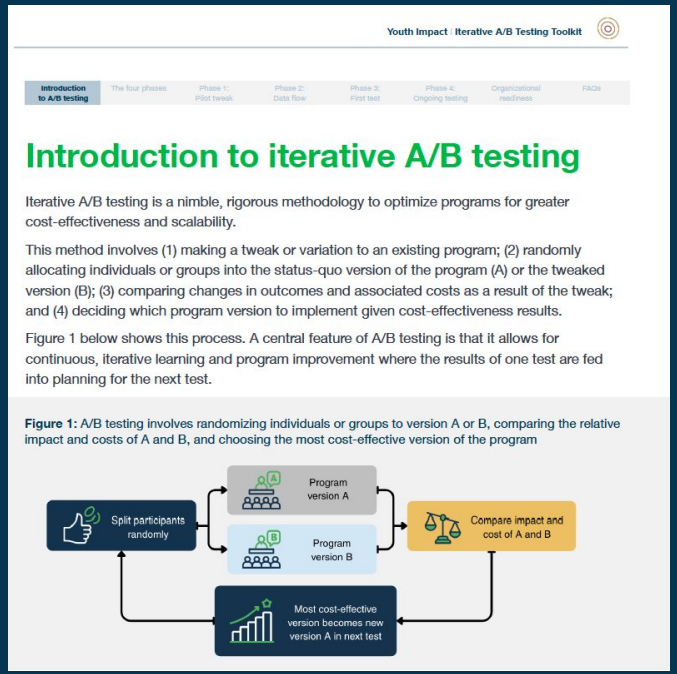
Iterative A/B testing: A practical toolkit to optimize what works at scale

A webinar with the WWH for Global Education

Amanda Beatty, Claire Cullen, Noam Angrist

11 March 2026

Check out our toolkit!



https://www.youth-impact.org/insights/a-b-testing-toolkit

A/B testing principles and groundwork

What is the goal of iterative A/B testing?

A/B testing is a rigorous evaluation methodology used to optimize and determine the “best” version of a program in terms of

Cost-effectiveness

Scalability

Fit for purpose/context

A/B testing is Rigorous, Rapid, Regular

Rigorous (like RCTs)

