reported over years using longer-term outcomes.
Regular	
🟢 Built into existing organizational M&E systems to directly inform program implementation and operations; multiple related tests in rapid succession to optimize cost-effectiveness.	🔴 Often a one-off high-stakes study testing novel ideas and involving external data collection.

Note: this table captures attributes of the typical RCT and A/B test, but there are exceptions. For example, some RCTs use shorter-time indicators and evaluate multiple cost-effective treatment comparisons.

When is A/B testing most useful?

A/B testing is designed for optimization and refinement, making programs cheaper, more effective, or more scalable through systematic tweaking. But not all questions are relevant for A/B testing. For example, if you're questioning the fundamental program model, you likely need different evaluation methods. Or if you are rolling out an innovation at low cost and low downside risk, it might not need a high-powered, rigorous test. Understanding what questions A/B testing can and cannot answer helps you use it effectively.

Questions A/B testing answers well

Enhancing effectiveness

- Does adding structure to coaching tools improve implementation fidelity?
- Does encouraging caregivers to engage in their child's education boost learning gains?

Reducing cost

- Can we deliver content or conduct teacher mentoring & monitoring via phone calls instead of in-person visits?
- Can we reduce training from 5 days to 3 days without losing impact?

Scale

- What is the optimal teacher-student ratio balancing effectiveness with cost?

Questions for which A/B testing is insufficient or inefficient

Deciding strategic direction

- Should we focus on health or education programming? *This is testing two different program types, not optimizing one with common indicators, and is a question of broader organizational priorities.*

Testing obvious, low-risk improvements

- Does a reminder SMS between sessions improve attendance? *If sending the SMS is cheap, easy, and highly unlikely to harm your intervention, you might want to just implement it without testing.*

Measuring long-term impact

- Does this program increase lifetime earnings? *Too long-term for rapid testing; use long-term impact evaluation.*

Case study: Meerkat Learning

Meerkat Learning 🔗 is an education-focused NGO implementing Teaching at the Right Level 🔗 (TaRL) programming in Namibia and a Youth Impact affiliate. Since 2023, Meerkat Learning has successfully built capacity to conduct A/B testing rigorously, rapidly, and regularly and is a strong example of what can be achieved with ongoing iterative A/B testing. They worked closely with regional education offices like the Khomas Regional Office to ensure buy-in and implementation feasibility. Table 2 shows highlights from several A/B tests they have implemented. Their testing philosophy emphasizes the importance of rapid iteration, implementation-driven questions emerging from field observations, and tests that both reduce costs and enhance effectiveness.

Table 2: Meerkat Learning identified large efficiency increases in several of its A/B tests

Test & research question	Versions	Results	Takeaway
1. Teacher motivational calls (2023) *Do encouragement phone calls improve data submission by teachers?*	A= In-person visits only B= Visits + calls encouraging teachers to submit TaRL data	23% reduction in missing data, 2-week faster data submission; improved learning outcomes	Simple, low-cost phone call monitoring interventions can significantly improve data submission ***Tweak was adopted***
2. Parent engagement calls (2024 - tested twice in two terms) *Do parent demonstration calls improve student learning?*	A= Standard program B= Standard program + parent demonstration calls (i.e., show parents how to do TaRL) and encourage TaRL attendance	No significant difference in learning or retention (replicated in 2 terms)	Not all promising interventions work; parent calls added burden without benefits ***Tweak was dropped***

3. Teacher coaching visits: in-person vs. remote (2024-2025)
Can phone call coaching for teachers replace in-person visits?
	A= Two school visits
	B= One visit + one support call
		Similar effectiveness, much lower costs, 30% better cost-effectiveness	Phone call coaching support maintains quality while significantly reducing costs
			Tweak was adopted

Meerkat Learning's systematic approach to A/B testing yielded actionable improvements. These lessons enabled the organization to identify large cost savings in ongoing teacher coaching at scale and to improve program and data collection efficiency. The A/B testing process made it possible for the organization to conduct multiple tests each year—a dramatic acceleration in learning. Even null results proved valuable by preventing misallocation of resources to ineffective interventions.

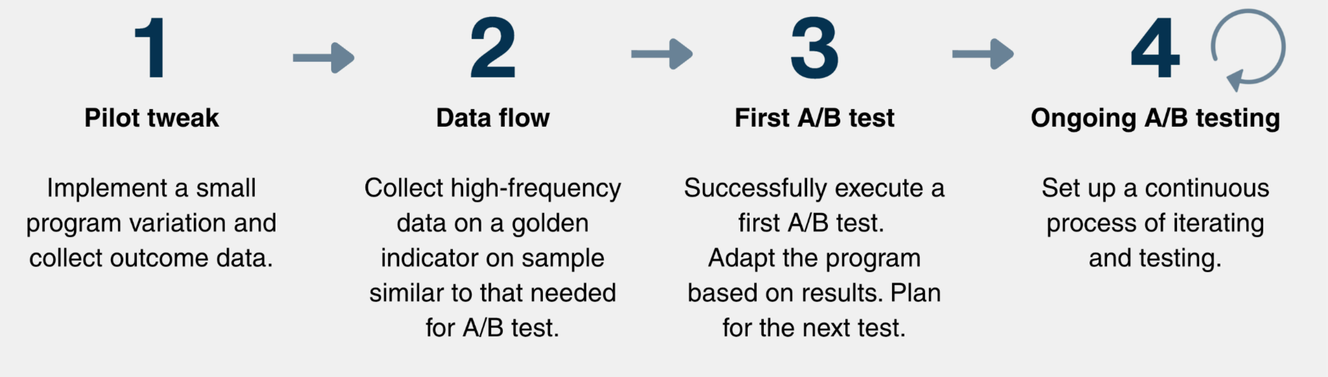
Key success factors included integrating A/B testing into routine monitoring rather than treating it as “special projects;” focusing on field-driven questions; maintaining realistic expectations about null results; and ensuring detailed cost tracking for comprehensive cost-effectiveness analysis. This experience suggests that systematic experimentation in education can compress traditional learning timelines while building institutional capacity for evidence-based decision making.

Four steps to jump-start A/B testing

We have developed a structured, four-phase approach to help organizations successfully start their A/B testing journey. We have found that organizations are most successful when they start with a few focused start-up steps, and expand into full A/B testing over time.

As shown in Figure 2 below, the process begins with *Phase 1: Pilot Tweak*, where you'll implement a simple program variation at small scale to practice the mechanics without the complexity of randomization. Next, in *Phase 2: Data Flow*, you'll establish or strengthen your data systems to ensure you can collect high-frequency, large-scale data on your golden indicators. In *Phase 3: First A/B Test*, you'll conduct your first formal A/B test, analyze the results, and begin your journey of iterative program improvement which is *Phase 4: Ongoing A/B Testing*. Each phase builds on the previous one, systematically developing your organization's capacity to use evidence for continuous program optimization. This toolkit explains each phase in detail, including specific steps, tips for success, usable tools, and what you'll want to complete before moving to the next phase.

Figure 2: We recommend a four-phase process to build the A/B testing muscle



1 Pilot tweak

Implement a small program variation and collect outcome data.

2 Data flow

Collect high-frequency data on a golden indicator on sample similar to that needed for A/B test.

3 First A/B test

Successfully execute a first A/B test. Adapt the program based on results. Plan for the next test.

4 Ongoing A/B testing

Set up a continuous process of iterating and testing.

Phase 1: Pilot tweak

Goal

The primary goal of this phase is to implement a small program variation (tweak) and collect outcome data, focusing on an easy-to-implement change to prepare for more ambitious tweaks in subsequent phases. This phase serves as a low-cost tryout to build organizational confidence in implementing variations and to generate learning questions for future A/B tests. Think of it as a "practice run" that helps teams experience the process of implementing different program versions before committing to full-scale testing. Success in this phase is primarily about gaining implementation and operational experience—not necessarily finding significant program improvements. Even if the tweak doesn't yet show promising results, the operational learning is the most important outcome of this phase.

Process

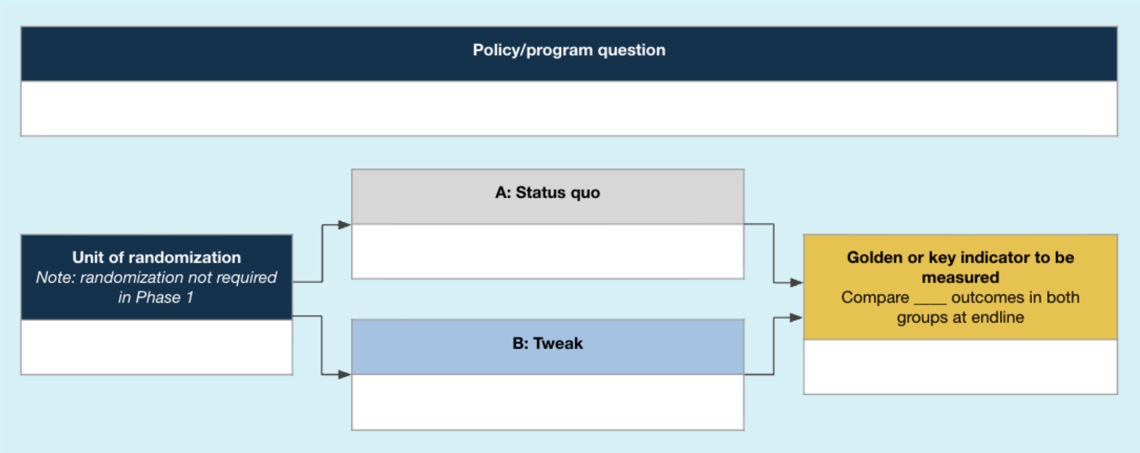
Plan

Implement

Reflect

Select a simple variation: Choose a straightforward program modification that doesn't require extensive resources or drastic changes to current operations. At this stage, the focus should be on practicing the process of implementing tweaks, not necessarily finding a groundbreaking innovation. Note you do not need to randomly allocate units to groups A or B; you can make this allocation purposefully. The objective of this phase is to demonstrate your ability to implement a variation rather than follow a randomization protocol. You can use Tool 1 🔗 to brainstorm ideas.

Tool 1: The tweak design tool allows you to brainstorm program variations against your business-as-usual implementation model



Policy/program question

Unit of randomization
Note: randomization not required in Phase 1

A: Status quo

B: Tweak

Golden or key indicator to be measured
Compare ____ outcomes in both groups at endline

Plan **Implement** Reflect

Implement at small scale: The tweak should take place in a small number of sites—at most 10 sites (total across status quo and tweak groups) or 100 individuals. This limited scale keeps the pilot manageable while still providing enough exposure to learn from the experience.

Keep the timeline short: The tweak implementation should be brief, taking at most two months from start to finish. This maintains momentum and prevents the pilot from becoming an extended, resource-intensive project.

Collect relevant data: During the implementation, collect data on relevant, high-frequency outcomes that could potentially serve as "golden indicator" candidates for future A/B tests. We discuss golden indicators in depth in Phase 2. For Phase 1, choose an indicator that you already collect data on. You do not need to determine its suitability for Phase 2 onwards yet. Aim to collect at least two data points during this period to practice the measurement process.

