*AI-generated image

A Practical Approach to Developing A/B Testing Systems for Digital-first Organizations

ipa Innovations for Poverty Action

March 2026

Authors

Andrés Parrado, Mariana Rodríguez, and Eduardo Vargas

Contributors

We thank our colleagues Amanda Beatty, Ben Tan, Daniela Sanchez, Elia Gandolfi, Noam Angrist, Rob On, Simon Rubio, Temina Madon, and Thomas Chupein for their careful review and valuable feedback throughout this work.

We also thank the Jacobs Foundation for their funding and support to the **Partnerships for Technology in Education (P4T-Ed)**, the What Works Hub for its commitment to actionable research and innovation, and our partners, Mentu Labs and TxC, whose collaboration, commitment to learning, and openness to testing shaped and strengthened this toolkit.

Editing and design support provided by the Innovations for Poverty Action (IPA) Communications team, including Ana Tamayo

Content

Executive Summary

Social impact organizations face growing pressure to learn and improve their programs with limited resources. A/B testing offers a rigorous and practical tool to test program adaptations aimed at improving cost-effectiveness. This brief provides a step-by-step guide for digital-first organizations seeking to embed A/B testing into their Monitoring, Evaluation, and Learning (MEL) system. Drawing on Innovations for Poverty Action's Right-Fit Evidence (RFE) Unit advisory experience and a growing body of evidence from the social sector, it introduces the Learning Roadmap for A/B Testing, a structured four-step process for developing the organizational capabilities, technological infrastructure, and learning culture needed to test, learn, and improve continuously.

KEY TAKEAWAYS

- A/B testing is a powerful method for generating rapid causal evidence, but it is not appropriate for every question. It is best suited for organizations that need to understand how to improve specific components of an intervention at the Refine, Adapt, or Scale stages of program development.
- Building an A/B testing system from scratch takes typically 6 to 9 months to establish the necessary processes and infrastructure. Once the system is operational, testing cycles can be completed within 2-3 months.
- The Learning Roadmap guides organizations through four core steps: aligning testing with learning needs through a theory of change; identifying and prioritizing what to test using a cost-effectiveness lens across reach, efficacy, and efficiency; preparing technological infrastructure and test design; and implementing tests with rigorous monitoring and structured analysis that can drive decision-making.
- Data infrastructure alone is not sufficient. Sustained adoption of experimentation depends on cross-team alignment between MEL, program, and technology functions, and a shared organizational commitment to acting on evidence.
- The ultimate goal is not to run experiments in isolation, but to build a sustainable learning system. Each test should inform the next, with documented findings feeding into program decisions, future test designs, and broader organizational strategy.

*AI-generated image

1. Introduction: What is A/B Testing?

Social impact organizations face growing pressure to learn faster, adapt smarter, and improve their programs with limited resources. In this context, the ability to test and learn quickly has become critical for improving programs with limited resources. A/B testing offers a rapid yet rigorous way for social sector organizations to learn what works and make evidence-based improvements.

> *At its core, **A/B testing is a rigorous method that compares variations of a program or product to determine which one generates a greater effect on a specific outcome.** In its simplest form, it tests the default version of an intervention (option A) against a variation of the default version (option B), randomly assigning participants to each and tracking performance on a pre-defined target metric (Angrist et al., 2024). When designed well, this approach is able to isolate the causal effect of the variation, generating rigorous evidence for decision-making. With some investment, organizations can leverage their existing monitoring systems to collect data on target metrics, enabling them to run regular testing cycles.*

A/B testing falls under the broader umbrella of randomized experiments, since it relies on random assignment across conditions to generate causal evidence. In this paper, however, we distinguish them more narrowly: we mention randomized controlled trials (RCTs) to refer to studies where an intervention is compared to a group that did not receive the intervention (pure control group) and is typically focused on long-term, final outcomes (e.g., income, learning gains). On the other hand, we use "A/B testing" to describe rapid experiments that usually rely on administrative or easily accessible monitoring and product engagement data. Both methodologies can be used for learning and evaluation purposes, but their strengths differ. A/B testing helps to answer questions such as "Which version works more effectively, efficiently, or scalably?" While RCTs are most often used to answer: "Does the program work?" In practice, A/B testing supports continuous, internal learning during implementation, while RCTs are typically used for summative, long-term impact evaluation. This practical distinction is important because A/B testing's potential for rapid, iterative learning is a major reason why organizations should seek to internalize this capacity.

While A/B testing has long been a core tool for product optimization in the technology sector, its application in the social sector has only more recently begun to gain traction. Emerging evidence shows that iterative A/B testing can be successfully embedded into Monitoring, Evaluation, and Learning (MEL) practices to generate causal evidence at implementation speed, particularly for questions related to cost-effectiveness, operational efficiency, and scalability

(Angrist et al., 2026). For instance, Angrist et al. (2025) found that A/B tests designed to improve the scalability of a phone-based tutoring program in Botswana produced efficiency gains in 58 percent of tested modifications (7 of 12), exceeding typical technology-sector discovery rates of 10-40 percent. Effectiveness-enhancing modifications like caregiver co-tutoring generated learning gains of up to 65 standard deviations (SD) per USD 100 spent. In addition, cost-reducing tests achieved up to 11 percent savings while maintaining learning outcomes. Similarly, Alvarez-Marinelli et al (2021) demonstrate how sequential randomized experiments can compound program effectiveness over time. Using feedback from each cohort to refine and test a remedial literacy intervention in Colombia, measured gains on literacy scores grew from 0.138 SD in the first cohort to 0.525 SD by the third, nearly a fourfold increase attributable to higher dosage and fine-tuning of pedagogical materials.

> *Despite this growing body of research,* ***the systematic use of A/B testing in the social sector remains limited.*** *Many organizations seek to develop experimentation capabilities, but it is often hard for them to integrate A/B testing into their measurement and evaluation practices. This guide aims to contribute to and complement emerging experimentation resources in the social sector like* ***Youth Impact's A/B Testing Toolkit (2025)****,* ***The Agency Fund's AI Evaluation in the Social Sector Playbook (2025)****, and* ***Evidential****, an experimentation engine developed by The Agency Fund and IDinsight. Specifically, it aims to help digital-first organizations bridge the gap between the demonstrated potential of A/B testing and its practical adoption as an internal learning tool.*

WHO IS THIS GUIDE FOR?

Every organization can invest in learning. That can mean trying new ideas, piloting variations of an intervention, user testing, and gradually building a more structured approach to understanding what works and why. Organizations do not need sophisticated technology to begin testing . They just need to be open to learning, willing to invest in iterative improvement, committed to evidence-based decision-making, and prepared to act on data. If you consider that your organization is mostly ready to embrace an experimentation and iterative learning culture, A/B testing may be an appropriate tool.[1]

That said, this guide is written for a specific audience, distinguishing organizations by delivery mode (digital vs. non-digital)[2] and data system maturity (nascent vs. mature). It will be most useful for organizations that either deliver their interventions through **digital channels (digital-first)** or already have relatively **mature data systems** (see dark gray cells in Figure 1). These conditions make it possible to track outcomes at the user level, run controlled comparisons, and interpret results with enough confidence to act on them. Much of the guidance will also be relevant for teams that are still strengthening their data infrastructure, as well as for organizations whose delivery is primarily in person but who already collect high-quality data in a systematic way. For organizations that are neither digital-first nor equipped with mature data systems, we strongly recommend starting this journey with Youth Impact's **Iterative A/B testing Toolkit**, which offers a phased and accessible pathway for building experimentation capacity.

Figure 1. Who is this guide for? [3]

	Nascent	*Level of maturity of data systems* →	**Mature**
Non-digital organizations 	Annual or less frequent monitoring data collected through manual processes (paper forms) and available months after program activities. No centralized system for storage or analysis.		Termly or frequently monitoring data collection on key indicators using mobile data collection tools. Near real-time monitoring of programs with integrated dashboards used for decision-making.
Digital-first Organizations 	Product generates user and usage data, but it is not systematically captured or stored. No analytics infrastructure or event tracking in place.		Automatic logging of monitoring data on key indicators (user interactions and outcome metrics). Real-time analytics pipeline connecting product data to visualization and analysis tools for decision-making.