Uses randomization to demonstrate causality

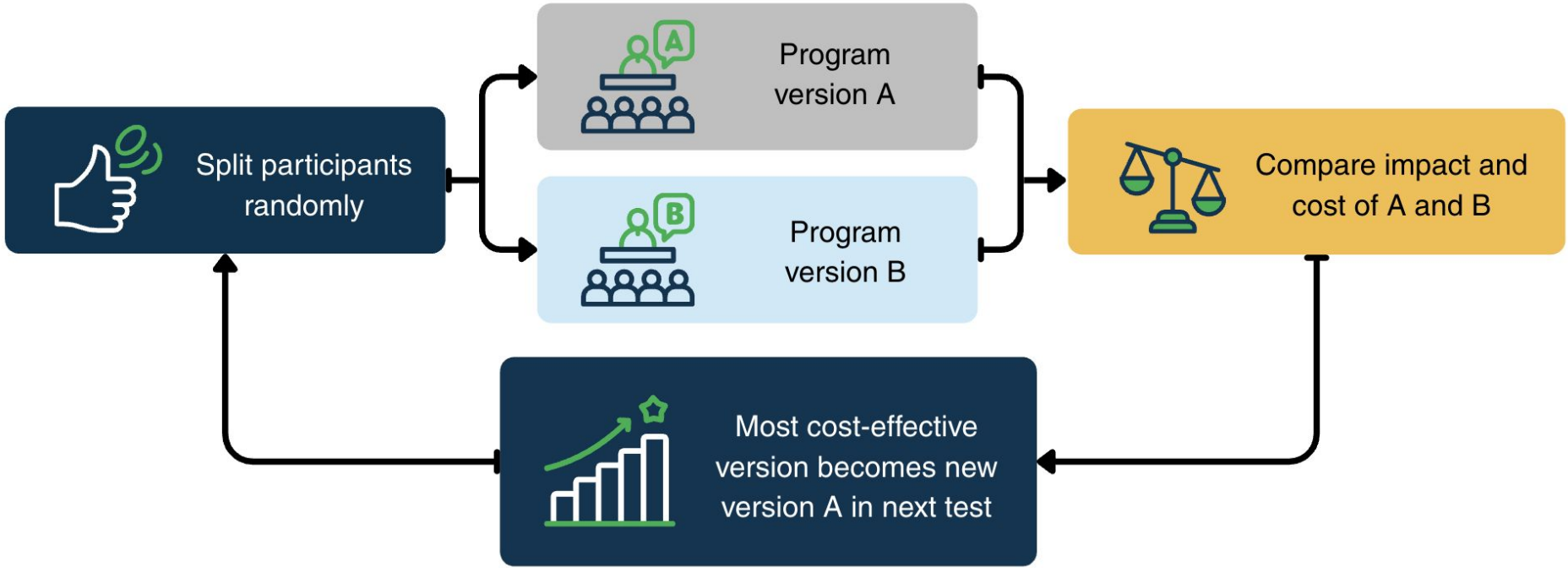
Rapid

Decision-informed results produced quickly

Regular

A/B testing becomes part of routine M&E

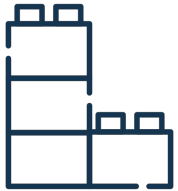
Randomize individuals or groups to version A or B, choose most cost-effective version, repeat



Two types of tests: cost-reducing & effectiveness-enhancing

Cost-reducing examples

- Does virtual training work as well as in-person training?
- Can we swap a complicated scheduling system for an “on-demand” program service?