Plan Implement **Reflect**

Document the experience: After completion, the team should reflect on what they've learned about implementing the tweak and any insights that might inform future A/B testing. Once you've completed your pilot tweak implementation, we recommend preparing a concise writeup (slides or narrative format) that includes:

- **Tweak description**: Clear explanation of your status quo (A) and tweak (B) versions
- **Implementation summary**: Timeline, number of implementation sites/units, and how you selected sites
- **Indicators tracked**: Which key indicators you measured and how you collected the data
- **Insights gained**: What you learned about implementation feasibility, measurement approaches, and any preliminary outcome observations
- **Plans for Phase 2**: Your approach to scaling up data collection systems based on Phase 1 learnings

This writeup will allow you to (a) keep track of your experience, (b) share your experience with colleagues, and (c) help structure thinking to plan for Phase 2. You can use Tool 2 🔗 to write up your results if helpful.

Phase 1 tips

Do	Don’t
🟢**Start simple**: Choose an easy variation that doesn't require extensive resources	🔴**Overthink the tweak**: Don't let perfect be the enemy of the good
🟢**Jump in quickly**: Begin implementation without excessive planning	🔴**Spend excessive time** debating the "right" variation to test
🟢**Emphasize operational learning**: Focus on process learning rather than results	🔴**Design overly complex** tweaks that are difficult to implement
🟢**Involve frontline staff:** Engage program implementers in tweak idea brainstorming	🔴**Try to perfect** the implementation before moving forward
🟢**Document challenges**: Record implementation difficulties to inform future test designs	🔴**Work in silos**: Ensure M&E staff are coordinating closely with program staff

Phase 2: Data flow

Goal

The objective of this phase is to pressure test and enhance your monitoring system by collecting high-frequency data on a useful indicator from a large number of sites, similar to the scale needed for future A/B testing. This helps ensure your team can collect data at the necessary scale and frequency for successful A/B testing and develop a mature enough M&E system to support rigorous iterative testing. Even if your system isn't yet A/B testing "ready," you could use the data flow phase for example to increase the number of observations you collect regularly, develop an ID system, and/or maximize survey response rates and/or streamlined access to administrative data. This phase does not involve randomizing units to A or B – the goal is to enhance your data system, or at a minimum to watch the data "flow" and ensure readiness for A/B testing.

When entering data flow, organizations typically fall into one of four stages of readiness.

Nearly ready to start A/B testing	
Data already flowing: Your organization already collects high-quality data on relevant indicators at the necessary scale and frequency. You may move quickly through this phase.	
More work required before starting A/B testing	
Monitoring system needs more frequently-collected data: Your organization has a golden indicator but needs to (a) collect data on more units, (b) collect data at higher frequency, or (c) improve data quality (e.g., reduce attrition, improve accuracy or reliability). Organizations in this stage may want to cycle through several rounds of data collection and implementation as they are testing out system improvements before they move to Phase 3. (See below for a discussion of golden indicators.)	**Missing a golden indicator:** Your organization collects data at high-frequency and on enough units for an A/B test but you have not identified an indicator that is sufficiently meaningful or relevant for programmatic decision making. Your team will likely need to work on indicator development, reflecting your theory of change. Organizations in this stage might need to spend several rounds of data collection piloting and iterating on an indicator to identify one that is "golden."
Significantly more work required before starting A/B testing	
Need a major monitoring system upgrade: Your organization will need to work on indicator development and M&E system strengthening before entering data flow. This will likely take multiple rounds of data collection and implementation to ready your system for A/B testing.	

Process

Plan Implement Reflect

Check your routine monitoring data: If you already collect routine monitoring data on potential golden indicators at scale, review your data using the below attributes that signal readiness for A/B testing.

Data attributes signaling readiness for A/B testing

✔ **Scale:** Sufficient number of observations (60-80 clusters or 1200+ individuals) needed for an adequately powered A/B test

✔ **Frequency**: Data collected weekly, monthly, or quarterly—not semi-annually or annually

✔ **Panel:** Data collected on the same units like schools, children, teachers, etc over time

✔ **System for tracking panel data**: An ID system to link units over time

✔ **Low missingness/attrition** between data collection points

✔ **Variation** in the golden indicator, i.e., there should be room for improvement

Select or confirm your "golden indicator:" Identify an outcome your leadership uses or could use to make decisions about program implementation and optimization.

Key features of a golden indicator

✔ **Meaningful for decision-making:** The indicator must be consequential enough that if it changes as a result of a program variation tested in your A/B test, decision makers will act on the results and adopt the more effective version.

✔ **Shows variation across the sample:** The indicator should have sufficient variation in your data, meaning beneficiaries show a range of outcomes. High variation indicates room for improvement and ensures you'll be able to detect meaningful differences between versions A and B in your test. If everyone scores the same or nearly the same on your indicator, you won't be able to see the impact of program changes.

✔ **Responsive within weeks or months, not years:** Golden indicators should be positioned in the upper half of your theory of change – downstream enough to reflect meaningful program effects, but not so far downstream that they take years to manifest. You need outcomes that respond to program changes within your testing timeframe so you can make decisions and iterate quickly.

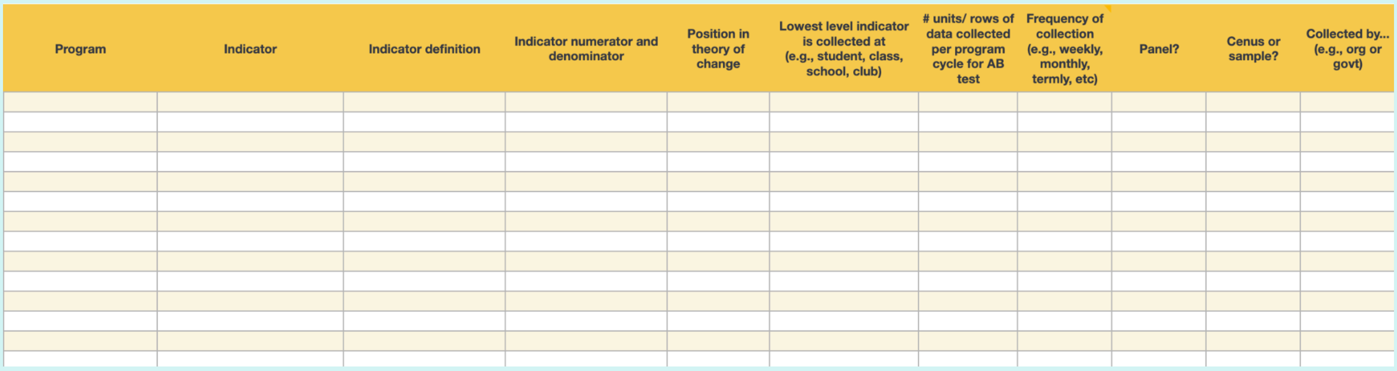
✔ **Feasible to collect at high frequency using internal systems:** Your organization must be able to collect data on this indicator regularly (ideally monthly, termly, or quarterly) using in-house capacity rather than relying on external data collection firms. If collecting the data requires hiring outside consultants or waiting for annual surveys, it won't support the rapid, regular testing cycle that makes A/B testing valuable.

We strongly recommend narrowing in on one primary indicator that you can use for repeated tests over time. Using the same indicator across multiple tests allows you to compare results from different program variations and track cumulative improvements in your program's performance.

Having one clear indicator also prevents selective reporting—there is one pre-specified outcome that serves as your north star for decision making, and success is clear from the outset. Of course, your team can and should examine results for other relevant indicators as part of your broader M&E system, but only one main indicator should drive your A/B testing decisions.

If you don't have a perfect "golden indicator," select a next-best "bronze" or "silver" indicator that can serve as a proxy for measuring program success. (See examples on the next page.) You can use the golden indicator shortlisting Tool 3🔗to guide your decision. If your organization cannot easily identify a golden, bronze, or silver indicator, it may need to work on indicator development before starting the data flow phase.

Tool 3: The golden indicator shortlisting tool helps identify the most useful indicator for your A/B tests

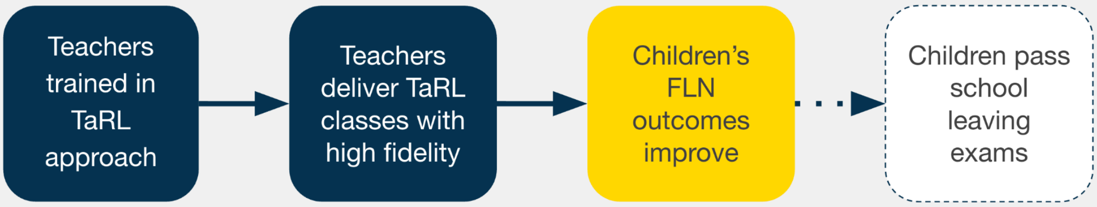


Program	Indicator	Indicator definition	Indicator numerator and denominator	Position in theory of change	Lowest level indicator is collected at (e.g., student, class, school, club)	# units/ rows of data collected per program cycle for AB test	Frequency of collection (e.g., weekly, monthly, termly, etc)	Panel?	Cenus or sample?	Collected by... (e.g., org or govt)

Golden indicator example: TaRL theory of change

Children's foundational literacy and numeracy (FLN) outcomes (i.e., the ability to read and comprehend a short story and compute basic mathematics operations) used in programs like TaRL 🔗 and structured pedagogy serve as the optimal golden indicator because they directly measure the program's primary goal while being rapidly responsive to program changes. Figure 3 below illustrates a basic theory of change for TaRL. Unlike input metrics (teachers trained) which don't guarantee short-term improvements in learning, or long-term impact metrics (exam pass rates) which take years to manifest, FLN outcomes sit in the sweet spot: they drive programmatic decisions and can be measured and respond within weeks or months to implementation changes. When selecting a golden indicator, look for metrics that occupy this ideal position in the theory of change – responsive to program changes within the testing timeframe but drive decisions about program effectiveness. That is, they are *fast enough* and *meaningful enough*.

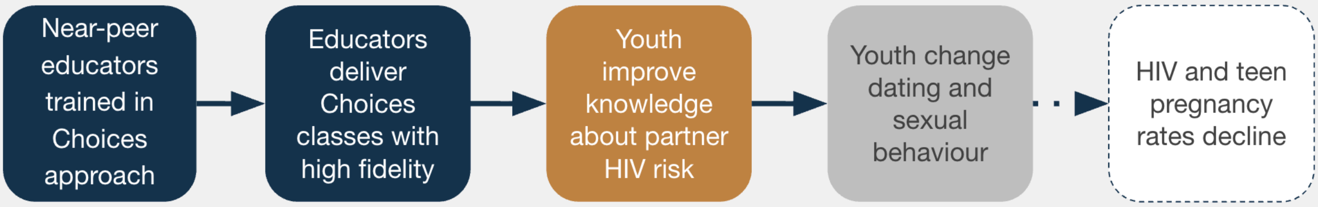
Figure 3: FLN outcomes are the golden indicator for TaRL


Teachers trained in TaRL approach
Teachers deliver TaRL classes with high fidelity
Children's FLN outcomes improve
Children pass school leaving exams

Bronze and silver indicator example: Choices theory of change

In Choices 🔗, our HIV and teen pregnancy prevention program, the theory of change progresses from inputs (near-peer educators trained) to impacts (reduced HIV/pregnancy rates) (Figure 4). With Choices, identifying the golden indicator is more challenging than with TaRL: knowledge about partner HIV risk is easier to measure frequently but may not predict impact, while HIV/pregnancy rates take too long to materialize for A/B testing. In this case, we use knowledge as a "bronze indicator" and the key indicator for rapid testing, since this can change over several weeks. We simultaneously monitor behavior change as a "silver indicator" at less frequent intervals to validate that knowledge gains lead to behavior changes. We confidently use these specific bronze and silver indicators for rapid testing because they were validated by two longer-term randomized controlled trials that established the relationship between the indicators and longer-term impact (reduced HIV and teen pregnancy) (Angrist et al., 2019; Dupas, 2011 🔗). This example illustrates that golden indicators exist on a spectrum —some programs have clear sweet spots like learning outcomes in TaRL, while others like Choices may need one key indicator (knowledge) with a supporting indicator (behavior) to ensure meaningful optimization.