[1] For an organizational readiness self-assessment, use Youth Impact's A/B testing readiness tool, specifically Section A Organizational culture & commitment.
[2] Non-digital organizations are considered those whose core intervention is delivered through in-person, analog, or field-based channels (schools, community meetings, home visits), even if digital tools are used for support.
[3] The spectrum for the level of maturity of data systems is designed in line with Youth Impact's A/B testing readiness tool, specifically Section B *M&E system capabilities.*

Before diving in, it is important to note that building an A/B testing system is not a quick win. For many organizations, moving from an initial idea to implementing a well-designed test takes time. In practice, the process to set-up the systems and process for testing from scratch takes **six to nine months**, as teams refine their theory of change (ToC), ensure their digital systems can support monitoring and testing, and put in place the workflows needed to deploy and analyze a test. Then, implementing the first test takes **two to three months.** These efforts also require sustained time and coordination across MEL, program, and technology teams, particularly if results are expected to meaningfully inform program or product decisions. This guide will return to these practical considerations in more detail in Section 3. For now, we encourage readers to start thinking early about the time, sequencing, and organizational commitment required to adopt A/B testing effectively.

> *The guide begins by discussing when this methodology is an appropriate method for learning, and outlines key organizational capabilities for the effective adoption of an A/B testing system. The next section introduces the Learning Roadmap for A/B testing, a structured approach developed by Innovations for Poverty Action's (IPA) RFE unit to help organizations design, prioritize, and implement A/B tests. This approach supports organizations in building the processes and technology infrastructure needed to experiment. The final section shares practical lessons that have emerged from the advisory engagements supporting partners in adopting A/B testing and refining their EdTech interventions. While the guidance applies broadly to organizations in any sector, the examples focus on education and education technology (EdTech)[4] interventions inspired by the Right-Fit Evidence Unit's (RFE) advisory engagements through the initiative* ***Partnerships for Technology in Education (P4T-Ed)****.*

[4] Based on Singh et al. (2025), we take a broad definition of EdTech to include all interventions that rely fundamentally on information and communication technology (ICT) tools, and that either directly affect student learning outcomes or serve as essential enablers of interventions that do so.

2. When is A/B Testing the Right-Fit Approach for Learning?

A/B testing is a powerful method, but it is not the right tool for every situation. Deciding when to do A/B testing requires reflecting on **the kind of learning questions** a program needs to answer **at its current stage of development,** and whether **the organization's current capabilities can support experimentation.**

WHEN TO DO A/B TESTING: STAGE-BASED LEARNING

To identify when to do A/B testing, IPA's Right-Fit Evidence Unit proposes an overarching framework that guides learning activities across programs' stages of development. The Stage-Based Learning framework helps organizations identify their learning needs based on program maturity.[5] Each stage –Ideate, Refine, Prove, Adapt, and Scale– is associated with particular types of learning needs, and those needs call for different methods that can generate insights that are credible, actionable, cost-effective, and are likely to generate transportable knowledge.

Figure 2. Stage-Based Learning

Ideate	Refine	Prove	Adapt	Scale
Problem is identified; potential intervention are **designed and tested with users.**	Interventions with early potential are **piloted**	Successfully piloted interventions are **tested for impact** on **final outcomes**	Proven **interventions are adapted** o new contexts and/or new implementers	Interventions with proven effectiveness are scaled
• Needs assessment • Evidence review • User testing/ prototyping	• Process Evaluation • **A/B Testing** • Monitoring	• Impact Evaluation	• Needs Assessment • Evidence review • **A/B Testing** • Monitoring • *Possibly* Impact Evaluation	• Ongoing monitoring (for implementation quality) • **A/B Testing**

In the early stages of an intervention, the focus of an organization should be on understanding the problem, rapidly testing potential solutions with users, and refining the design. In later stages, the priority shifts to demonstrating impact, adapting to new contexts, and ensuring consistent performance at scale. A/B testing is one of several learning methods, and it is particularly relevant for programs at the Refine, Adapt, and Scale stages of the framework,

[5] IPA Right-Fit Evidence Unit. (2024, October). Enabling stage-based learning: A funder's guide to maximize impact. https://poverty-action.org/sites/default/files/2024-10/Enabling-Stage-Based-Learning-Full-Guide.pdf

when organizations face questions about how to improve or adapt specific components of an intervention in a different context or at a greater scale.

Table 1. Learning Needs at the Refine and Adapt Stages

Stage	Stage identification	Learning focus	Key learning questions
Refine	The intervention's **initial design is complete** and its key components have been **tested** with a small **group of users.**	Iteratively test the intervention to assess **implementation quality** and **early signs of effectiveness.**	• Which delivery channels or engagement strategies are most effective at increasing take-up and/or reach among the target population? • Which components of the intervention are driving changes in early outcomes? • What elements can improve before the intervention is evaluated for impact and/or scaled?
Adapt	The intervention has **demonstrated positive impact** through rigorous impact evaluation and will be implemented in a **new context[6].**	Adapt the model ensuring **alignment** with challenges and characteristics in the **new context,** while aiming for efficiency.	• What adaptations are needed for the intervention to be effective and relevant in the new context? • How can the same or similar level of impact be achieved at lower cost?
Scale	The intervention has **demonstrated positive impact** and **has been adapted** to be implemented with quality at scale.	Ensure that the **implementation quality** is maintained when **reaching more participants.**	• Are there more efficient ways to deliver the program at scale maintaining the quality?

In the Refine stage, A/B testing can support learning needs by enabling organizations to systematically test variations[7] to the program that could address these questions. For example: An EdTech organization piloting an AI-enabled teaching assistant may want to test whether adding a classroom feedback feature leads to improved instructional quality. By running an A/B test comparing two groups–one with and one without the feedback feature–and tracking student performance, the organization can assess whether this addition meaningfully enhances outcomes.

In the Adapt stage, A/B testing can support learning goals by comparing different versions of the intervention or delivery model to assess which approach works best in a new context. For example, a proven student learning platform expanding to a new country might test different tones for its AI-enabled coach to determine which one is most effective at engaging students.

[6] This often involves delivering the intervention in new geographies, to new populations, at a later time, and/or through new partners.
[7] A recent thought piece by IDinsight highlights an important pitfall: the risk of launching A/B tests before giving enough time to work out the kinks of new program designs. To avoid this, it is essential to ensure that variations to be tested are grounded in evidence or user research, as discussed in the next section. Additionally, when testing major components of a program, it is good practice to pilot them first to confirm they work as intended. This helps ensure that resources are not spent testing a variation that is clearly ineffective or infeasible to implement.

In the Scale stage, A/B testing can support learning goals by testing adaptations to a delivery model to determine which approach is more cost-effective at scale. For instance, an EdTech program scaling nationwide might compare SMS reminders (costly) versus in-app notifications to see which drives higher student retention at lower cost.

The insights gained from A/B testing help organizations strengthen interventions based on rigorous evidence from real-world implementation before making larger investments in an impact evaluation or afterwards to keep adapting the intervention at scale. A/B testing can be a very useful methodology for refining and adapting programs, but its potential depends on organizations' ability to design credible tests, interpret results accurately and act on them.

> *This often limits or restricts its use, so our approach consists of developing an A/B testing system that builds the capabilities necessary to test and refine continuously.*

WHEN AN A/B TESTING SYSTEM IS A GOOD FIT FOR A DIGITAL-FIRST ORGANIZATION

Selecting methodologies that align with the organization's existing capabilities is crucial to balance the learning-to-cost ratio. This means focusing on the most useful data for the lowest possible cost. While A/B testing can offer powerful insights, it is a tool more appropriate for **organizations that have, or aim to build, the capabilities needed to design, implement, and learn from rigorous experiments.** These include both technical and operational requirements that ensure the tests are feasible and the results are reliable.
Section 3 provides a step-by-step process on how to develop these capabilities as organizations build and embed an A/B testing system within their MEL toolkit.

Drawing on experience developing A/B testing systems with organizations, we have found that developing or strengthening the following capabilities is especially important:

PROGRAM CONDITIONS

- **Clarity** on what the organization wants to learn with the A/B test (e.g. improving uptake or retention rates, increasing meaningful engagement, or reducing costs).
  - Quick rule-of-thumb: If you can state the decision you'll make under each possible result before launching, you have clear learning goals. If not, you likely need discovery research first (needs assessment, prototyping, user research).
- **Clear plans to implement the progra**m, as A/B testing can only be done while an intervention is running or a product is live.
  - If the product isn't live yet, do user tests or prototyping first.
- **A stable and sufficiently large user base**, to support rigorous comparisons.
  - If the sample size does not support statistical rigor, prioritize qualitative methods.