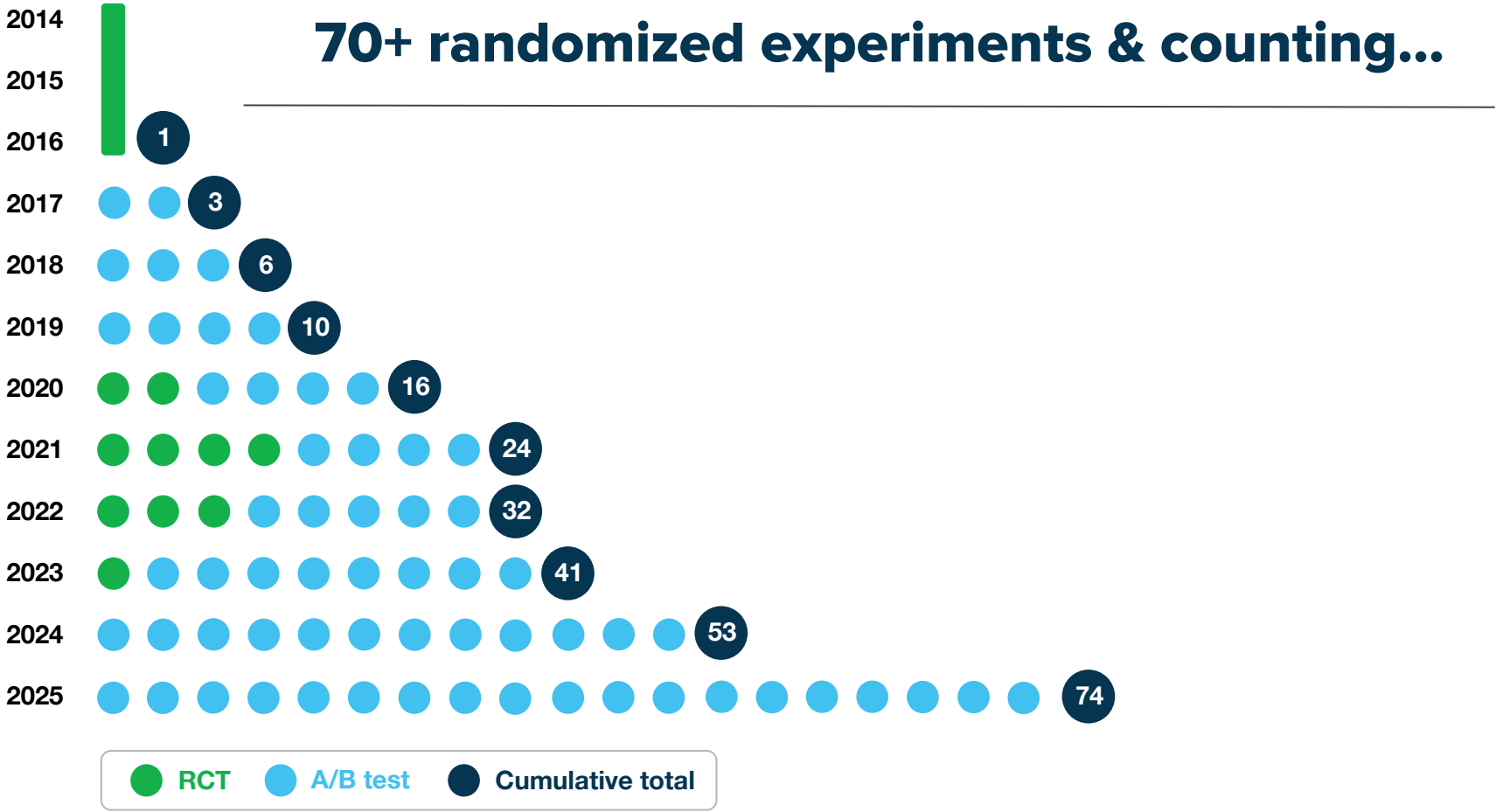


Effectiveness-enhancing test examples

- Does more frequent small group work in class improve learning?
- Does adding caregiver engagement improve learning?

A/B testing at Youth Impact

70+ randomized experiments & counting...