Figure 4: Knowledge and behavior change indicators are the silver and bronze indicators for Choices


Near-peer educators trained in Choices approach
Educators deliver Choices classes with high fidelity
Youth improve knowledge about partner HIV risk
Youth change dating and sexual behaviour
HIV and teen pregnancy rates decline

Plan **Implement** Reflect

Utilize a robust ID system: When planning to match individuals or clusters (e.g., students in classrooms and schools) across datasets, for example a baseline and endline survey, you will want to establish a tracking system *before* data collection. You will want to make sure that IDs are unique and consistent. We recommend using existing IDs when possible; for example if implementing in-school programming, the schools may already have school IDs or a way of classifying classes or streams like class roster numbers.

Collect data at high frequency: Gather data on your selected indicator, optimally every two to six weeks, though collecting several times a year might be feasible when first getting started. The key is establishing a regular rhythm of data collection that aligns with your program cycles and provides enough data points to track changes over time. You might use the data flow phase to increase the frequency of data collection.

Collect data at scale: Gather data from a sample size similar to what you'll need for A/B testing: at least 60-80 sites (schools, classrooms, clinics) or 1200 individuals (students, teachers, beneficiaries) depending on the unit of randomization. Large samples are essential for ensuring you can detect meaningful differences between program variations in future A/B tests.

Extend collection over time: Continue data collection over several months, ensuring you collect multiple data points for each unit (child, center, school, etc.). This timeline helps establish the consistency of your measurement approach.

Minimize data attrition: Since A/B testing means that data is collected at high frequency, it should be relatively easy to ensure that you can collect data on the same units at baseline and endline. But your team will want to take steps to minimize the possibility of seeing attrition, such as scheduling data collection at times when you know respondents are most likely to be available (e.g., not close to school holidays or during harvest time if relevant).

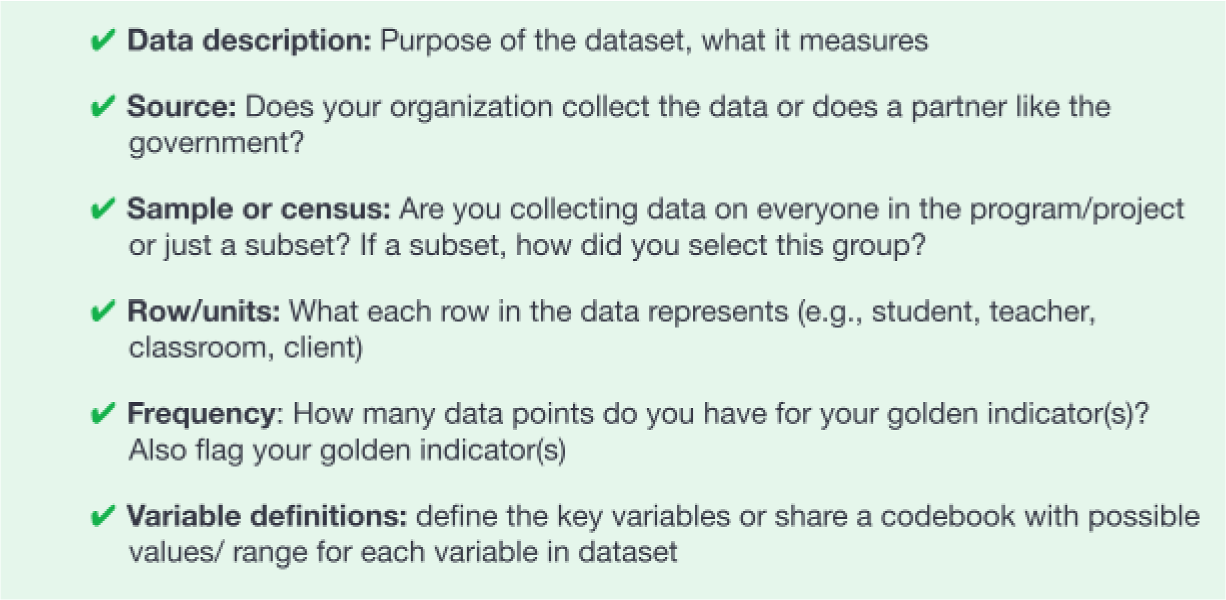
Process data efficiently: Develop systems to clean, organize, and analyze your data relatively quickly after collection, ideally within about two weeks. This rapid turnaround ensures that insights can inform timely program decisions and prepares you for the quick iteration cycles of A/B testing.

Plan Implement **Reflect**

Document the experience: After Phase 2 completion, the team should reflect on what they've learned about scaling data systems and any insights that will inform the upcoming A/B test. When you've completed data flow, it is helpful to write up your experience (in slides or narrative format). You can use Tool 4 🔗 as a template to write up the experience, including:

- **Golden indicator description:** Clear explanation of which indicator(s) you're tracking, indicator definition(s), rationale for selection; if you used the same indicator(s) as you used in Phase 1 and why/why not
- **Implementation summary:** Timeline, number of data collection sites/units, frequency of data collection, data collection process, challenges encountered, and solutions implemented
- **Data quality observations:** Describe sample characteristics (share basic descriptive statistics), reliability issues, validation approaches, attrition
- **Insights gained:** What you learned about indicator usefulness for A/B testing, data system scalability, and data system readiness for rigorous testing
- **Plans for A/B testing:** Your approach to the experiment based on Phase 2 learnings, including timeline and next steps

Collect and review your data: This step allows you to assess data quality, identify potential analysis approaches, and tailor the next steps to your specific context and needs. We recommend including the following if you consider asking for help to review your data or organizing your data in this way if you review internally.



✔ **Data description:** Purpose of the dataset, what it measures

✔ **Source:** Does your organization collect the data or does a partner like the government?

✔ **Sample or census:** Are you collecting data on everyone in the program/project or just a subset? If a subset, how did you select this group?

✔ **Row/units:** What each row in the data represents (e.g., student, teacher, classroom, client)

✔ **Frequency**: How many data points do you have for your golden indicator(s)? Also flag your golden indicator(s)

✔ **Variable definitions:** define the key variables or share a codebook with possible values/ range for each variable in dataset

Phase 2 tips

Do	Don't
🟢 **Consider the relationship** between golden indicators and impact carefully	🔴 **Choose indicators** that your team will not use for decision making
🟢 **Ensure golden indicators can change** over weeks or several months, not many months or years	🔴 **Collect data on** too many indicators
🟢 **Choose stable indicators** that can be measured regularly over time in existing M&E systems	🔴 **Establish your data tracking system** after you have already started data collection
🟢 **Engage implementation staff** in indicator selection and results interpretation	🔴 **Work in isolation** from program teams
	🔴 **Start ad hoc** or one-off surveys to capture data

Phase 3: First A/B test

Goal

The main goal of this phase is to successfully execute a first randomized A/B test with adequate sample size, assess the mechanics of running the test, adapt the program based on results, and plan for the next test. This is the beginning of your journey of iterative program improvement and optimization. The first test need not be your "best" test. It is the start of a process to build up your capacity for A/B testing.

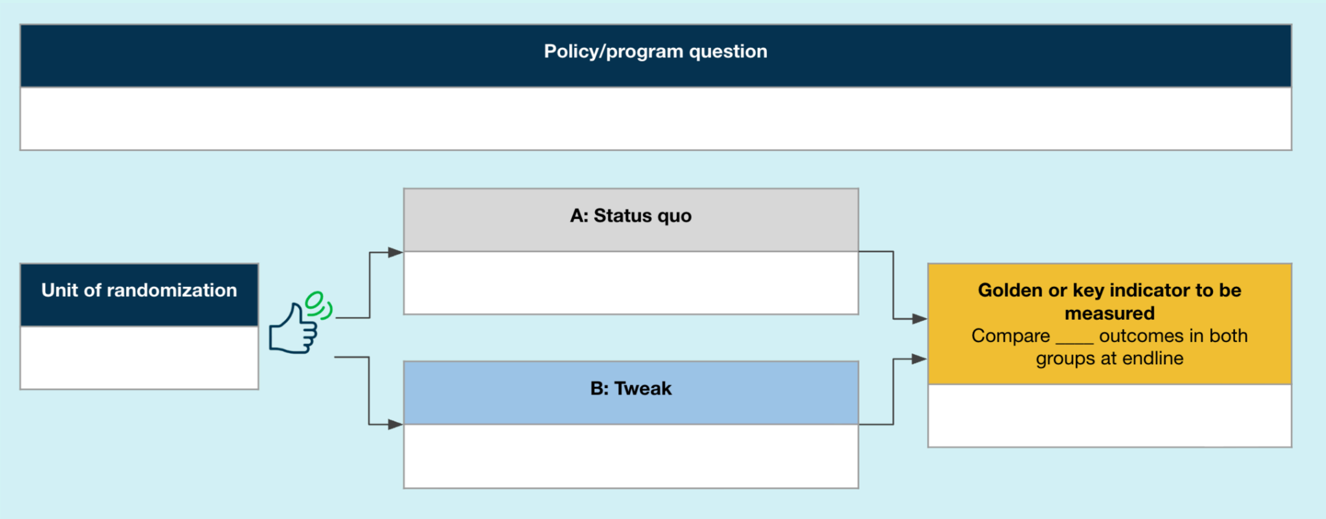
Process

Plan Implement Reflect

Select a simple question for your first test: The variation you test (Version B) should be realistic and implementable within your current operations; something your team can implement with fidelity. Avoid overly ambitious modifications for your first test. The goal is to help you learn the mechanics of testing.

You can use Tool 5 🔗 to illustrate your test ideas. Note this is slightly different from the tool used in Phase 1/tweak since in Phase 1 you did not need to use randomization. In Phase 3 this first test will be randomized.