SYSTEMS AND TOOLS

- **Reliable data systems** to track relevant outcomes, such as user interaction with the platform, impact-related metrics, and experiment performance indicators.
- **Experimentation tools or software,** developed internally or integrated from a third party, that allow for random assignment and the management of multiple tests.

TECHNICAL CAPACITY

- **Technology capacity,** to build or set-up the infrastructure for testing , including data workflows, tool integration, and the delivery of test variations in live environments.
- **Analytical capacity,** a person in the team that knows both the program and the technological infrastructure and can allocate time to **design valid tests , interpret statistical results, and turn findings into actionable recommendations.**

Organizations do not need to master all these capabilities at once. Many can start with other low-cost learning methods to build capacity over time, using early experiences as opportunities to strengthen data practices and foster a culture of learning. These early steps help teams build capacities and confidence in using evidence for decision-making, laying the groundwork for more structured experimentation.

BOX 1. WHEN NOT TO RUN A/B EXPERIMENTS AND WHAT TO DO INSTEAD:

- When analytical tools and know-how to run experiments are not available, explore using other methods such as user research, prototyping, or simple analog tests, and start building analytical and data capabilities.
- When the user base is small or volatile (underpowered results won't be credible), do usability tests, interviews, focus groups or small pilots.
- When you want to measure program effects on long-term outcomes (e.g. changes in income), consider an RCT to assess impact on final outcomes.
- When you expect high interference/spillovers between users, consider cluster-level tests or use system-level methods.
- When outcomes are highly predictable or stakes are trivial, try the obvious change and track performance with strong monitoring.

The next section outlines a step-by-step approach for building an A/B testing system. The approach illustrates how key capabilities can be developed over time in a way that aligns with an organization's stage of readiness. While designed to support teams starting from scratch, the process can also be adapted by organizations that already have some capabilities in place.

3. How to Implement A/B Testing in Digital-first Organizations?

When an organization determines that A/B testing is the right tool to address its intervention's learning needs, a structured approach becomes essential to integrating experimentation into MEL processes and organizational culture. Successful adoption of A/B testing requires aligning learning goals with experimentation, establishing appropriate technological infrastructure, and designing processes that generate reliable and actionable insights. For organizations starting this journey, the challenge is not only how to run a test, but how to build an integrated experimentation system that supports decision-making. Importantly, organizations do not need to have all the required capabilities in place from the outset. This roadmap is designed to support the capability-building process, providing **sequential guidance** to develop these step by step. For most digital-first organizations, this process from beginning to end could take **six to twelve months** and it aims to build lasting capacity to conduct faster cycles of A/B testing into the future. Note that this is an indicative estimate and depends on several factors like the maturity of the data systems and how developed an organization's Theory of Change is. Ultimately, once the organization has built the technology and procedural capacity to run A/B tests, **subsequent testing cycles can be conducted as fast as two to three months.**

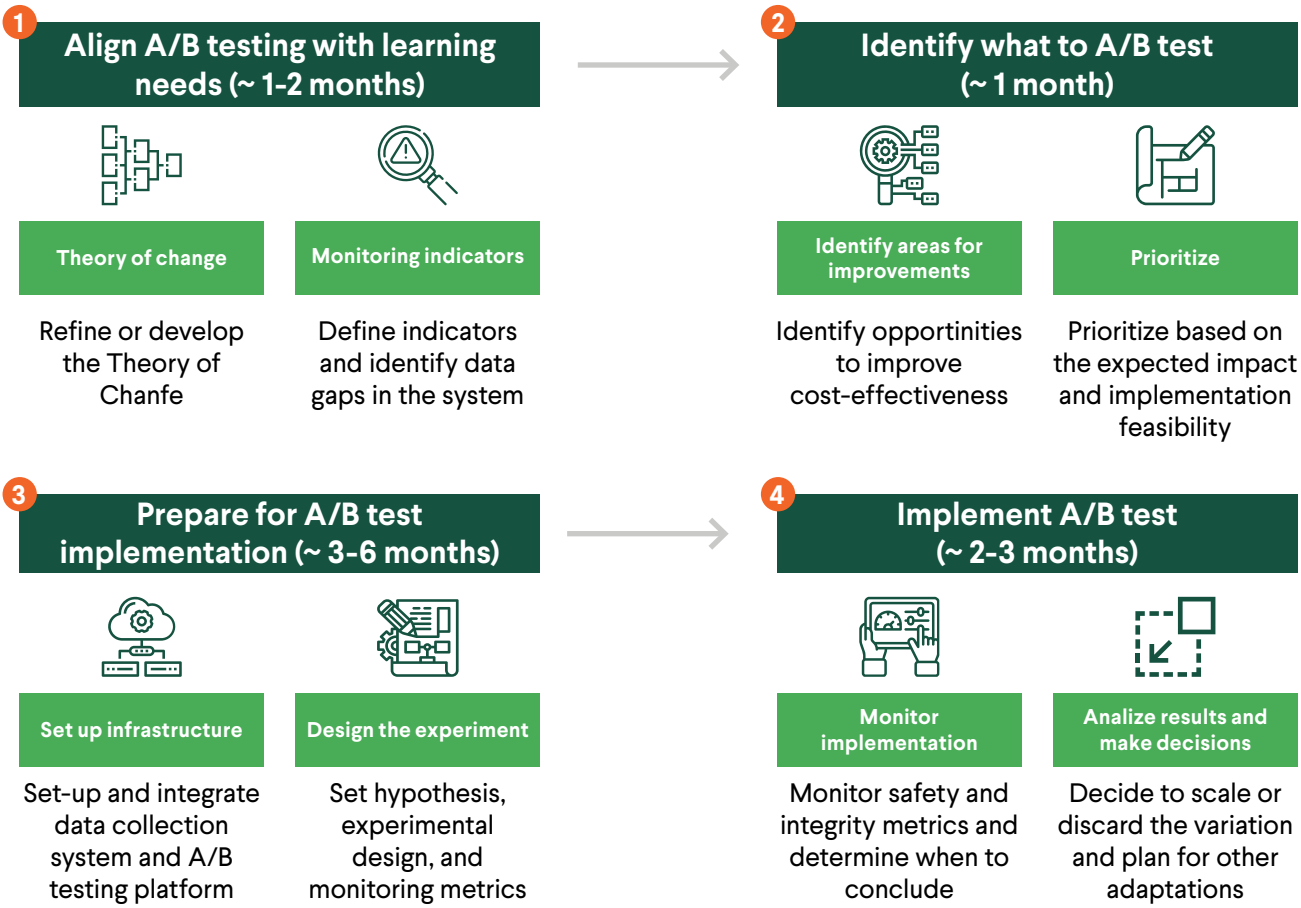
The **Learning Roadmap** offers a structured approach to adopting A/B testing by guiding organizations through the development of the capabilities required to do it well. It helps organizations strategically identify what to test, build necessary capabilities, and conduct experiments that are both rigorous and practical. It ensures that product, technology, and MEL activities move in coordination, so that A/B testing becomes a reliable learning tool for informing program and product decisions.

The Learning Roadmap consists of four core steps:

I. Aligning A/B testing with learning needs
II. Identifying and prioritizing what to test
III. Preparing for implementation
IV. Implementing and learning from the A/B tests for decision-making

In the sections that follow, we describe each step in detail, drawing from our experience supporting organizations that were new to A/B testing but eager to make it a meaningful part of their strategy.

Figure 3. Learning Roadmap for A/B Testing



STEP 1: ALIGNING A/B TESTING WITH LEARNING NEEDS (~1-2 MONTHS)

WHO SHOULD BE INVOLVED IN THIS STEP AND AT WHAT CAPACITY?