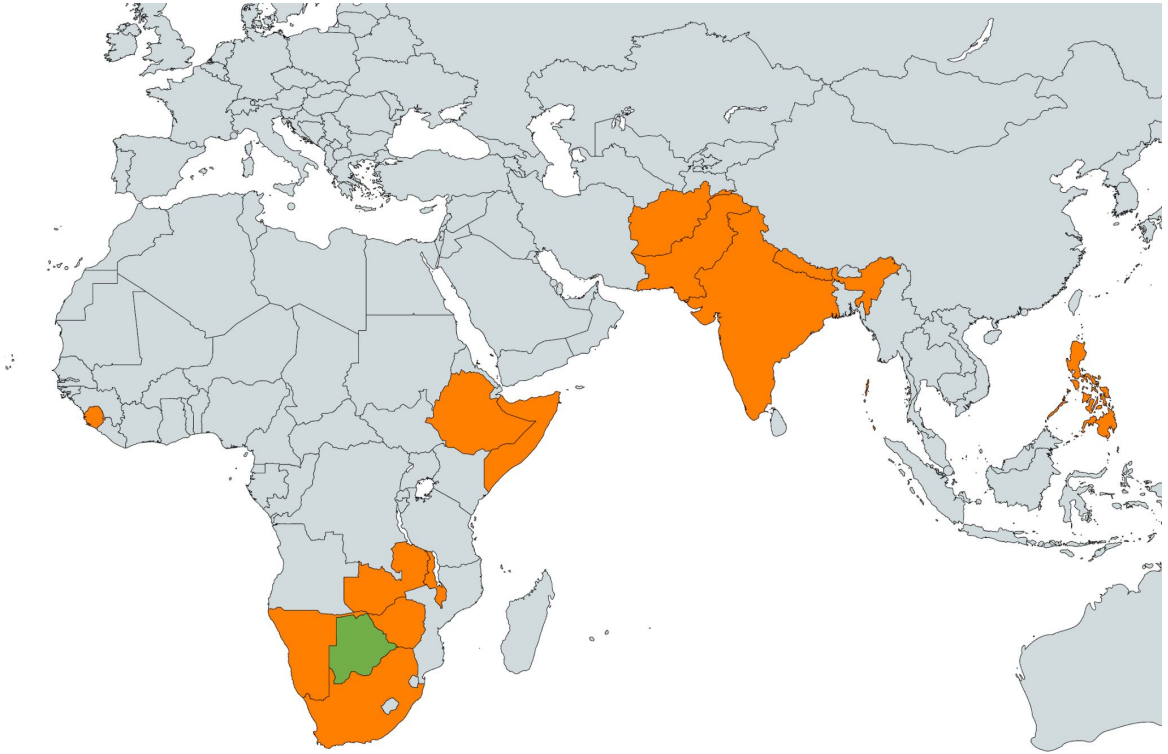
ConnectEd is an at-home FLN program

Tutor calls caregiver and child to deliver weekly lesson

ConnectEd sends caregiver practice problems weekly to complete with student

RCT established efficacy → A/B testing enables scaling across & within countries

- Implemented RCTs in 7 countries with partners
- A/B testing in 5+ countries implementing ConnectEd



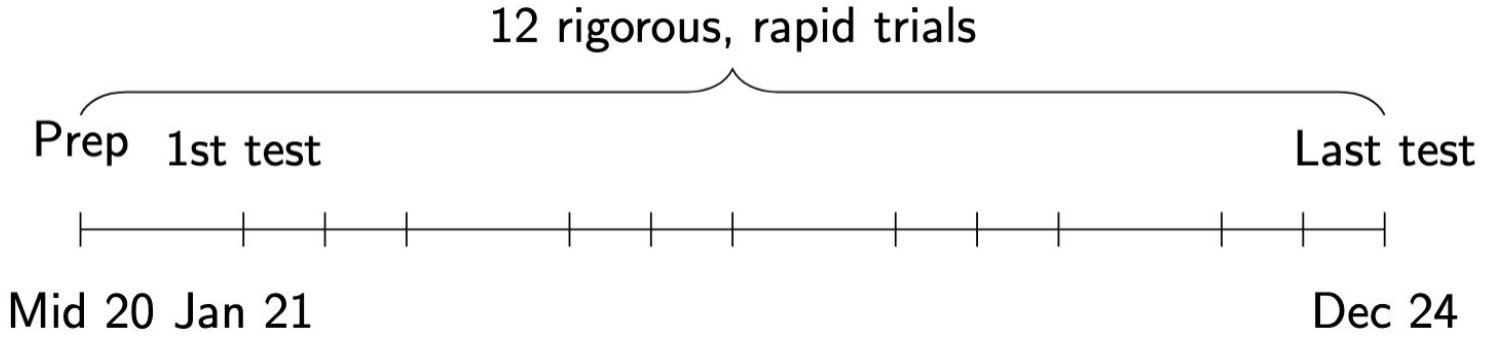
A recent WWH working paper focuses on results of 12 ConnectEd A/B tests