Tool 5: This tool will allow you to visually depict your version A and B ideas



Policy/program question

Unit of randomization

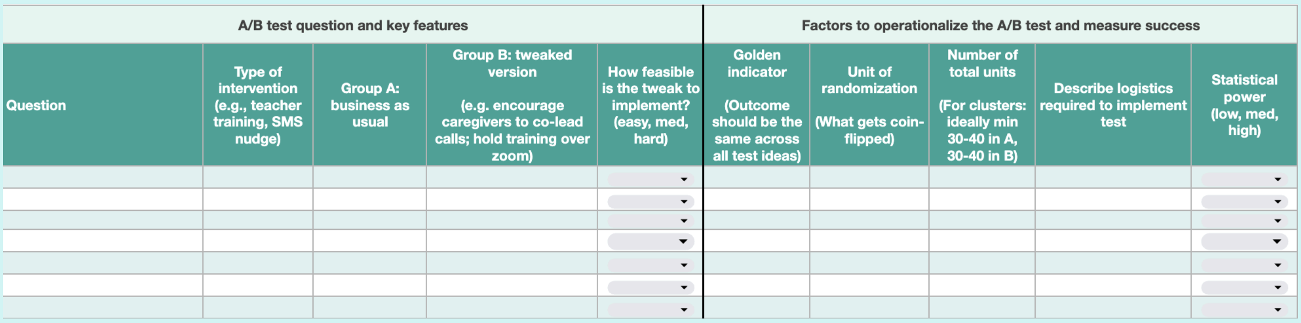
A: Status quo

B: Tweak

Golden or key indicator to be measured
Compare ____ outcomes in both groups at endline

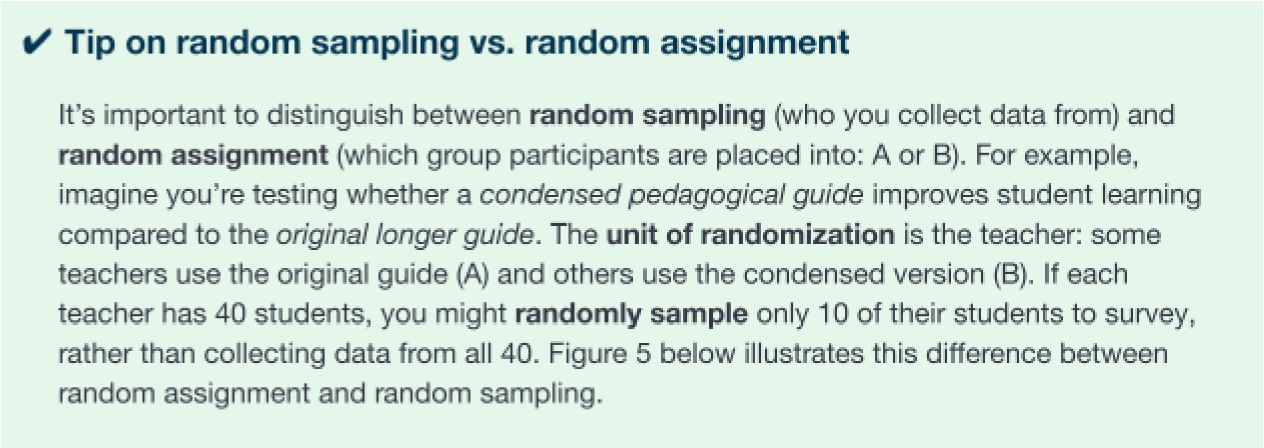
Once you have illustrated several ideas, you can use this brainstorming tool to help narrow them down. Importantly, the tool asks you to consider feasibility since this is a key factor in getting started with A/B testing. See the A/B testing question brainstorming tool for Phase 3 in Tool 6 🔗. Tool 6 also helps you consider other important factors like your golden indicator, unit of randomization, statistical power that this test will have, and other logistics needed to make this test possible.

Tool 6: This sheet will allow you to see your A/B testing ideas in one place and weigh strengths and weaknesses of each



A/B test question and key features					Factors to operationalize the A/B test and measure success				
Question	Type of intervention (e.g., teacher training, SMS nudge)	Group A: business as usual	Group B: tweaked version (e.g. encourage caregivers to co-lead calls; hold training over zoom)	How feasible is the tweak to implement? (easy, med, hard)	Golden indicator (Outcome should be the same across all test ideas)	Unit of randomization (What gets coin-flipped)	Number of total units (For clusters: ideally min 30-40 in A, 30-40 in B)	Describe logistics required to implement test	Statistical power (low, med, high)

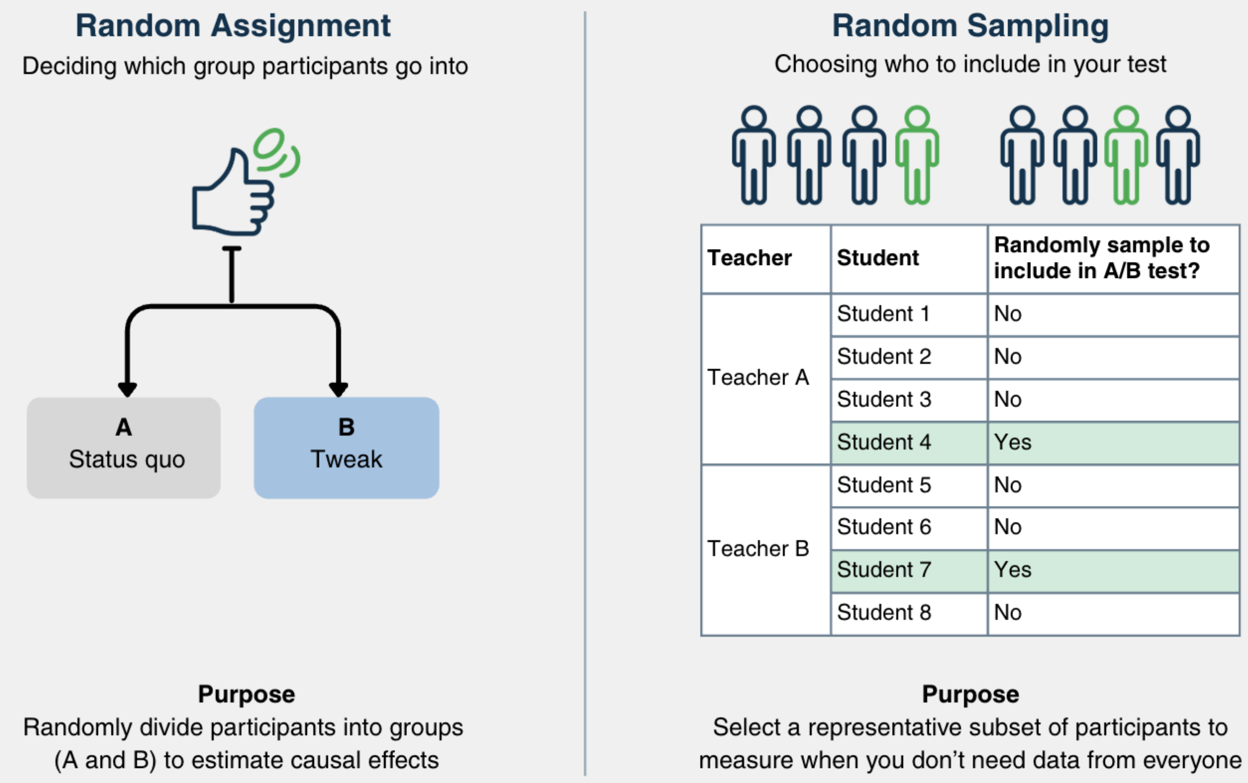
Determine the unit of randomization: When feasible, randomizing at smaller units (individuals rather than clusters) can increase statistical power and provide more precise estimates of program effects, especially when there is low likelihood of spilling over to the other group in your test. The choice of randomization unit must balance statistical considerations (maximum units) with implementation feasibility, budget, and ethical or political concerns. Organizations should select the smallest randomization unit possible while avoiding contamination between A/B test groups. Programs implemented in schools might randomize at the classroom or student level rather than school level when feasible, while programs implemented in communities might be able to randomize at the household level rather than community level.



✔ Tip on random sampling vs. random assignment

It's important to distinguish between **random sampling** (who you collect data from) and **random assignment** (which group participants are placed into: A or B). For example, imagine you're testing whether a *condensed pedagogical guide* improves student learning compared to the *original longer guide*. The **unit of randomization** is the teacher: some teachers use the original guide (A) and others use the condensed version (B). If each teacher has 40 students, you might **randomly sample** only 10 of their students to survey, rather than collecting data from all 40. Figure 5 below illustrates this difference between random assignment and random sampling.

Figure 5: Random assignment determines what version of the program someone receives, and random sampling determines who you collect data from



Random Assignment
Deciding which group participants go into
A
Status quo
B
Tweak
Random Sampling
Choosing who to include in your test

Teacher	Student	Randomly sample to include in A/B test?
Teacher A	Student 1	No
	Student 2	No
	Student 3	No
	Student 4	Yes
Teacher B	Student 5	No
	Student 6	No
	Student 7	Yes
	Student 8	No

Purpose
Randomly divide participants into groups (A and B) to estimate causal effects

Purpose
Select a representative subset of participants to measure when you don't need data from everyone

Ensure adequate sample size: A/B testing requires large enough samples to detect meaningful statistical differences between program variations, especially when program improvements might be subtle. Inadequate sample sizes lead to inconclusive results and missed opportunities to identify valuable program improvements. Organizations should plan for sample sizes that provide sufficient statistical power to detect realistic effect sizes within their program context. Organizations with smaller programs may need to repeat a test across multiple program cycles to achieve necessary sample sizes, pooling data for analysis.

Some helpful tools for determining adequate sample size include a guide 🔗 by the Abdul Latif Jameel Poverty Action Lab (J-PAL), a helpful tool 🔗 from the Evidence in Governance and Politics (EGAP) network or a similar tool 🔗 from J-PAL.

Conduct randomization: Randomly assign units to either the status quo (Version A) or the tweak (Version B). You can use whatever method works best for your team to conduct randomization or our Tool 7 🔗. (See more guidance on how to conduct randomization from J-PAL 🔗.) Be sure to check that your A and B groups are balanced using baseline variables. Document your randomization process for transparency. Be sure to safely file your original randomization assignment!

Tool 7: You can use this tool to divide units into Groups A and B

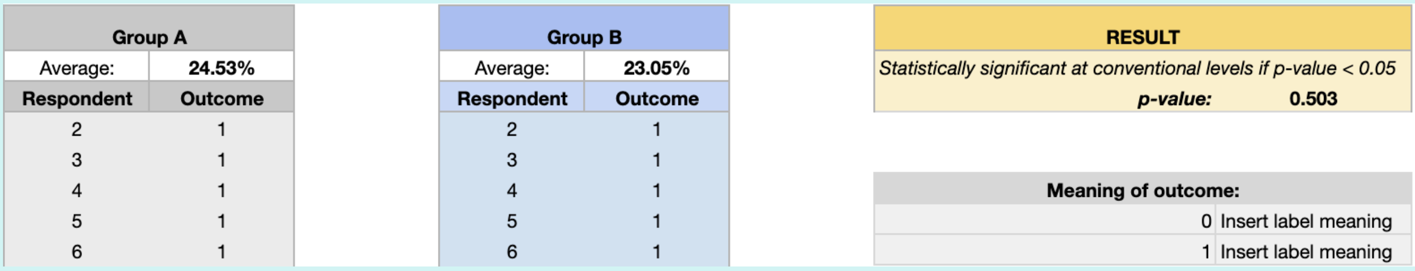
Unique student/ class/school ID	Random number generator	Randomized assignment
1001	0.79	Tweak
1002	0.06	Status quo
1003	0.31	Status quo
1004	0.49	Status quo
1005	0.70	Tweak

Implement with fidelity: Ensure that the program variations are delivered as planned to their assigned groups. Monitor implementation to confirm that the status quo and variation are being implemented correctly and that there is no contamination between groups. For example, if testing the added impact of sending homework assignments via SMS, you could verify fidelity by asking a small sub-sample of tweak group households whether they received the SMS and whether the child completed the assigned homework. Similarly, if testing the additional effect of teacher mentoring, you could ask mentors to share photos of their weekly mentoring notes and confirm with teachers that mentoring sessions occurred as planned.

Collect outcome data: Gather data on your golden indicator for all units in both Version A and Version B groups. Ensure that both groups are asked identical outcome questions, since the impact of the tweak is calculated as the difference in average outcomes between Group A (status quo) and Group B (tweak). Use the same data collection procedures established during your data flow phase.

Analyze results, document, and share learnings: M&E staff should analyze the data, comparing outcomes between the two groups using appropriate statistical methods. Tool 8 🔗 is a simple tool for calculating p-values for a difference in means across your A and B groups. While academic research usually looks for a p-value less than 0.05 to be confident that differences between A and B groups are statistically significant, A/B testing decisions can use more flexible thresholds. For example, 0.10 or higher might make sense, depending on factors such as the cost or risk of the tweak (if low, maybe you can accept higher p-values and more uncertainty) and your team's appetite for decision making under uncertainty. In practice, the acceptable level of certainty depends on the potential risks and rewards of adopting Version B.