Indicative estimates:

- 20-30 percent of MEL leadership to understand drivers of impact and key steps in the ToC and monitoring indicators
- 10-20 percent of a product or program manager to understand drivers of impact and key steps in theToC
- ‹5 percent of senior leadership to secure strategic alignment on ToC and organizational buy-in

A/B testing is most valuable when it is anchored to a clear hypothesis of how an intervention is expected to generate impact. The first step is to revisit the **Theory of Change**[8] **(ToC) to clarify how the product or program is expected to generate meaningful outcomes**. The ToC helps teams identify the **core levers of impact,** the early outcomes that signal the intervention is working, and the parts where there is **uncertainty or room for improvement.** To start this journey, you can use **this guide** to design or refine your intervention's ToC. Additionally, a **user funnel**[9], which is a model that illustrates the step-by-step journey taken by users, can complement this exercise nicely by capturing granular dynamics within a portion of the ToC, a specific user segment, or the transition from one funnel stage to the next (e.g. product discovery to account activation). Hence, the user funnel serves as complementary tools to **identify the areas with the highest potential for A/B testing.**

Anchoring testing in this structured way of thinking ensures that it is used to improve strategic levers of change, not just make superficial tweaks that do not lead to improved outcomes.

> *We suggest using A/B testing to assert early linkages in the ToC, focusing on instances or components of the user journey that are relevant to early and intermediate outcomes, such as changes in knowledge, attitudes, beliefs and behavior. However, if the organization is able to rapidly and credibly collect final outcomes, these can also be used for testing. For example, Youth Impact has A/B tested implementation variations and measures how these affect learning outcomes such as foundational numeracy skills (Angrist et al., 2024; Angrist et al., 2025).*

[8] A theory of change is a structured, logical representation of how an intervention is expected to create impact. It highlights the activities and outputs that generate the hypothesized sequence of outcomes that lead to long-term goals, with early outcomes serving as key signals that the intervention is on the right path.
[9] For more guidance on how to design a user funnel consult the User Funnel Playbook published by The Agency Fund.

With a clear vision of how the product aims to create impact, grounded on a ToC and complementary user funnels, the next step is to define the metrics that the data system should produce to monitor and enable testing over early outcomes, and if possible intermediate and final outcomes. A/B tests remain valuable to test whether a product or intervention is strengthening early drivers of impact, such as whether students complete a practice quiz, teachers open feedback reports, or users return to the app within a week. For defining meaningful metrics, select those that reflect signals of positive change towards program cost-effectiveness in terms of reach, meaningful user engagement and experience, and cost.

For example, if your product is facing barriers to uptake, define how to track the number of "active users" as a meaningful measure of engagement. Typically, this represents a discrete user interaction, or a threshold amount of time spent with the product, indicating the user received some value (The Agency Fund, 2025). Specificity is key to capture data that actually matters. An "active user" could be defined as someone who created an account and logged-in once in the last year, or as someone who logs in and spends at least 30 minutes per week in the app, metrics that reflect very different levels of engagement.

> *Making these distinctions early helps identify critical data gaps and inform what investments may be needed to strengthen the data infrastructure required for effective testing. The goal at this stage is to identify what key signals of change need to be captured in a timely and reliable manner.*

STEP 2: IDENTIFYING A/B TESTING PRIORITIES THROUGH COST-EFFECTIVENESS (~1 MONTH)

WHO SHOULD BE INVOLVED IN THIS STEP AND AT WHAT CAPACITY?

Indicative estimates:

- 20-30 percent of a product or program lead to identify and prioritize experimenting priorities
- 20-30 percent of MEL leadership to identify and prioritize experimenting priorities
- ‹5 percent of the implementing or programs team to support the prioritization process
- ‹2.5 percent of an engineer to support the prioritization process
- ‹5 percent of senior leadership to secure strategic alignment

Once a team has developed a solid understanding of their intervention's ToC and identified key performance and outcome metrics, they will be better positioned to make strategic decisions about what is worth testing. To focus efforts, teams can use a **cost-effectiveness**[10] lens to identify the most promising areas for optimization. Broadly, cost-effectiveness can be improved by:

- **Expanding reach**[11] when the user base is small or uptake is low.
- **Enhancing efficacy**[12] when expected changes in early outcomes are not achieved or when key assumptions are not holding.
- **Increasing efficiency** by reducing costs while achieving the same results.

Each lever is linked to specific metrics that guide what to test (see Table 2):

- For expanding reach, this might mean **increasing registration rates or improving accessibility and onboarding.** For example, testing changes such as offering other login options (WhatsApp or email based), setting a referral scheme, or making the landing page more appealing to users.
- For enhancing efficacy, it might involve focusing on optimizing early outcomes and other relevant assumptions that enable or lead to final outcomes. Such as **increasing meaningful task-related**[13] **engagement, optimizing changes in knowledge, attitudes, beliefs** or short term **behaviors**, or **enhancing consistent-usage** aligned with the ToC. For example, testing nudges, changes in content, or impact-related features.
- For increasing efficiency, it may involve **reducing cost-drivers** like **time of service delivery** by testing lighter-touch delivery models, streamlining intensive components, or shifting to lower-cost formats (e.g. digital, AI-enabled), while maintaining the same level of impact per user.

[10] Cost-effectiveness measures how efficiently an intervention converts resources into meaningful impact by considering the number of people reached and the depth of change achieved on a specific outcome.
[11] The extent to which an intervention is accessible to its target population.
[12] How well is the intervention addressing users' needs? Refers to the depth or magnitude of the positive change achieved per user.
[13] There is still mixed evidence about the correlation between engagement and learning outcomes, but it is clear that the impact of educational programs on student learning is contingent upon the quality and fidelity of their implementation (Vanacore et al., 2023). Additionally, high-quality engagement, such as thoughtful completion of assessments, effective use of hints, and meaningful interactions with platform content, has proven to be a stronger predictor of gains in learning outcomes than simply measuring time spent or the number of interactions (Ruipérez-Valiente et al., 2018; Muñoz-Merino et al., 2013; Kelly et al., 2013).

Table 2. Identifying A/B Testing Priorities Through a Cost-Effectiveness Framework

Lever	Focus areas for changes	Learning focus
Reach	Increase scope of reach and acquisition	Number of registered users
	Ease accessibility	• % of active users • % of users who complete onboarding
Efficacy	Increase meaningful engagement	• Outcome-related task completion rate • Feature utilization rate • Frequency and depth of interaction (time spent) • Follow-up question rate • % of users who transition from passive to active usage
	Optimize knowledge gains or behaviour change	Changes in users' knowledge, beliefs, perceptions and behaviors[14]
	Enhance capacity to retain users engaged over time	• Session duration • Return rate • Day-n retention rate[15]
Efficiency	Reduce costs or time of service delivery	Total operational costs

Once a priority lever (reach, efficacy, efficiency) and its respective target metric have been selected, product and program teams can lead a light-touch ideation process to identify specific variations to test. These teams, being closest to delivery and user experience, are well positioned to suggest changes that are both feasible and relevant to improve the metric. While we do not present a detailed brainstorming approach, even brief, structured conversations can help surface strong ideas for research.

[14] The Agency Fund's AI Evaluation Playbook provides detailed guidance on measurement tools for assessing psychological and behavioral changes in users.[11] The extent to which an intervention is accessible to its target population.
[15] The % of users who return on day "n" after their first interaction. Similarly, you could define week-n or month-n retention metrics (The Agency Fund, 2025).

PRIORITIZING A/B TESTING OPPORTUNITIES

After identifying several A/B testing opportunities, the next step is to prioritize them using two simple criteria:

I. **Relevance:** How likely is this variation to improve a key outcome metric? Focus on ideas with the potential to improve results in areas linked to your cost-effectiveness priorities. Use data, prior tests, predictive models, or relevant research to make an informed judgment of the potential impact.
II. **Feasibility of implementation:** How easy or costly will it be to develop and deploy this variation? Consider, with the implementation and product teams, the costs, effort, and risks of designing and implementing the variation.