WORKING PAPER

20 October 2025

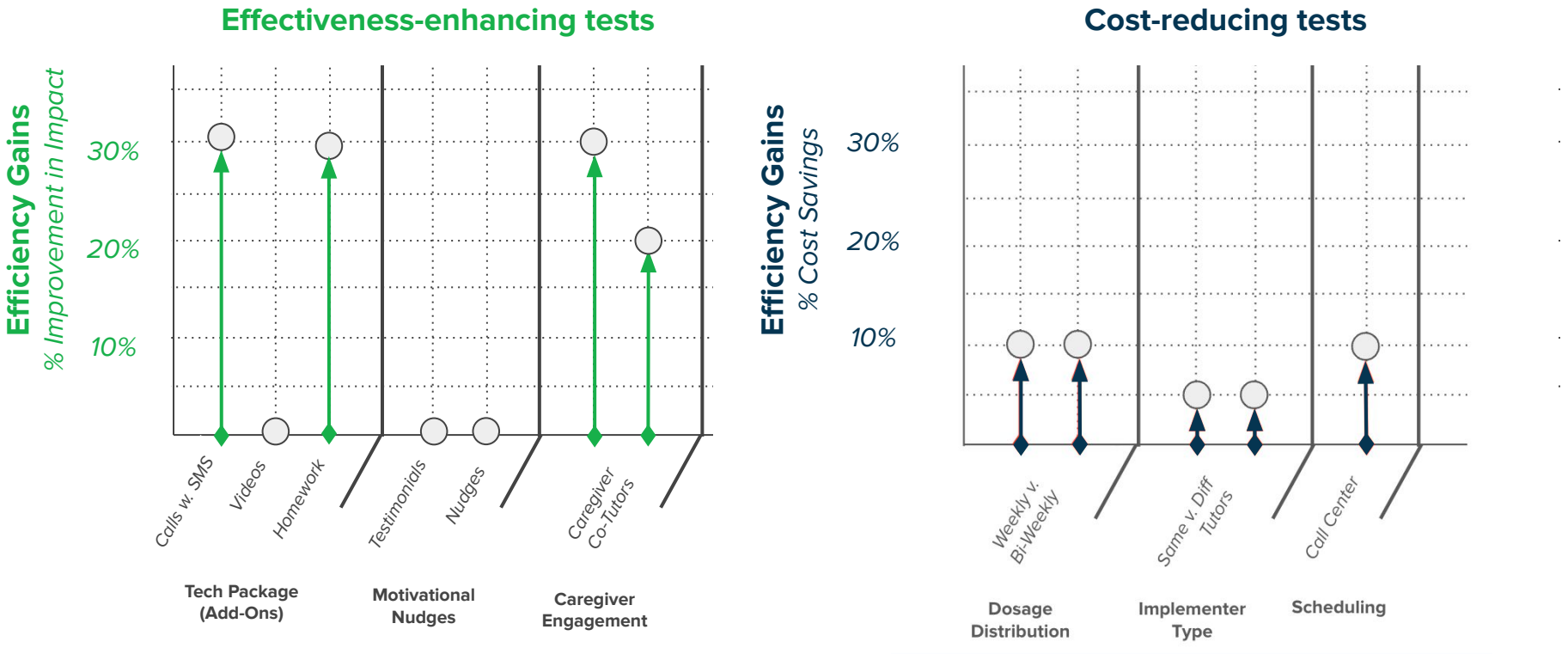
Cheaper (and more effective) by the dozen: Evidence from 12 randomised A/B tests optimising tutoring for scale