Tool 8: This tool will help you calculate the difference in means between the A and B groups, which will help you determine if there is a significant difference between A and B



Group A	
Average:	24.53%
Respondent	Outcome
2	1
3	1
4	1
5	1
6	1

Group B	
Average:	23.05%
Respondent	Outcome
2	1
3	1
4	1
5	1
6	1

RESULT	
Statistically significant at conventional levels if p-value < 0.05	
p-value:	**0.503**

Meaning of outcome:	
0	Insert label meaning
1	Insert label meaning

Calculate cost-effectiveness gains generated by the innovation: If your organization routinely tracks program costs, you can estimate how much additional impact each dollar buys under the new model compared to the status quo. This helps determine whether to drop, adopt, or refine the tweak. (See useful costing guidance from J-PAL 🔗, the What Works Hub for Global Education 🔗, and IDinsight 🔗.)

For example, suppose the status quo model of your program – including only light phone-based teacher mentoring – costs USD 15 per child, while your tweaked program – with in-person intensive mentoring – costs USD 18 per child. If the tweak produces a 0.15 standard deviation improvement in learning relative to the status quo, you can calculate the gain in impact per dollar spent (effect size ÷ cost). In this example, the tweak generates 5 standard deviations of additional learning per USD 100 spent (0.15 SD difference ÷ USD 3 difference x 100), helping you judge whether the higher impact justifies the higher cost. If both versions achieve similar learning outcomes, the lower-cost version is more cost-effective; if the tweak achieves greater impact at a proportionate cost, it represents an efficiency gain worth adopting.

Make decisions based on evidence: Review results together with decision making implementation staff, and together agree on findings that will inform program decisions and the next test. One helpful way to inform decisions is to measure cost-effectiveness (above).

Plan your next test: Based on what you learned from your first test, design your next A/B test. After each test, the version that is most impactful or cost-effective is typically adopted as the new status quo for subsequent tests. This iterative process allows for continuous optimization. We discuss the process for ongoing testing in Phase 4.

Document the experience: After completing the A/B test, we recommend that the team reflect on what they've learned from the A/B test and how findings inform future decisions. We recommend a concise writeup (slides or narrative format). We share Tool 9 🔗 that can be used as a guide for structuring your presentation and decision summary. We also share an example 🔗 of a results presentation using the tool. These include:

- **A/B test question and design:** Clear explanation of your question or problem statement and randomization approach including explaining your status quo (A) and tweak (B) versions
- **Implementation and data summary:** Note the timeline, unit of randomization (e.g., students, classrooms, sites), and number of units in each group. Include response rates by group (e.g., share of the sample successfully surveyed or assessed at endline) and note whether response rates were balanced across A and B.
- **Results:** Present key findings on impact differences between A and B using your golden indicator data, including effect sizes and statistical significance.
- **Implementation fidelity**: If you measured fidelity, summarize any key fidelity statistics showing whether implementation occurred as planned.
- **Takeaways:**
  - **Insights**: Summarize what the team learned about the tweak's effectiveness, implementation challenges, and implications for practice. Also include cost-effectiveness results if possible.
  - **Recommendation**: Based on the evidence, the analytical staff recommend a way forward based on three options: *drop*, *adopt*, or *refine & retest*.
  - **Program management decision and next steps**: Record the actual decision taken by program leadership. Describe plans to action the decision.
- **Next A/B test:** Outline ideas for next A/B tests and proposed timelines.

This documentation ensures results are actionable, decisions are traceable, institutional learning is preserved, and it builds your organizational knowledge about what works in your program.

Phase 3 tips

Do	Don't
🟢 **Start** with a clear, simple question	🔴 **Test too many variations** in your first test
🟢 **Document** your randomization process carefully	🔴 **Get discouraged** if your first test shows no difference
🟢 **Monitor implementation fidelity** throughout the test	🔴 **Forget to document** what you learned for next time
🟢 **Involve program staff** in interpreting results and decision making	🔴 **Allow contamination** between treatment groups
🟢 **Source questions** from frontline implementers	🔴 **Just run one test** and fail to maintain the momentum

Case study in starting simple and moving onto a more advanced test: Save the Children Bangladesh

Save the Children Bangladesh 🔗 operates Catch Up Clubs 🔗, a foundational literacy and numeracy intervention addressing learning gaps. The program combines TaRL methodology with child protection services and family financial support, targeting marginalized children at risk of dropping out of school.

Recognizing that rapid iteration would maximize impact, Save the Children sought A/B testing support to generate quick, actionable insights. As shown in Table 3, their first test examined whether voice text message reminders to parents would improve children's learning outcomes. Results showed no impact—but this finding proved incredibly valuable to the team. Staff learned to design randomized experiments, implement protocols, and analyze data while fostering a cultural shift toward testing assumptions rather than relying solely on intuition. Building from this experience, Save the Children designed a more sophisticated test: does the sequencing of whether literacy versus numeracy instruction comes first matter? If sequencing doesn't affect outcomes, the program could allow flexible delivery, reducing logistical complexity.

Table 3: Save the Children Bangladesh conducted a first proof of concept test and progressed to a second more ambitious test

First test	Second test
Research question: Do text message voice note reminders to parents improve children's learning outcomes in Catch Up Clubs? **Version A:** Standard Catch Up Club implementation **Version B:** Standard implementation plus nine text message voice note reminders to parents over several months **Primary outcome:** Numeracy learning outcomes **Result**: Voice reminders had no effect on numeracy	**Research question:** Does the sequence of whether literacy or numeracy instruction come first in Catch up Clubs matter for learning outcomes? **Version A:** Children must complete literacy club before accessing numeracy club **Version B:** Children can access numeracy clubs before literacy clubs **Primary outcome:** Numeracy learning outcomes **Result**: Available end of 2025

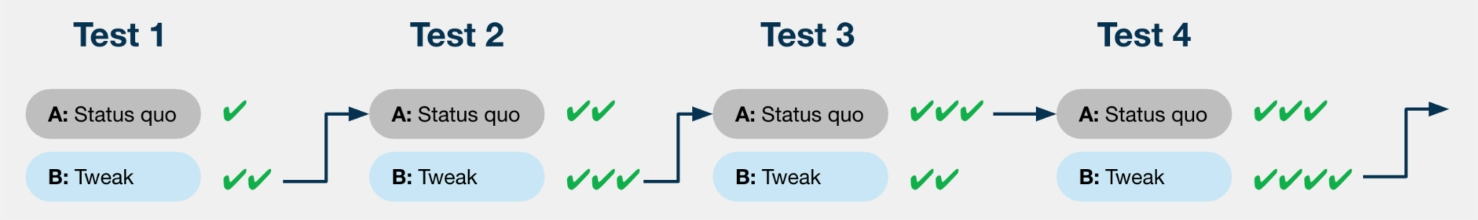
Save the Children's experience demonstrates that starting simple builds organizational confidence and an underlying learning infrastructure before tackling more complex program questions. Both tests addressed real implementation challenges with timelines aligned to program needs, showing how organizations can embed rapid, rigorous optimization into routine operations.

Phase 4: Ongoing testing

Goal

This phase is at the heart of iterative A/B testing. A/B testing is not a single evaluation or event, but an ongoing iterative process. In this section, we walk you through the process of developing more complex, ambitious A/B testing questions, as well as maintaining the momentum of A/B testing and carrying out repeated tests. Figure 6 illustrates how tests can cascade, with one test building on the next.

Figure 6: If version B is more effective, it can become the standard operating model in the next test



Plan | Implement | Reflect

Choose more complex, program-relevant A/B testing questions. The key characteristics of the first A/B test were that the test was simple, accessible, and doable. The goal for the second and subsequent tests is to dive deeper into **effectiveness-enhancing** and **cost-reducing tests** that are core to optimizing programs. Not every test results in an improvement and it's especially important in earlier phases to move quickly towards action rather than dwelling on the question. However in this phase and during the ongoing iterative A/B testing cycle, it is recommended to think more deeply and carefully about the questions being asked, in order to capitalize on the rapid testing muscle your organization now has.

As mentioned earlier in the toolkit, cost-reducing tests ensure scalability and sustainability, while effectiveness-enhancing tests maximize program impact. Both types of tests are essential for program optimization, because they both contribute to improved cost-effectiveness. While we do recommend cycling between both types of test, if a program is starting at an already high effectiveness bar, it might be preferable to prioritize cost-reducing tests. Similarly, if a program is very cheap to run but less effective, it would be worth prioritizing impact-enhancing tests.

Designing cost-reducing tests

This approach is similar to the game "jenga" where elements are carefully removed and made "lean," while maintaining the overall structure or in the case of A/B testing, maintaining impact. Cost-reducing tests typically remove or simplify program components, in order to reduce the cost of delivering a program. The goal is to identify costs or program elements that the program team believes can potentially be reduced or simplified without significantly reducing program impact. Identifying significant program cost drivers is a good place to begin: the most significant reductions naturally stem from the largest cost drivers; for example, if a large proportion of program costs are staff costs, considering staff-related questions will be important. If a large proportion of program costs are related to material production, questions around how to reduce this will be key, and so on. Costs can include those directly related to program implementation, like reducing transport or staffing, or those indirectly related, like adjusting staff allocation to improve efficiency.

Examples of cost-reducing test questions

- Does virtual training work as well as in-person training?
- Can weekly program sessions be delivered bi-weekly instead of weekly?
- Can we swap a complicated scheduling system for an "on-demand" program service?

Designing impact-enhancing tests

This approach is similar to building with Lego, where elements are carefully added to increase the height or "impact" of a program. Impact-enhancing tests typically add a program component, in order to enhance the effectiveness of a program. The goal is to identify additional program components that could potentially improve program impact, at no or low marginal cost.

Examples of impact-enhancing test questions

- Does including more structure in coaching checklists improve implementation fidelity?
- Does more frequent small group work in class improve learning?
- Does adding caregiver engagement improve learning outcomes?

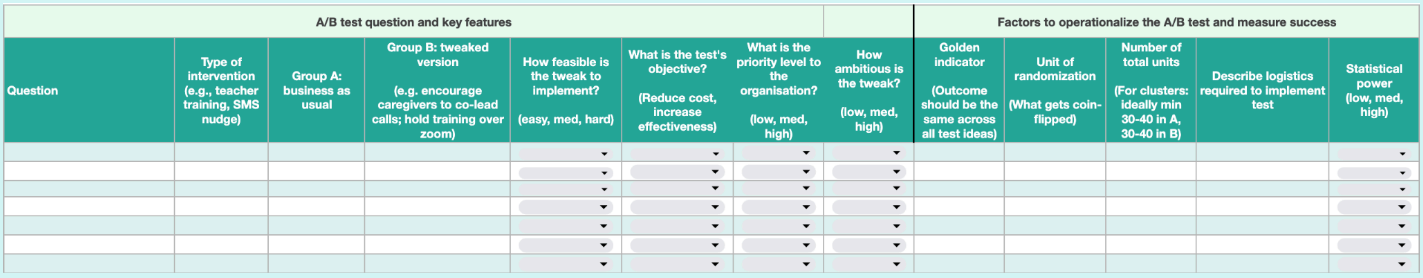
✔ Tip on sourcing questions for ongoing learning agendas

When generating A/B testing question options, a recommended primary source is frontline implementers, who know implementation efficiency margins best and have lots of ideas about ways to improve the program, or program features that are frictions to scale. Sourcing questions directly from frontline implementers can help populate strong A/B test question learning agendas. Engaging implementers in the test question development also leads to stronger take up of A/B test results.