For each potential change, rate both relevance and implementation feasibility on a scale from 1 to 3 (see Figure 4):

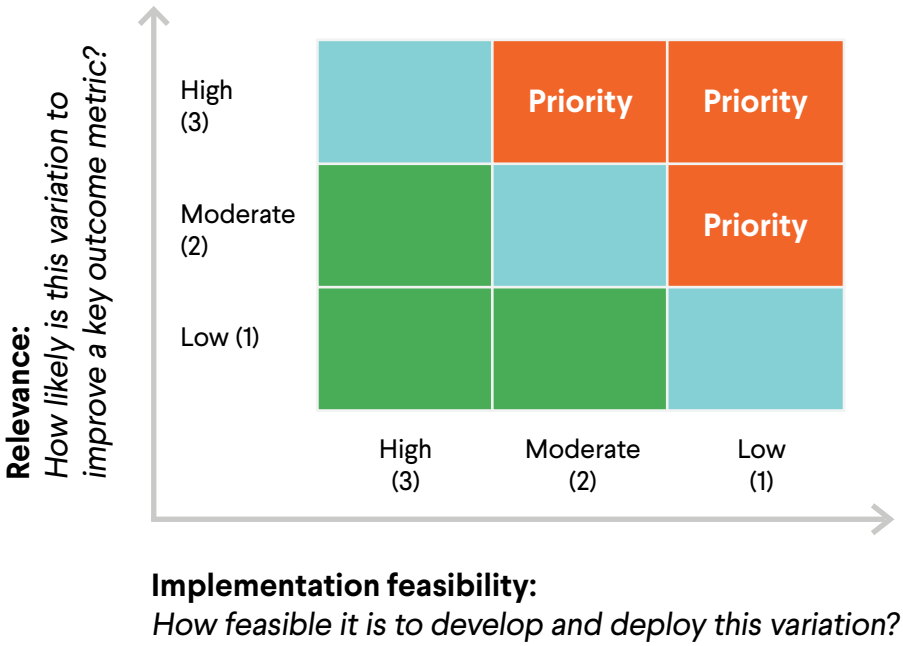
- 1 is low relevance / low implementation feasibility
- 2 is moderate relevance / moderate implementation feasibility
- 3 is high relevance / high implementation feasibility

During the prioritization process, include the product manager and representatives from field and engineering teams, as these perspectives are essential for credibly assessing each variation's implementation feasibility. Use existing research and data trends to assess how likely a change is to improve the key outcome metric you care about. Where evidence is limited, informed judgment might be necessary. While these criteria provide a useful framework, additional contextual factors may also influence prioritization decisions. Once each variation has a score, select **two to three A/B tests** that fall within the **priority area** (see Figure 4). **Focus first on variations with high relevance and high feasibility and then consider those with moderate to high relevance and feasibility** (see Box 2 for a detailed example).[16]

After the prioritization exercise, you will typically have around four to six A/B tests in the pipeline. Sequence these tests based on (1) your organization's decision-making priorities and (2) the maturity of your data systems and experience with experimentation. If this is your first time A/B testing, start with a low-stakes test to build experimentation muscle. Then, sequence the remaining tests starting with the highest priority and define a timeline for subsequent testing cycles. We recommended following an iterative learning approach, where the results of one test inform the design of the next test. Account for the time needed to design each variation, implement the test, and observe meaningful results. For example, an A/B test aimed at improving user registration may run for a shorter period than one targeting learning outcomes, since learning gains take longer to materialize.

[16] To organize the variation list and prioritization process you could use a template similar to Youth Impact's AB testing question brainstorming tool.

Figure 4. Prioritization Criteria



BOX 2. PRIORITIZATION EXAMPLE

An organization is considering three A/B testing opportunities that aim to increase platform use. The scoring might look like this:

- **Improving the onboarding process** is expected to have a *high (3)* relevance for increasing platform use and is *moderately (2)* feasible, as it requires investment in developing an improved onboarding experience.
- **Sending email reminders** is expected to have a *low (1)* relevance for increasing platform use and is *highly (3)* feasible because it is low cost and doesn't require development effort.
- **Gamifying the platform** is expected to have a *high (3)* relevance for increasing platform use and has *low (1)* feasibility because it requires a significant investment and poses some risks to platform performance.

Based on this assessment, improving the onboarding process would be prioritized first, as it offers both high relevance with moderate feasibility.

STEP 3: PREPARING FOR IMPLEMENTATION (~3-6 MONTHS)

WHO SHOULD BE INVOLVED IN THIS STEP AND AT WHAT CAPACITY?

Indicative estimates:

- 20-30 percent of a data scientist* to design tests, define the monitoring metrics, and manage and analyze data
- 20-30 percent of an engineer* to set-up the data infrastructure and integrate the experimentation platform
- 10-15 percent of the technology team lead* to guide the set-up the technological infrastructure
- 20-30 percent of MEL leadership to support the design of the tests and define the monitoring metrics
- 10 percent of a product or program manager to support the design of the tests

*These estimates assume the organization is building the data analytics and experimentation systems from scratch. In some organizations the data infrastructure and data analysis roles are concentrated in one single team member.

Careful preparation is essential to running A/B tests that generate clear, actionable insights. This involves first setting up the necessary technological infrastructure to collect data on selected metrics and deploy experiments; defining the experiment's design including the hypothesis and success criteria; designing and developing the variation; and finally setting a monitoring plan.

SETTING UP THE TECHNOLOGICAL INFRASTRUCTURE[17]

The goal at this stage is to ensure your organization can capture key metrics reliably and in a timely manner, manage tests, monitor implementation, and analyze and visualize results. In most cases, organizations need two core components:

1. An **analytics system** to collect, store, and visualize program data
2. An **A/B testing platform** to randomly assign users to variations and deploy experiments

The process to build this infrastructure is: first establish the data collection system to monitor your program continuously, then select and integrate the experimentation tool, and finally pressure test the system (e.g., through an A/A test) before implementing your first A/B test. Organizations with mature data systems may skip or accelerate some of these steps.

This infrastructure is a long-term investment, so aligning the setup with your organization's specific needs, scale, and technical capacity is key to sustained adoption. Whether your organization has a nascent or mature data system, the guidance that follows will help you select appropriate tools and configure them. Based on our advisory experience, building the

[17] Setting up the technological infrastructure for A/B testing can run in parallel with steps one and two as data and infrastructure gaps emerge.

technological infrastructure required for monitoring key metrics and experimenting may take three to six months.

Selecting the tools for the analytics system:

To be ready for testing , your organization needs a functional data analytics system that reliably collects program data at scale. The system must also store the data and make it accessible for analysis and decision-making. For this you will need:

- **Data collection system:** Captures user interactions, behaviors, outcomes, feedback, and platform performance.
- **Data warehouse:** Stores the raw and processed event and outcome data so it can be queried and analyzed. It is important to review the data security that different warehouses offer.
- **Data cleaning, analysis, and visualization tool:** Enables cleaning raw data, querying data to calculate metrics, run statistical analysis, and visualize data through dashboards (graphs and tables).[18]

Digital-first products offer an advantage at this stage, as they make it easier to collect rich on-platform and interaction data within the app or website. However, some organizations also collect impact-level outcomes outside the digital product, for example through classroom observations, in-person surveys, or written assessments. At this step, ensure that all relevant data sources are integrated and stored in a single data warehouse, using unique identifiers that allow datasets to be merged. Your organization may already have all or some of these tools in place as part of your monitoring system; the next step is to ensure they function together as a coherent analytics system.

BOX 3. CHECKLIST TO CONFIRM YOUR ANALYTICS SYSTEM IS CAPABLE OF SUPPORTING EXPERIMENTATION[19]

- ☐ **Unique identifiers:** Each record has a unique identifier (e.g., user ID, student ID) that allows data to be linked across sources and tracked over time
- ☐ **Frequency:** Data is collected on a regular timeline (e.g. continuously, monthly, quarterly), rather than in an ad hoc basis
- ☐ **Scale:** Your system can handle your full user base, not just a subset or pilot group
- ☐ **Integration & automation:** Data flows automatically through your system, without manual entry or file uploads
- ☐ **Storage & accessibility:** Data is securely stored with PII encrypted or anonymized, and accessible to relevant team members
- ☐ **Metric definition:** Each metric has a documented definition that is applied consistently across the organization
- ☐ **Data analysis:** You can calculate key metrics and compare them across user groups
- ☐ **Visualization:** Dashboards display key metrics and experiment results in accessible formats
- ☐ **Reliability:** Data can be validated against trusted sources (e.g., server logs, historical records)

[18] To conduct these three processes (data cleaning, analysis and visualization) you could use one or more tools, depending on the capabilities of the tools that you choose.
[19] For complementary guidance on assessing whether routine monitoring data is suitable for A/B testing see Youth Impact's Toolkit *Phase 2: Data Flow.*

SELECTING THE A/B TESTING PLATFORM:

An A/B testing platform typically provides two essential capabilities:

- **Variation deployment:** Most commercial and open-source platforms offer interfaces to create and deploy front-end changes (e.g., changing button colors, showing messages, or rendering conditional UI elements). Some platforms also support server-side testing, where backend modifications (e.g., enabling voice messaging or triggering a different workflow) are deployed to different groups. Server-side changes usually require more engineering effort and integration work.
- **User randomization:** There are two common approaches to assigning users to test variants.
    - **Pre-assigned randomization:** Users are randomly assigned to variants before the test begins, and this assignment is stored and later accessed by or uploaded to the A/B testing platform. This is the standard approach when the intervention happens outside the digital product (e.g., in-person workshops or phone-based outreach) and assignment must be determined before delivery. This approach is also common in batch-based interventions where you know who the participants are before the intervention starts, so you can randomize them in advance. Or, when you have baseline data on participants and want to stratify randomization by key characteristics (e.g., ensuring equal distribution of high- and low-engagement users across groups).
    - **On-the-fly randomization:** Users are randomly assigned to variants at the moment of interaction with the system and they receive the appropriate variation (A or B) based on a pre-specified event, behavior, or condition[20]. This approach works well when users arrive continuously and unpredictably (e.g., new app sign-ups or website visitors), you want assignment tied to a specific user action (e.g., clicking a button, completing a lesson), or when you don't have prior information about users before they interact with the product.