Authors:
Noam Angrist, Claire Cullen and Janica Magat



They looked at efficiency gains across two types of tests

Up to 30% improvement in learning outcomes and 11% reduction in program costs



Operationalizing A/B testing

Steps to ramp up to ongoing A/B testing

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

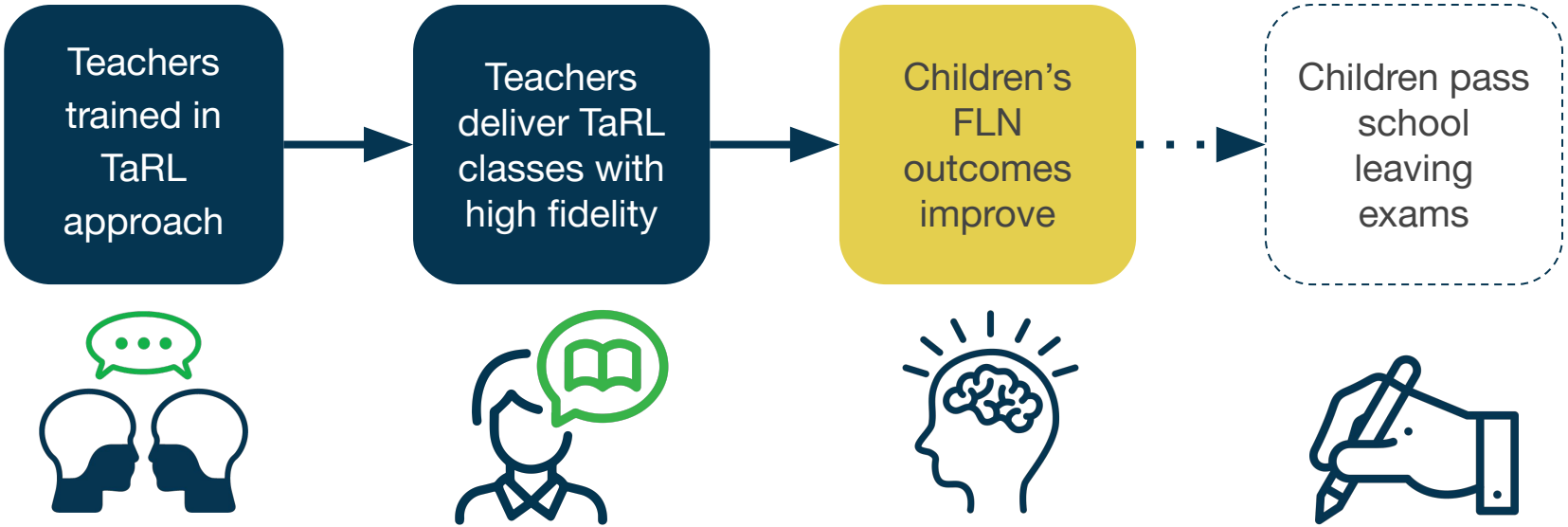
Successfully execute a first A/B test. Adapt the program based on results. Plan for the next test.

4

Ongoing A/B testing

Set up a continuous process of iterating and testing.

Step 2: Golden indicator outcomes change within weeks or months



Step 2: Shortlisting tool helps narrow down a golden indicator

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?
TaRL	Student literacy and numeracy	The share of students who are numerate or literate at endline	# students at division level (scoring 4 on 0-4 scale) / # of students assessed	Outcome	Student	80	6	Yes

Step 3: Optimize for feasibility in a first test (tool)

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)
Do SMS with homework problems sent in between tutoring sessions improve learning?	Homework over SMS	Children receiving tutoring do not receive homework via SMS between tutoring sessions	Children receiving tutoring receive homework via SMS between tutoring sessions	Easy

Steps 3&4: Other important considerations for each test

Factors to operationalize the A/B test and measure success				
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)**
Share of children who are numerate (ie can divide)	Child	1200	Determine system for sending SMS problems to 600 students in between three tutoring sessions, so two SMS total	Med

Step 4: Consider other factors in ongoing testing (tool)

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)
Does using an improved weekly checkpoint (check for understanding) tool improve learning outcomes?	Teacher instructional tool	Teachers use older, less-structured checkpoint tool	Teachers use new, more-structured checkpoint tool	Med	Increase effectiveness	High	Med

A good A/B testing question is

- ✓ Feasible to implement
- ✓ High priority
- ✓ Effectiveness-enhancing or cost-reducing
- ✓ Sufficiently ambitious
- ✓ Implementer-driven
- ✓ Answering a question with an unknown answer

What does an organization need to do A/B testing?

Organizational attributes

- ✓ Learning culture
- ✓ Willingness to flip a coin
- ✓ Longer-term investment
- ✓ Tolerance for making decisions under uncertainty
- ✓ Internal data-savvy team

M&E system attributes

- ✓ Golden indicators to drive decision making
- ✓ Large-scale data collection
- ✓ High-frequency data collection