This A/B testing question brainstorming Tool 10 🔗 will allow you to weigh important considerations for deciding which second and subsequent A/B tests to run, such as:

1. **Feasibility**: How easy or hard is the tweak to implement?
2. **Objective:** Effectiveness-enhancing or cost-reducing. Is version B going to improve impact or reduce costs? The strongest tests fit into one of these categories.
3. **Priority**: How much of a priority is answering this question to the organization?
4. **Ambitious yet achievable**: Is the question sufficiently "big" to help the organization make transformational changes? Version B should represent a meaningful change that matters to decision makers but not be so large that it requires extensive resources or months of piloting before testing. If a change is minor, low-risk, and likely to generate only marginal improvements, consider simply adopting it rather than testing. (Small tweaks also often require impractically large sample sizes to detect effects.) The best questions are typically "medium" in scope: substantial enough to be persuasive but realistic enough to implement with A/B testing's rapid cadence.

Tool 10: This tool helps weigh the pros and cons of different testing options for your subsequent tests



A/B test question and key features								Factors to operationalize the A/B test and measure success					
Question	Type of intervention (e.g., teacher training, SMS nudge)	Group A: business as usual	Group B: tweaked version (e.g. encourage caregivers to co-lead calls; hold training over zoom)	How feasible is the tweak to implement? (easy, med, hard)	What is the test's objective? (Reduce cost, increase effectiveness)	What is the priority level to the organisation? (low, med, high)	How ambitious is the tweak? (low, med, high)	Golden indicator (Outcome should be the same across all test ideas)	Unit of randomization (What gets coin-flipped)	Number of total units (For clusters: ideally min 30-40 in A, 30-40 in B)	Describe logistics required to implement test	Statistical power (low, med, high)	

As with the Phase 3 tool, the Phase 4+ tool also considers a range of technical questions for operationalizing the A/B test like the golden indicator, unit of randomization, total number of units, statistical power and any logistical details that would affect decision making. As your organization develops an A/B testing practice, you will develop your own format for noting questions and prioritizing.

Another A/B testing question trait to keep in mind as your question prioritization process becomes more sophisticated is to consider how much uncertainty is there at the organization about the effectiveness of the tweak. If prior beliefs are strong (i.e., decision makers are sure Version B will or won't work), changing them will be hard and require a lot of testing. If prior beliefs are more weakly held (i.e., decision makers do not have strong expectations about the impact of a tweak), the test will provide useful new information and the team might be more likely to accept evidence and take up the evidence. It can often be preferable to choose tests where decision makers have more uncertain prior beliefs.

In summary, a good A/B testing question is:

✔ **Feasible to implement**: Consider cost and feasibility of version B.

✔ **High priority**: Testing out changes in version B should be important to decision makers.

✔ **Effectiveness-enhancing or cost-reducing**: Tests should fall into one of these categories.

✔ **Sufficiently ambitious**: A/B testing questions should answer pressing implementation questions.

✔ **Implementer-driven**: Those closest to the programming often have the best ideas about how to improve it.

✔ **Answering a question with an unknown answer**: Decision-making teams shouldn't have strong priors about the outcome.

Plan **Implement** Reflect

Implementation for your second and subsequent tests is identical to your first test, so follow the key steps discussed in Phase 3, which in summary are:

1. Conduct randomization
2. Implement with fidelity
3. Collect outcome data

A/B testing in government implementation

A/B testing can be valuable in government partnerships because government implementation contexts are where sustainable scale happens, making them ideal environments for optimization. Government scale is also where there are often substantial implementation barriers – A/B testing is well suited to help tackle these barriers through further optimization. However, government partnerships also present unique challenges, including less implementation control, more stakeholders, and sometimes lower staff bandwidth. While we don't necessarily recommend starting your A/B testing journey in a government context if you have alternatives, we encourage organizations to conduct tests in government settings once they've built basic testing capacity.

Conditions that enable A/B testing in government contexts

✔ **Strong buy-in:** Government officials who understand the value of testing, are willing to randomize, and are committed to using results to inform decisions about program adoption or continuation.

✔ **Sufficient implementation oversight:** Your organization has enough influence over program delivery to help ensure both versions A and B are implemented with fidelity according to the randomization protocol, even when working within government systems.

✔ **Government staff capacity to support implementation:** Government personnel (teachers, health workers, extension agents, etc.) have the capacity and willingness to deliver different program versions as assigned and can maintain implementation fidelity throughout the test period.

✔ **Political feasibility of randomization:** Government stakeholders are comfortable with the randomization approach. This is often possible because all beneficiaries receive programming (just different optimized versions) rather than having a control group that does not receive the program.

✔ **Data systems infrastructure:** Government systems can support the high-frequency data collection needed for A/B testing, or your organization can supplement government data collection without creating parallel systems that aren't sustainable.

✔ **Alignment with government priorities:** The A/B test questions address genuine government concerns about cost, feasibility, or effectiveness, making results directly relevant to their decision-making needs.

Plan Implement **Reflect**

Reflection for your second and subsequent tests is also very similar to your first test, so you should follow the key steps discussed in Phase 3. You can use the results template Tool 9 🔗shared in Phase 3 to document.

1. Analyze results, document, and share learnings
2. Calculate cost-effectiveness gains
3. Make decisions based on evidence
4. Plan your next test
5. Document the experience

In Step 3 above (decision making), as you run more tests, you'll develop a clearer sense of which *p*-value thresholds feel appropriate for different situations. We introduced p-values in Phase 3; in Phase 4, as you reflect more deeply on them, consider how their interpretation might differ between tests aimed at improving effectiveness and those focused on reducing costs.

For **effectiveness-enhancing tests**, while academic research usually looks for a *p*-value less than 0.05 to be confident that differences between A and B groups are significant, when A/B testing your team might want to use 0.10 or an even higher *p*-value if the costs to adopt a tweak are low or low-risk and your team is more comfortable with decision making with less certainty.

For example, if your team is testing whether sending reminder texts two days before tutoring appointments reduces no-shows (compared with sending one day before), the tweak costs essentially nothing and is easily reversible. If for example you get $p = 0.12$ and a 3 percentage point drop in no-shows, you might decide to adopt the two-day reminder even if a bit above statistical significance thresholds, since the downside is negligible and there's a reasonable signal it might help.

For **cost-reducing tests**, the goal is often to confirm that a lower-cost version performs *no worse* than the standard model. In this case, a larger p-value can be desirable, suggesting no statistical difference between groups. For instance, suppose you test whether virtual training is as effective as the costly status quo of in-person training. You find a p-value of $p = 0.40$ and only a 3 percent lower learning level for the virtual group. Assuming your test was adequately powered, this indicates no statistically significant difference in learning, and you could confidently adopt the virtual training model, achieving the same impact at lower cost. To be sure that a lack of statistical difference in effects with the cheaper tweak is a true equivalence rather than being underpowered, you could repeat the test and also pool results across rounds.

Phase 4 tips

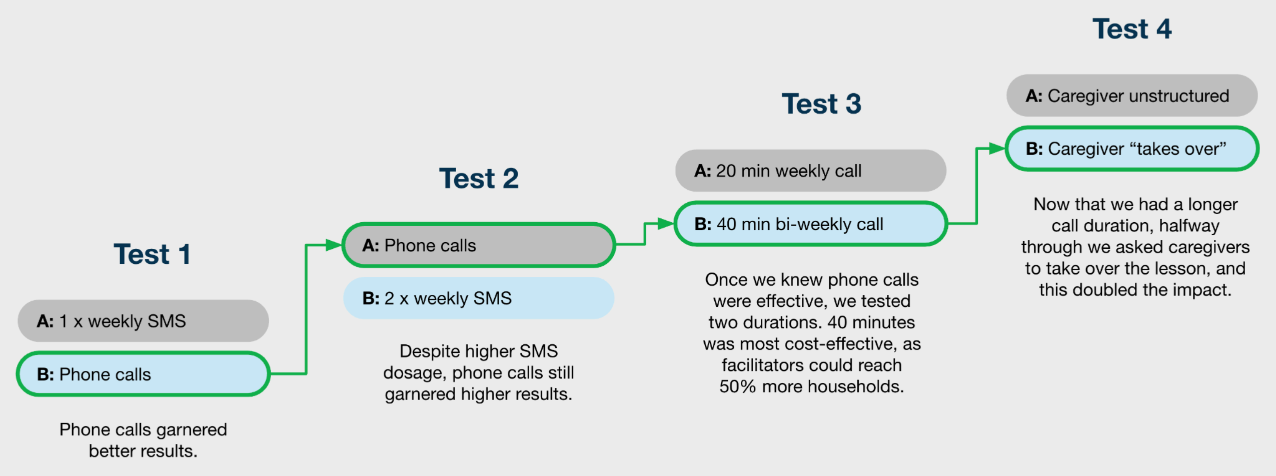
Do	Don't
🟢 **Apply all the dos and don'ts from Phase 3** in your second and subsequent tests	🔴 **Test something** that you already have high degree of certainty about
🟢 **Use the Phase 4 brainstorming tool** to help you refine the most relevant next test	🔴 **Test something** that is not operationally relevant
🟢 **Plan** for a series of tests, not just one	🔴 **Test something** that will not support your scaling (i.e., isn't cost-reducing or effectiveness-enhancing)
🟢 **Aim for future tests** to build on prior test, facilitating cumulative learning	🔴 **Carry out an A/B/C test** unless you do not have the option to carry out a series of A/B tests
🟢 **Carry out A/B tests with government partners** if they are sufficiently bought in and your organization can adequately support the test.	

Case study: a dozen A/B tests boost cost-effectiveness for tutoring

Since 2020, the ConnectEd 🔗 program has embedded rigorous, rapid, and regular A/B testing into every school term to optimize its phone-based math tutoring for scale and cost-effectiveness. The program delivers targeted instruction to primary school children via weekly tutorial phone calls and text messages—building foundational skills in addition, subtraction, multiplication, and division. Earlier randomized controlled trials across six countries established the model's effectiveness (Angrist et al., 2022 🔗; Angrist et al.; 2023 🔗). The next frontier was making it cheaper and even more impactful as it scaled, through iterative A/B testing.

Over 12 successive A/B tests conducted across school terms, the program reduced costs substantially while improving impact (Angrist et al., 2025 🔗), avoiding the typical "voltage drop" that programs often face as they scale (List 2023 🔗). Figure 6 shows four of these iterative tests. In our RCT in Botswana, we had tested a combination of phone calls and math problems delivered by text message. We found that the phone calls along with text messages had an impact (0.12 standard deviations) but the text messages alone had no impact (Angrist et al., 2022 🔗). For our first A/B test, we repeated this test (calls and text messages against text messages alone) as a baseline. In Test 2, we tried to double the dosage of text messages; since texts are so low-cost, we didn't want to give up on them, so we tested a higher dose of text messages. Results showed far greater impact still with the calls relative to text messages, reinforcing the importance of the phone calls for the model's effectiveness.