At a minimum, ensure the A/B testing platform supports:

- Experiment setup and variation deployment
- Randomization and assignment logic
- Targeting, segmentation, and stratification[21]
- Data export or integration with your analytics system

Note that some more comprehensive A/B testing platforms offer data analysis and visualization capabilities to monitor experiments and visualize results, which can be very useful. However, through our partnerships we have learned that having an internal data analysis and visualization set-up outside the tool can be especially useful to make broader analysis and replicate results.

[20] Alternative approaches such as multi-armed bandits and adaptive testing adjust allocation probabilities over time based on observed outcomes. This mechanism dynamically assigns more users to better-performing variants.
[21] Stratification means dividing users into homogeneous groups (strata), for example, men and women, and then randomly assigning individuals within each group to variant A or B, to ensure that men and women are balanced across the variants.

General key considerations when selecting tools:

- **Compatibility:** Choose tools that integrate smoothly with your organization's existing technological infrastructure, including data pipelines, front-end frameworks, and backend services.
- **Cost:** Evaluate licensing fees, implementation costs, hosting,[22] and support expenses. Open-source options may be free or charge reduced fees.
- **Feature Requirements:** Define the organization's essential "must-have" vs. optional "nice-to-have" features.

There are various A/B testing tools available, ranging from open-source tools such as **Evidential** and **UpGrade**, where you can find non-profit specific solutions, to proprietary tools, to developing the system in house. The costs of each option can vary significantly based on your usage (e.g., number of users or experiments), storage requirements, and support expectations. We recommend evaluating at least two open-source and two commercial options to make a well-informed decision based on functionality, compatibility, and total costs. Building the experimentation system in house is least recommended, as it typically requires one full-time staff member to build, maintain, and monitor the system. In addition, in-house solutions are more prone to errors while the system becomes robust, which can create additional cost and undermine trust in results. Given the availability of well-tested tools maintained by specialized teams, building and maintaining a custom experimentation system rarely offers enough benefits to justify the long-term cost and risk.

BOX 4. SELECTING AN A/B TESTING PLATFORM

Example 1: Commercial and open-source options

A P4T-supported EdTech organization running an AI-enabled teacher assistant on the web explored several commercial A/B testing platforms. While many were technically compatible with their stack, licensing costs were a barrier. With IPA's support, the team evaluated open-source alternatives and ultimately chose **GrowthBook**. The free tier allowed them to run unlimited experiments, and a modest upgrade provided access to premium support, meeting their experimentation needs without locking them into high recurring costs. Building a custom solution was considered but ruled out due to the long-term development and maintenance burden.

Example 2: When an A/B testing platform isn't needed

Another P4T-supported organization delivers its intervention through a WhatsApp chatbot managed via a messaging platform. During the tech set-up, we realized that a standalone A/B testing tool was unnecessary because the messaging platform already supported random assignment to different chatbot flows, which is the core requirement for A/B testing. Because content changes and logic updates could be implemented directly within the messaging platform, the organization was able to run experiments without adding new tools, keeping the setup simple and aligned with their operational capacity.

[22] Hosting refers to where the data and system will be stored.

Once you've selected the tools for the data analytics system and A/B testing platform, set them up and integrate them, making sure that data flows seamlessly between these two components. Finally, before launching the test, run a mock to validate that the system is working correctly and that the data pipelines between tools are well integrated. Faults in the system can lead to misleading results and false conclusions.

BOX 5. HOW TO CHECK IF YOUR EXPERIMENTATION SYSTEM IS READY TO GO LIVE?

Before launching a real A/B test, run a mock test (known as A/A test) to validate that the system and its pipelines are working correctly. In this test, users are randomly split into two groups receiving the same version of the intervention, which in essence is not testing anything. Verify the following:

- **Check** that the **system is collecting all the necessary data** to monitor the experiment.
- **Compare the metrics collected** against trusted sources, such as server logs or known historical data. Any discrepancies such as misaligned timestamps, missing values, or inconsistent values should be resolved before an experiment goes live.
- **Check for proper random assignment,** where the distribution of users between the two groups is split according to the initial experimental design, for example 50% of the sample of users go to one group and 50% to the other.
- **Check for important differences between the groups across key variables,** such as a group with a larger portion of users with access to a computer at home. In theory, there shouldn't be an actual difference between the groups, so any differences observed may indicate flawed randomization, inaccurate data collection, or hidden bugs.

DESIGNING THE EXPERIMENT

A good experiment begins with a clear learning question and a testable hypothesis. The hypothesis is a well-reasoned guess that articulates how the change being tested might influence an outcome metric. A good hypothesis would be: *If we send myth-busting messages about the vaccine via WhatsApp (X), the intention to vaccinate (Y) will increase because adolescents' erroneous beliefs will be corrected (Z).* It clearly states:

- The **change being tested** (e.g., sending myth-busting messages)
- The **outcome you aim to influence** (e.g. intention to vaccinate)
- The **reason why** the change should work (e.g. correcting false beliefs)

Grounding hypotheses in user research, existing data, or prior studies strengthens the validity of the test. In this example, knowing adolescents' beliefs, trusted information sources, and decision-making patterns helps identify both what to test and why it might work.

Next, define a **primary outcome metric** that measures whether the change worked and identify how to capture the data. Additionally, identify **secondary complementary metrics** that can help explain why the change did or did not generate the desired effect.

[23] To review the power calculations for the test, you can use *J-PAL's Quick Guide to Power Calculations*.

Table 3. Example: Defining metrics for the experiment

Metric	Data source
Primary outcome metric: Percentage of adolescents who request an appointment to get vaccinated	Internal health appointment system
Secondary metric: Change in the percentage of adolescents who believe that the vaccine generates cancer.	Pre and post intervention perception surveys

Plan the test's statistical requirements[23] to ensure reliable results. A key factor is sample size which is the number of users (or sessions, depending on the unit of analysis) needed in each group (A and B) to be able to reliably detect an effect if one exists. If the sample size is too small, the test might fail to detect a real effect, leading to the wrong conclusion that the change didn't work when in fact the treatment was beneficial. Ultimately, leading you to wasting resources on a test that can't deliver a clear and reliable answer to your hypothesis.

To calculate the right sample size, you'll need to define the minimum effect size you want to be able to detect, for example, a 5 percentage point increase in completion rates. Free online calculators or your A/B testing platform's built-in calculator can estimate the required sample based on your baseline metric, expected effect size, and desired statistical confidence.[24]

Finally, plan the operational details of the experiment:

- Define when and how to assign users to the test. As mentioned above, there are two possibilities: predefined assignment or "on-the-fly" assignment. Some examples of the latter include:
  - Page load: A user is assigned to a variation when they visit a specific webpage (page load event).
  - Button click: The test starts when a user interacts with a specific element, such as clicking a button or link.
  - Session start: The user is assigned to a variation at the beginning of their session, and the same variation persists throughout the session.
  - Custom event: Assignment occurs when a specific event is logged, such as completing a form, scrolling to a certain point, or watching a lesson.
  - Time-based: The test triggers at a specific time or during a defined time window.
- Define when the A/B test will start and how long it will run to avoid result peeking. Running tests long enough is critical to avoid misleading results. When defining the timeline, account for the implementation context and the expected time required to observe changes in the target outcome. For example, if testing a feature that affects weekly engagement (e.g., streaks or weekly challenges), run the test for at least 3-4

[24] Some A/B testing platforms, like Evidential, have built-in sample size calculators. If yours doesn't, free tools like Evan Miller's A/B test calculator can be useful.

weeks to capture natural variation in user behavior across weekdays and weekends. Alternatively, If testing a variation aimed at improving learning outcomes, plan for a longer duration, typically 6-8 weeks or more, to allow sufficient exposure to the intervention and time for measurable learning gains to emerge.

DEFINING MONITORING METRICS

Once the experiment design is clear, define the metrics you will monitor during implementation and later use for decision-making:

- System and product health metrics are used to check whether an experiment is unintentionally harming the user experience or platform performance. Monitoring these metrics helps detect potential issues early. For example, if during an experiment you see a sharp rise in the percentage of users starting but not completing a curriculum, this might indicate that the new content is disengaging or confusing users. In such cases, the team can pause or stop the experiment to avoid further harm.
- Research integrity metrics assess how reliable the results of the test are. They help check both whether the test groups are comparable and whether the intervention was delivered as intended. These metrics typically fall into three categories:
  - Assignment integrity checks whether users were correctly randomized and assigned to groups as planned. For example, if you planned a 50-50 split, in a sample of 500 users about 250 should be assigned to each group. If one group is much bigger than the other (e.g. 400 vs 100), it is likely that the randomization is not working well. Check this in your A/B testing platform dashboard or by running a simple count query in your database during the first few days of the experiment. If you detect an imbalance, pause the test and investigate the randomization code before continuing.
  - Group comparability checks whether the groups have a similar mix of key characteristics that could affect outcomes. Typical characteristics to check include user demographics (age, gender, location), prior engagement levels (active vs new users), or device type (mobile vs desktop). Identify and review these metrics in your baseline or historic data before launching the test. For example, if you define that gender might affect the outcome, check that both groups have about the same number of females and males.
  - Implementation fidelity checks how closely the experiment was delivered as intended. For example, whether the correct variation (features or content) was shown to the right users. This can be checked by querying which variation each user was shown and comparing it against their assigned group in your experimentation platform. Or, if your program is non-digital this implies checking whether each participant received the program version they were pre-assigned to.
- Results metrics, which you will have defined when designing the experiment, are used to evaluate the impact of the change. The primary metric shows whether the change worked, while secondary metrics help explain why it worked (or why it didn't).

Once you've defined the metrics, design how you will structure your data for analysis. Define whether each row in your dataset represents a user, a session, or another unit of analysis, and list which variables (columns) you need to calculate your metrics. This ensures that you have

all required data and that it is organized at the appropriate unit of analysis, enabling proper analysis and visualization of the results. Then, set-up or customize the monitoring dashboards in your analytics system to visualize health, research integrity and results metrics during and after the test. Finally, with a well-designed experiment in place and monitoring plan, the next step is to design the variation.

DESIGNING AND DEVELOPING THE VARIATION

When designing test variations, such as the content of the myth-busting messages, draw on evidence and user insights to inform design choices. Also, plan for the budget to develop and implement the variation, keeping in mind that more complex changes may require greater investment. Before launching, validate that the variation works as intended and is understandable to implementation teams (if relevant).

BOX 5. GOOD PRACTICES FOR DESIGNING THE VARIATION:

- Gathering user input to inform design
- Review evidence of what has worked in similar interventions
- Reviewing past behavior data related to the target metric
- Testing variation functionality and observing user experience with it
- Refining hypotheses based on this feedback

STEP 4: IMPLEMENTING A/B TESTS (~2-3 MONTHS)

WHO SHOULD BE INVOLVED IN THIS STEP AND AT WHAT CAPACITY?

Indicative estimates:

- 5 percent of the technology team lead to support troubleshooting during the experiment
- 5 percent of an engineer to support troubleshooting during the test
- 30 percent of a data scientist to monitor the test and analyze results
- 20-30 percent of MEL leadership to support monitoring and analysis, and lead interpretation and decision-making sessions
- 5 percent of a product or program manager to support decision-making when test results are ready

With the experiment design and variation in place, your organization is ready to launch. Once the A/B test is live, it is important to monitor it carefully to verify that it is working correctly. When the test has concluded, analyze and interpret the results in a way that supports sound decision-making.

MONITORING KEY METRICS

During the test, regularly monitor the execution and some key metrics, defined in your experimentation plan. This ensures that the test is running correctly and that the results will be reliable.

- Verify that the microdata is being collected and stored securely and completely.
- Continuously monitor safety metrics throughout the test to quickly detect and address any negative impacts on user experience or system performance. If the problem can't be fixed, consider stopping the test.
- Monitor the number of users participating in the test . If the sample is smaller than expected, consider actions to increase reach, such as sending reminder messages inviting users to participate in the intervention or allowing the test to run for a few more weeks than originally planned.
- Check integrity metrics at least two weeks into the test, to verify random assignment and comparability between groups, as well as implementation fidelity.
- Analyze results metrics after allowing sufficient time for data stabilization. This helps avoid drawing conclusions too early or mistaking short-term patterns for real effects.

Timely and careful monitoring supports accurate interpretations and enables informed decision-making.

ANALYZING AND INTERPRETING RESULTS FOR DECISION-MAKING

Once the test has run for the predetermined period, gather the relevant team members to analyze the results. A structured analysis process should guide the discussion and decision-making process, which involves three key steps:

- **Assess the credibility of the test:** Begin by reviewing the integrity metrics to evaluate whether the results are trustworthy. First, check if the sample size is large enough to detect a meaningful effect based on your power calculations. Then, check whether the groups are statistically comparable on key characteristics. Based on the sample size and balance between the samples, discuss the level of confidence you have in the validity and reliability of the results.
- **Assess the effectiveness of the variation:** Examine the outcome metric to determine the impact of the tested variation. What does the data suggest about the effect of the change? Is there a clear difference in performance between groups? Try to identify a plausible mechanism that could explain how the variation produced the observed result.
- **Interpret the data and make a decision:** Based on the evidence, decide whether to extend or conclude the test. You may decide to extend it to reach a larger sample size. If concluding, confirm or reject the hypothesis based on the data and decide:
    - **If results are inconclusive or null,** consider repeating the test if you suspect issues with the design, implementation, or measurement that may have affected the outcome. Otherwise, you may choose to discard the variation and move on to the next test.
    - **If the results show a positive effect** on the results metric, you can scale the variation to all users or relevant segments, or continue testing to find an even more effective variation.
    - **If the results show a negative effect** on the results metric, discard the variation and use the findings to inform future iterations.

Make sure to use findings from the A/B test to make adaptations to the product or intervention. It is not only about scaling or discarding the variation, but understanding in depth why something worked better (or not) and using those insights to make other adaptations and decisions, test other things, and even share learnings with the broader development community.

BOX 6. EXAMPLE:

An education platform is testing whether adding personalized feedback increases student engagement. After three weeks they analyze results:

- Safety metrics have shown that system performance remains stable, with no increase in load time or crash rate.
- Integrity checks confirm that the test has been implemented as planned, and both groups are balanced by grade level and prior average engagement time.
- The sample size is sufficient to detect a 5 percent change in the primary outcome metric (time spent in the platform).
- The results metrics suggest a 6 percent increase in engagement in the group that received personalized feedback compared to the control group, with additional evidence from user click data showing that students are actively reading the feedback.

Based on these findings, the team may decide to conclude the experiment and scale the change across all users. However, if the sample size had been too small or the groups had been unbalanced, they might extend the test before making a final decision.

Given the significant effect of personalized feedback on user engagement, the organization decides to test whether video or text feedback is more effective in helping students answer the next set of questions correctly.

Finally, allocate time to reflect on the full process. Document key challenges, unexpected findings, and lessons learned during the design and implementation phases. Use these insights to plan improvements for the next cycle of testing. This reflective practice is essential to strengthen your organization's A/B testing capabilities and ensure continuous learning.