Figure 7: In ConnectEd, we conducted related, iterative A/B tests

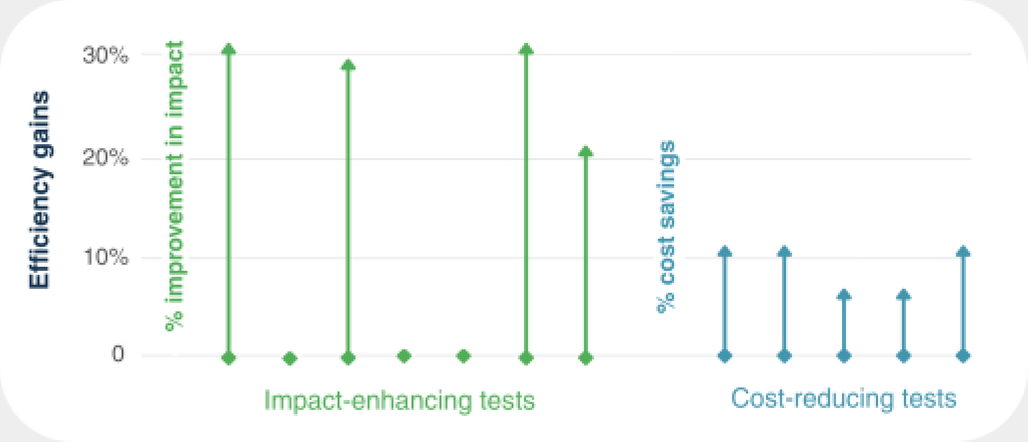

Test 1
A: 1 x weekly SMS
B: Phone calls
Phone calls garnered better results.
Test 2
A: Phone calls
B: 2 x weekly SMS
Despite higher SMS dosage, phone calls still garnered higher results.
Test 3
A: 20 min weekly call
B: 40 min bi-weekly call
Once we knew phone calls were effective, we tested two durations. 40 minutes was most cost-effective, as facilitators could reach 50% more households.
Test 4
A: Caregiver unstructured
B: Caregiver "takes over"
Now that we had a longer call duration, halfway through we asked caregivers to take over the lesson, and this doubled the impact.

For Test 3, we attempted a cost-reducing test. We explored whether shifting tutoring calls from weekly 20-minute sessions (version A) to biweekly 40-minute ones (version B) —keeping total instructional time constant—could preserve learning while saving time and cost scheduling in between calls. Lengthening calls maximized instructional time once tutors got a hold of households and minimized the time needed to schedule in between sessions. Results showed that there was no difference in outcomes, but significant reductions in staff time and coordination costs. Concretely, this resulted in a 11 percent reduction in cost at the same impact (Angrist et al., 2025 🔗). We repeated the test to be sure and confirmed the result.

The next test was an effectiveness-enhancing test. After implementers observed that tutoring calls went more smoothly when caregivers were involved, we tested this hypothesis. In version B, caregivers were encouraged to lead part of the tutoring call. This was also made possible by having longer 40-minute calls, building on prior tests. The result was striking: when caregivers took over calls part-way through, learning outcomes more than doubled at nearly no marginal cost. This tweak generated learning gains that were up to 30 percent more efficient, and yielded up to 66 standard deviations gained in learning per $100 spent. This breakthrough opened a new line of subsequent tests focused on maximizing caregiver engagement.

Figure 8 illustrates all 12 tests in this case study example and how iterative tests can help determine the most efficient and cost-effective model of a program. Each dot in the figure represents a test and efficiency gains as a result of the test. For effectiveness-enhancing tests, efficiency gains are captured by the gain in impact relative to the prior impact. For cost-reducing tests, efficiency gains are captured by reductions in cost relative to the prior cost. We found that all cost-reducing tests produced efficiency gains; effectiveness-enhancing gains were effective in only a subset of cases, but when they worked, they generated large efficiency gains.

Figure 8: Both cost-reducing tests and effectiveness-enhancing tests generated efficiency gains



While cost-reducing tests produced consistent efficiency gains, they remain rare in the social sector. To achieve scalability, removing program components while preserving impact is essential to develop “minimum viable products” (MVPs) that are easier for organizations to deliver at low cost and more likely to be adopted by governments at scale. Moreover, while a common assumption is that effectiveness is often reduced as programs are rolled out at ever larger scale, we find that effectiveness can be increased over time, as long as organizations continue to iterate and improve, mitigating and even reversing “voltage drops.” These examples illustrate how iterative A/B testing can both reduce costs and enhance effectiveness, and how various tests can build on each other, improving cost-effectiveness cumulatively over time. It can take multiple tries to identify a breakthrough innovation, so ongoing iteration is key. These examples are also a powerful reminder of the value of the rigor A/B testing brings to scaling decisions—and why we’ve made it core to how we work.

Organizational readiness

Before diving into the process, organizations should consider several criteria to assess their readiness for A/B testing. These are related to the organization's learning orientation, commitment, staffing, and existing M&E system. You can take our quiz 🔗 to assess your readiness.

Organizational attributes

Learning culture
A learning culture means the organization demonstrates a genuine commitment to making program changes based on data rather than tradition or intuition. Leaders and staff view "failures" and null results as valuable learning opportunities. The organization practices transparent communication about results and encourages implementing staff to suggest improvements based on their observations. This culture extends beyond the M&E team to program implementers who ultimately need to embrace and act on evidence-based recommendations.

Willingness to flip a coin
Staff should be willing to follow the randomization protocol and implement according to the random assignment. This means accepting that a "coin flip" determines whether a unit receives the "status quo" (Version A) or the "tweak" (Version B), even when staff have personal preferences for one version. Implementation fidelity is critical—staff should implement exactly as assigned and understand the importance of preventing contamination between treatment groups. Staff leading the A/B testing should be prepared to explain the randomization rationale to beneficiaries or external stakeholders, emphasizing that both versions are expected to be beneficial and no one is being denied the program.

Long-term commitment
A/B testing is a part of a learning system, not a one-off research activity. Organizations should plan to carry out multiple tests (our recommendation is at least five to ten rounds) to optimize a program and identify key results. This requires dedicating budget and staff time over multiple program cycles and persisting even when tests may show no difference—often it takes several tries before finding an efficiency improvement. Successful organizations view optimization as an ongoing process and build institutional memory by documenting tests and learnings to preserve knowledge.

Tolerance for making decisions under uncertainty
Organizations conducting A/B testing will make the most of the process if they are comfortable making decisions with lower levels of statistical certainty than traditional research. If an organization finds 70 percent certainty that a lower-cost approach is equally effective, they should be willing to switch to that model rather than requiring 95 percent+ confidence. This involves balancing rigor with

the need for timely decisions, assessing the potential risks of making changes based on preliminary evidence, maintaining flexibility to adjust program plans as new evidence emerges, and sometimes implementing approaches that differ from sector norms when evidence supports it.

Internal data-savvy M&E team, closely coordinated with program teams
Having committed M&E staff, aligned with organizational leadership, is crucial. We recommend at least one key person dedicated to A/B testing (this can be a program team member or M&E lead) who drives the process. This person should have or develop the ability to conduct basic statistical analyses and present findings in clear, actionable formats. Management should champion A/B testing and support implementation of findings, and program and M&E teams should work together closely throughout the process. Organizations should establish regular processes for reviewing results and incorporating learnings into operations.

M&E system attributes

Golden indicators to drive decision making
The outcomes of interest for A/B testing should be indicators your leadership uses to make programmatic decisions. These should generally be middle- to higher-level outcomes in your theory of change—they are downstream enough to reflect meaningful change but not so far downstream that they take years to manifest. Ideally, your organization should collect data on these indicators at least termly (every 2-4 months), although collecting several times a year might also be feasible when an organization is just getting started with A/B testing. If a true "golden indicator" isn't immediately feasible, organizations can identify "bronze" or "silver" indicators that serve as the next-best proxy for program success.

Large-scale data collection
Your tests should include a large enough number of units (e.g., students, classrooms, schools) to generate statistically valid results. While we would carry out calculations to determine your specific needs, as a rule of thumb, we recommend at least 60-80 clusters (classrooms, schools) if randomizing at the group level, or at least 1,200 students if randomizing at the individual level. This sample size requirement applies both during data flow (Phase 2) when establishing your measurement systems and during actual A/B testing (Phase 3). Having adequate scale is critical for detecting meaningful differences between program variations, especially when differences might be subtle but still important for cost-effectiveness. Organizations with smaller programs may need to collect data across multiple program cycles to achieve the necessary sample size.

High-frequency data collection
Data should ideally be collected at least termly (every 2-4 months) to be useful for A/B testing and decision making. This allows for rapid iteration, potentially conducting tests every school term or program cycle. High-frequency data collection creates a feedback loop that enables organizations to quickly learn from results and apply those learnings to the next iteration. While traditional evaluations might measure outcomes over years, A/B testing requires regular measurement that can show incremental progress and allow for real-time course correction. Organizations need measurement systems that can efficiently collect, clean, and analyze data within timeframes that align with program implementation cycles.

Frequently asked questions

Questions related to A/B testing generally

1 How is A/B testing different from a randomized controlled trial (RCT)?

While both use randomization, A/B testing is designed for rapid, iterative program optimization with results in weeks or months, not years. RCTs typically ask "does the program work?" by comparing treatment to a no-program control group. A/B tests ask "how can the program work more effectively, cheaply, and scalably?" by comparing different optimized versions of the same program. A/B testing ensures all participants receive programming (just different variations), making it more feasible for programs already operating at scale. A/B tests are also integrated into regular M&E systems for ongoing learning, while RCTs are typically one-off high-stakes studies.

2 Is A/B testing just for the tech sector?

While A/B testing originated in tech, it's increasingly recognized as a powerful tool for social sector organizations. The principles of rapid, rigorous testing apply to any organization implementing programs at scale and seeking to optimize cost-effectiveness. Education, health, agriculture, and other development sectors can all benefit from iterative A/B testing. The key difference is adapting the methodology to your context—using indicators that matter for social impact, not just clicks or conversions.

3 What scale does my program need to be to conduct A/B testing?

We recommend that organizations have at least 60-80 clusters (such as classrooms or schools) if randomizing at the group level, or at least 1,200 individuals if randomizing at the individual level. This scale is necessary to detect meaningful statistical differences. Organizations with smaller programs may need to collect data across multiple cycles to achieve necessary sample sizes.

4 What if I don't have a data system set up yet?

You don't need a perfect data system to get started. Phase 1 (Pilot Tweak) helps you identify gaps in your current system, while Phase 2 (Data Flow) focuses specifically on building the data infrastructure you need. We work with you to strengthen existing systems rather than building entirely new ones. Most organizations already collect some relevant data—the key is making it more frequent, systematic, and decision-focused.

5 Can we start A/B testing if our data systems aren't very strong?

It just might take your organization longer to get started if you need to work on system strengthening. Organizations typically fall into one of three categories: (1) Data already flowing—you can move quickly through Phase 2; (2) Need a small push—you collect some data but need to expand scale, increase frequency, or improve quality; (3) Need major system development—you require substantial overhaul of M&E systems before A/B testing is feasible.

6 How long does it take to get results from an A/B test?

Results typically come in weeks to months, depending on your program cycle and data collection frequency. Most tests produce actionable insights within one program cycle (e.g., one school term, one implementation round). This rapid turnaround allows you to make evidence-based decisions in real-time rather than waiting years for evaluation findings.

7 Can I just do one A/B test and then decide later if I want to do more?

Our experience is that A/B testing works best as an ongoing system for learning rather than a one-off activity. The real value comes from building your organization's capacity to conduct regular tests and continuously improve your programs. A/B testing is designed to be iterative—each test builds on the last, with the winning version becoming the new status quo for subsequent tests.

8 Can A/B testing help me identify whether my program works in the first place?

A/B testing is designed to optimize programs that have already demonstrated proof of concept. If you haven't yet established that your program produces impact, you'll want to first conduct a traditional evaluation or RCT comparing your program to a no-program control group. Once you have evidence of impact, A/B testing helps you make the program more cost-effective, scalable, and impactful as you grow.