Implementing A/B testing is not about getting everything perfect from the start. It is about building capabilities that allow organizations to test, learn, and adapt over time. The Learning Roadmap provides a structured guide to embark on this process, helping organizations progressively strengthen their experimentation capabilities. Once an organization has completed all the steps in the Roadmap, including running its first test, it is ready to experiment iteratively. The number of tests an organization can run per year depends on their systems and analytical capacity, but the following timeline is indicative: one month for test preparation, one to three months for implementation (depending on how long it takes to observe changes in the outcome of interest), and one month for analysis and decision-making. Preparation for the next test can begin while another test is live, provided they do not influence each other. Otherwise, organizations should first analyze results and make decisions before using those insights to inform the next test. Remember that the goal is not to run A/B tests indiscriminately, but to use research as a strategic decision-making tool to improve program cost-effectiveness and scalability.

The next section offers additional practical guidance drawn from early experiences supporting organizations in adopting A/B testing.

4. Lessons learned: Learning Roadmap Implementation

Adopting A/B testing is not just a technical task. It is about building the mindset, processes, and capabilities that allow your organization to test, learn, and improve continuously. In our work with digital-first organizations,[25] we have learned that effective A/B testing works best when it focuses on strategic questions, relies on evidence-based design, is supported by reliable systems, and is embedded in a culture of continuous learning. Without these, even well-run tests may fail to drive meaningful program improvements.

This section offers practical guidance for teams seeking to build organizational capacities for implementing A/B tests that generate actionable insights.

Focus on learning that matters: A/B testing should help your team answer questions that lead to better programs or products. To do this, anchor A/B testing in a theory of change and cost-effectiveness lens. Focus on testing changes that can meaningfully improve impact, not just minor user experience tweaks.

A/B testing is not always the right tool: A/B testing requires data infrastructure, engineering effort, and user exposure, so it's important to use it for questions where the potential insights justify the investment. Think of policy relevant questions and not just marginal changes. When outcomes are predictable or the expected learning value is low, other methods may be more appropriate and less costly.

Imagine an EdTech platform is considering whether the interface color has an effect on student learning outcomes. In this case, running an A/B test is not the best approach. Instead, doing a survey to understand user preferences and reviewing existing human-computer interaction literature would be more appropriate and cost-effective methods to inform this decision before investing in an A/B test.

Keep tools and systems simple: A/B testing works best when your tools integrate smoothly with your existing systems. Prioritize tools that work well with your current data and engineering setup. Simple integrations reduce costs and errors in data collection and analysis. Keep in mind that tools with strong documentation and support will be easier to maintain over time.

Effective monitoring requires upfront planning: Good A/B testing depends on good data. Before launching a test, define what data will be collected and from which sources, and ensure that your data systems can support the analysis you will need.

[25] These learnings have emerged through work with partner organizations, Mentu Labs and Tirando x Colombia, during the development of their A/B testing systems.

Align teams on metrics and technical aspects: Align teams (program, MEL, data) on metric definitions and data needs to ensure consistent collection and valid interpretation of results. Clarify roles for data collection, monitoring, and analysis.

Build experimentation capacity through iteration: A/B testing is a capability that grows with practice. Begin with A/A tests to verify that randomization and data flows work correctly. Run simple, low-stakes A/B tests first to build team confidence. Use early tests as learning opportunities to refine your processes.

Ensure tests are valid and reliable: For results to be trustworthy, check that test groups are balanced in size and across key contextual variables, and verify that the intervention was delivered as intended in each group.

Interpret results for practical value, not just significance: When analyzing results, focus on whether observed differences are meaningful for program or product decisions, not just whether they are statistically significant.

Document and share learnings: Institutionalize a process for documenting processes, recording test designs, results, and key takeaways so that learnings inform future iterations and broader strategy.

A/B testing is a powerful method when applied thoughtfully and with a clear learning agenda. Success depends on selecting the right questions, building reliable data systems, ensuring data quality, and embedding learning into decision-making. Organizations should view A/B testing as a capability to be developed iteratively. By starting small, building alignment across teams, and investing in systems and structured processes, organizations can turn A/B testing into a sustainable driver of program and product improvement. With practice, A/B testing can become a key part of how organizations deliver greater impact.

References

Abdul Latif Jameel Poverty Action Lab. (n.d.). Quick guide to power calculations. https://www.povertyactionlab.org/resource/quick-guide-power-calculations

Alvarez-Marinelli, H., Berlinski, S., & Busso, M. (2021). Remedial education: Evidence from a sequence of experiments in Colombia. Journal of Human Resources, 56(4), 1137–1186. https://doi.org/10.3368/jhr.0320-10801R2

Angrist, N., Beatty, A., Cullen, C., & Nkwane, T. M. (2026, January). Iterative A/B testing for social impact: Rigorous, rapid, regular. Stanford Social Innovation Review. https://ssir.org/articles/entry/iterative-a-b-testing-social-impact

Angrist, N., Cullen, C., & Magat, J. (2025, October). Cheaper (and more effective) by the dozen: Evidence from 12 randomised A/B tests optimising tutoring for scale (Working paper). What Works Hub for Global Education. https://www.wwhge.org/wp-content/uploads/2025/10/Cheaper-by-the-dozen-tutoring-AB-testing_WP_2025001_updated.pdf

Angrist, N., Beatty, A., Cullen, C., & Matsheng, M. (2024). A/B testing in education: Rapid experimentation to optimise programme cost-effectiveness. What Works Hub for Global Education. Insight Note 2024/001. https://doi.org/10.35489/BSGWhatWorksHubforGlobalEducation-RI_2024/001

Gugerty, M. K., & Karlan, D. (2018). The Goldilocks challenge: Right-fit evidence for the social sector. Oxford University Press. https://doi.org/10.1093/oso/9780199366088.001.0001

Innovations for Poverty Action Right-Fit Evidence Unit. (2024, October). Enabling stage-based learning: A funder's guide to maximize impact. https://poverty-action.org/sites/default/files/2024-10/Enabling-Stage-Based-Learning-Full-Guide.pdf

Innovations for Poverty Action. (2026, January). Theory of change. IPA Knowledge Hub. https://data.poverty-action.org/monitoring-evaluation-learning/theory-of-change.html

Kasy, M., & Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. Econometrica, 89(1), 113-132.

Kelly, K., Arroyo, I., & Heffernan, N. (2013). Using ITS generated data to predict standardized test scores. In Proceedings of the 6th International Conference on Educational Data Mining (pp. 3–4). https://www.educationaldatamining.org/EDM2013/papers/rn_paper_62.pdf

Kohavi, R., Tang, D., & Xu, Y. (2020). Trustworthy online controlled experiments: A practical guide to A/B testing. Cambridge University Press.

Kohavi, R., Deng, A., & Vermeer, L. (2022). A/B testing intuition busters: Common misunderstandings in online controlled experiments. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22) (pp. 3168–3177). Association for Computing Machinery. https://doi.org/10.1145/3534678.3539160

Muñoz-Merino, P. J., Ruipérez-Valiente, J. A., & Delgado-Kloos, C. (2013). Inferring higher level learning information from low level data for the Khan Academy platform. In Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13) (pp. 112–116). ACM Press. https://doi.org/10.1145/2460296.2460318

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado-Kloos, C. (2018). Improving the prediction of learning outcomes in educational platforms including higher level interaction indicators. Expert Systems. Advance online publication. https://doi.org/10.1111/exsy.12298

Singh, A., Navarro-Sola, L., & Oreopoulos, P. (2025). Education technology. VoxDevLit, 20(1). https://voxdev.org/voxdevlit

The Agency Fund. (2025). User Funnel Playbook for the Social Sector. https://theagencyfund.substack.com/p/user-funnel-playbook-for-the-social

The Agency Fund. (2025.). AI evaluation in the social sector: A living playbook for evaluating AI products in the social sector. https://eval.playbook.org.ai/

Vanacore, K., Ottmar, E., Liu, A., & Sales, A. (2024). Remote monitoring of implementation fidelity using log-file data from multiple online learning platforms. Journal of Research on Technology in Education. Advance online publication. https://doi.org/10.1080/15391523.2024.2303025

Innovations for Poverty Action (IPA) is a research and policy nonprofit that discovers and promotes effective solutions to global poverty problems. IPA designs, rigorously evaluates, and refines these solutions and their applications together with researchers and local decision-makers, ensuring that evidence is used to improve the lives of the world's poor. Our well-established partnerships in the countries where we work, and a strong understanding of local contexts, enable us to conduct high-quality research. This research has informed hundreds of successful programs that now impact millions of individuals worldwide.