9 How can I build buy-in for A/B testing within my organization?

Building buy-in requires demonstrating that randomization leads to better decisions than intuition or anecdotal evidence alone. Start small with Phase 1's pilot tweak to show quick wins and build momentum. Emphasize that A/B testing is easier to integrate than traditional RCTs since everyone receives programming—there's no pure control group being denied services. Building this "learning muscle" takes time, but organizations that commit to the process consistently find the evidence compelling once they see results from their first few tests.

10 What's the difference between cost-reducing and effectiveness-enhancing tests?

Both types of tests are essential for program optimization, and organizations often cycle between them. **Cost-reducing tests** aim to reduce costs while maintaining program impact. They typically remove or simplify a program component—for example, reducing staff time or materials. These tests make programs more scalable and attractive for government adoption.

Effectiveness-enhancing tests aim to increase program impact at minimal additional cost by adding program components. These tests improve outcomes while maintaining cost-efficiency, since enhancements typically have low marginal cost.

11 What if our leadership is skeptical about making decisions based on a "coin flip"?

This is a common concern. Leadership may worry about fairness, implementation complexity, or whether randomization is necessary. We find it helpful to emphasize that: (1) all beneficiaries receive programming under A/B testing—just different optimized versions, (2) randomization allows you to rigorously determine which version works better, and (3) guessing wrong can be costly and affect many beneficiaries when programs scale. Building a "learning muscle" within your organization takes time, as does building leadership buy-in throughout the process.

12 What staff roles are essential for A/B testing?

We find three key staff roles are critical for successful A/B testing:

1. Senior leadership champion: A leader who advocates for A/B testing and commits the organization to using results from tests. They must be willing to randomize, stick to randomization protocols, devote resources to A/B testing, and most importantly, act on A/B test results. Without leadership commitment to evidence-based decision-making, tests won't translate into program improvements.

2. Implementation staff: Program implementers who are willing to randomize their work, embrace A/B testing ideas, and generate program tweak ideas based on their field observations. These staff members implement different versions with fidelity and help identify which questions are most operationally relevant to test.

3. M&E or data-savvy lead: Someone who can enhance your existing monitoring data system if needed, execute the randomization, manage data collection, run analyses, and interpret results. This person doesn't need to be a PhD-level researcher, but they need practical skills in data management and basic statistics, plus the ability to communicate results clearly to decision makers.

13 Can we do A/B testing in partnership with the government?

Yes! A/B testing is particularly valuable for government partnerships because it demonstrates cost-effectiveness and scalability—key considerations for government adoption. Many organizations implementing programs do so in partnership with government systems (such as regional education offices or health departments), and A/B testing works well in these contexts.

We encourage A/B testing with government partners, though we don't necessarily recommend starting in this context if you have alternatives—it's often easier to build your testing capacity first in a context where you have more implementation control. That said, government contexts are where scale happens, and testing in these environments ensures your findings are directly relevant to sustainable, scalable programming.

One major advantage: A/B testing is often more agreeable to governments than traditional RCTs because all beneficiaries receive a version of the program rather than having a control group that receives nothing. Tests that reduce costs while maintaining impact are especially persuasive for government scale-up decisions. We've successfully supported partners working with regional offices and government systems to run A/B tests.

14 I want to do a three-arm trial (A, B, C) for the first test. Is that okay?

For your first A/B test, you should keep it simple with just two variations (A and B). We find that simple A vs B sequential trials, conducted in rapid succession at high frequency, often outperform less frequent multi-armed trials for several reasons: (1) they have larger samples for each treatment, so one can be more certain about differences they produce, (2) they are easier and more feasible to implement and monitor fidelity on, and (3) they are simpler to interpret and thus to rapidly convert into decisions. Even after you've conducted multiple successful A/B tests, we generally recommend sticking with two-arm designs. The simplicity, clarity, and statistical power of A/B tests make them more practical for program optimization. Multi-arm trials should be reserved for rare circumstances where sequential testing is truly not feasible.

15 What is a "golden indicator" and how do I know if I have one?

Golden indicators are outcomes that decision makers view as consequential enough to make programmatic decisions. They must strike a balance—proximate enough to detect changes quickly (within weeks or months), yet meaningful enough to drive important decisions. These are typically middle-indicators in your program's theory of change. For example, in education programs, foundational literacy and numeracy outcomes often serve as golden indicators.

If you don't have a perfect golden indicator, you can use "silver" or "bronze" indicators as next-best proxies. Bronze indicators (like knowledge assessments) may be easier to measure frequently but further from your ultimate impact. In that case, you might want to use bronze indicators as the leading indicator for your A/B test but also track a silver indicator, one that is closer to the final outcome but may take longer to measure. Tool 3 the “golden indicator shortlisting tool” helps you identify which indicator(s) work best for your context.

16 Can we use more than one golden indicator?

We recommend that organizations choose one primary indicator that serves as the leading indicator for A/B testing over time. This should be an indicator that triggers programmatic changes if the indicator moves—your north star for decision-making. Having one clear golden indicator ensures your team has clarity about what success looks like.

Of course, your organization can and should collect several indicators of interest and relevance as part of your broader M&E system. The key is that one indicator drives your A/B testing decisions. Tool 3 the “golden indicator shortlisting tool” helps you identify your primary golden indicator while recognizing that other indicators remain valuable for your overall monitoring.

17 Should we prepare pre-analysis plans or pre-specification documents before each test?

Your design, hypotheses, and outcome(s) can be documented using tools like Tool 5 (A/B test design tool), Tool 6 or Tool 10 (the A/B testing question brainstorming tools). These tools encourage you to clearly state your research questions, specify your golden indicator, and document your approach before conducting the test. You'll also present these components in Tool 9 (the A/B testing results deck) once the test is complete. Because you're using one golden indicator and one clear measure of success, there is also less risk of "p-hacking" among indicators.

18 What happens if we can't implement version B with fidelity?

Implementation fidelity is critical, which is why we recommend starting with simple, realistic questions in Phase 1. We recommend selecting program variations that your team can implement with confidence. During testing, we help you monitor implementation to ensure both versions are being delivered as planned. If fidelity is a concern, it's better to choose a simpler test question or strengthen implementation systems before proceeding. Testing only works if you're actually implementing what you intend to test.

19 Should I hire an external data collection firm to collect data for A/B testing or do the data collection in-house?

We strongly recommend conducting data collection in-house rather than hiring external firms. A/B testing is designed to be an internal learning system that produces rapid, actionable insights for programmatic decision-making. Using external data collection firms introduces delays, increases costs, and creates dependencies that undermine the core purpose of A/B testing.

You'll be running tests regularly—potentially every program cycle or school term—which makes outsourcing data collection impractical and unsustainable. The costs would quickly become prohibitive, and waiting for external firms to mobilize, collect data, and return results would eliminate the "rapid" advantage that makes A/B testing valuable.

This upfront investment of building a strong internal M&E system pays dividends over time as you conduct repeated A/B tests. Your internal team will develop expertise, data collection becomes more efficient with each round, and you maintain full control over timing and quality. Phase 2 (Data Flow) of this toolkit is specifically designed to help you strengthen your internal data systems to support ongoing A/B testing.

In-house data collection also keeps A/B testing integrated with your program operations rather than treating it as a separate research activity. Your implementation staff understand the context, can troubleshoot issues in real-time, and can quickly iterate based on what they're learning. This integration is essential for building a true learning culture within your organization.

20 What if our A/B test shows no difference between variations?

Null results are a normal and valuable part of A/B testing—not every test will be transformative. This is why we emphasize a "fail fast" approach and recommend committing to multiple tests . When a test shows no difference, you've still learned something important: either the tweak didn't work as hypothesized or both versions are equally effective. For cost-reducing tests, finding no difference can be excellent news—it means you can adopt the cheaper version without sacrificing impact. Innovation requires iteration, and we find it often takes five to ten tries to identify a breakthrough. Documenting null results prevents your organization from repeatedly testing ineffective approaches.

21 Once an organization has run an A/B test and wants to do a subsequent test, should we re-randomize using a new cohort or continue testing within the same sample?

This issue is relatively uncommon for most implementers. Many organizations—especially those delivering programs like foundational literacy and numeracy—work with a new cohort of beneficiaries (e.g., new students) each implementation cycle, so you re-randomize with each new cohort. However, this question is more relevant for organizations doing individual-level randomization or working in EdTech contexts where the same individuals may participate across multiple rounds. In these cases, we recommend re-randomizing for each new test. It's fine to have a mix of individuals who were previously exposed to versions A and B in an earlier test —this doesn't compromise your new test as long as randomization is done properly for the current round. Each individual will have an equal chance of being in A or B each new randomisation, so any difference between them at the new endline will be attributable only to the latest test. However, previous tests can add more variation to the data, and it's nice to guarantee equal representation of previous test participants in your new AB test randomised groups, so it is advisable to stratify by previous participation in an A/B test, if possible.

Questions related to partnering with Youth Impact

1 How do I get started?

Read this guide! After you have read through this guide, you can reach out to Youth Impact by filling out our online form 🔗. We'll discuss your program, assess your readiness for A/B testing, and determine how we can best support your journey toward evidence-based program optimization.

2 What are the requirements for partnership?

We look for organizations that: implement programs at sufficient scale (minimum 60-80 clusters or 1,200 individuals); have commitment to evidence-based program improvement; can dedicate staff time to A/B testing implementation; have leadership buy-in for making program changes based on evidence; and are willing to invest in data systems improvements if needed.

3 How much time does A/B testing require?

If partnering with Youth Impact: Organizations should be available to work with us over several months (approximately 9 months for the initial phased process), with key staff potentially committing 2-4 hours per week during busier periods. The process is structured in phases to build capacity gradually, starting with small-scale pilots and expanding to full A/B testing over time.

If conducting A/B testing independently: Time requirements depend on your existing systems and capacity. Most organizations find that once systems are established, A/B testing integrates into regular program cycles without substantial additional time burden. Initial setup (identifying indicators, strengthening data systems) requires more upfront investment, but ongoing testing becomes part of routine M&E.

4 How much does it cost to partner with Youth Impact on A/B testing?

Youth Impact's A/B testing support is currently funded by partner donors. We provide technical assistance, tools, and in some cases resources to partner organizations (subgrants). However, organizations should budget for their own implementation costs, staff time, and any data system improvements needed to support A/B testing.

5 How do you support us through the partnership?

We provide tailored support based on your needs, including: regular consultation calls throughout your A/B testing journey; review of test designs and analysis plans; guidance on randomization and implementation fidelity; support with data analysis and interpretation; help integrating results into program decisions; tools for planning, randomization, and analysis (in this guide!); connections with other organizations implementing A/B testing; and learning sessions featuring case studies and best practices.

6 Can you help my organization develop new indicators for A/B testing?

Generally we lean towards leveraging indicators you already collect rather than developing new ones. This allows us to get started quickly and ensures the indicators are already integrated into your operations. We find that most organizations already have suitable indicators that can serve as "golden" or "bronze" indicators for A/B testing, and we're happy to help you identify these from your existing data. However, if you need to develop new indicators, we can support that process, though it may extend your timeline for Phase 2.

7 Will you keep my data confidential?

Yes, we take data privacy and confidentiality seriously. We have protocols in place to protect your organization's data and will only share results with your explicit permission. We recommend only sharing anonymized data with us. We can also sign a data sharing agreement if that would be helpful for your organization.

Youth Impact

www.youth-impact.org