

March 2026



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

.....

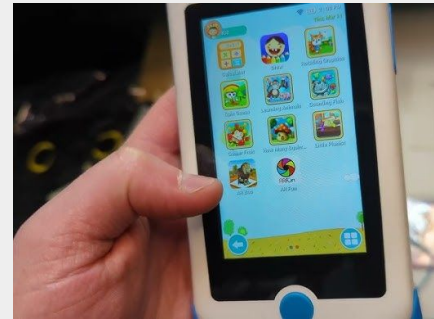
Andrés Parrado

Most interventions suffer a “voltage drop” at scale

Between 50-90% of programs experience a “voltage drop” when scaled, where the benefit-cost ratio shrinks when moving from a small-scale trial to larger, real-world implementation (List, A., 2023)

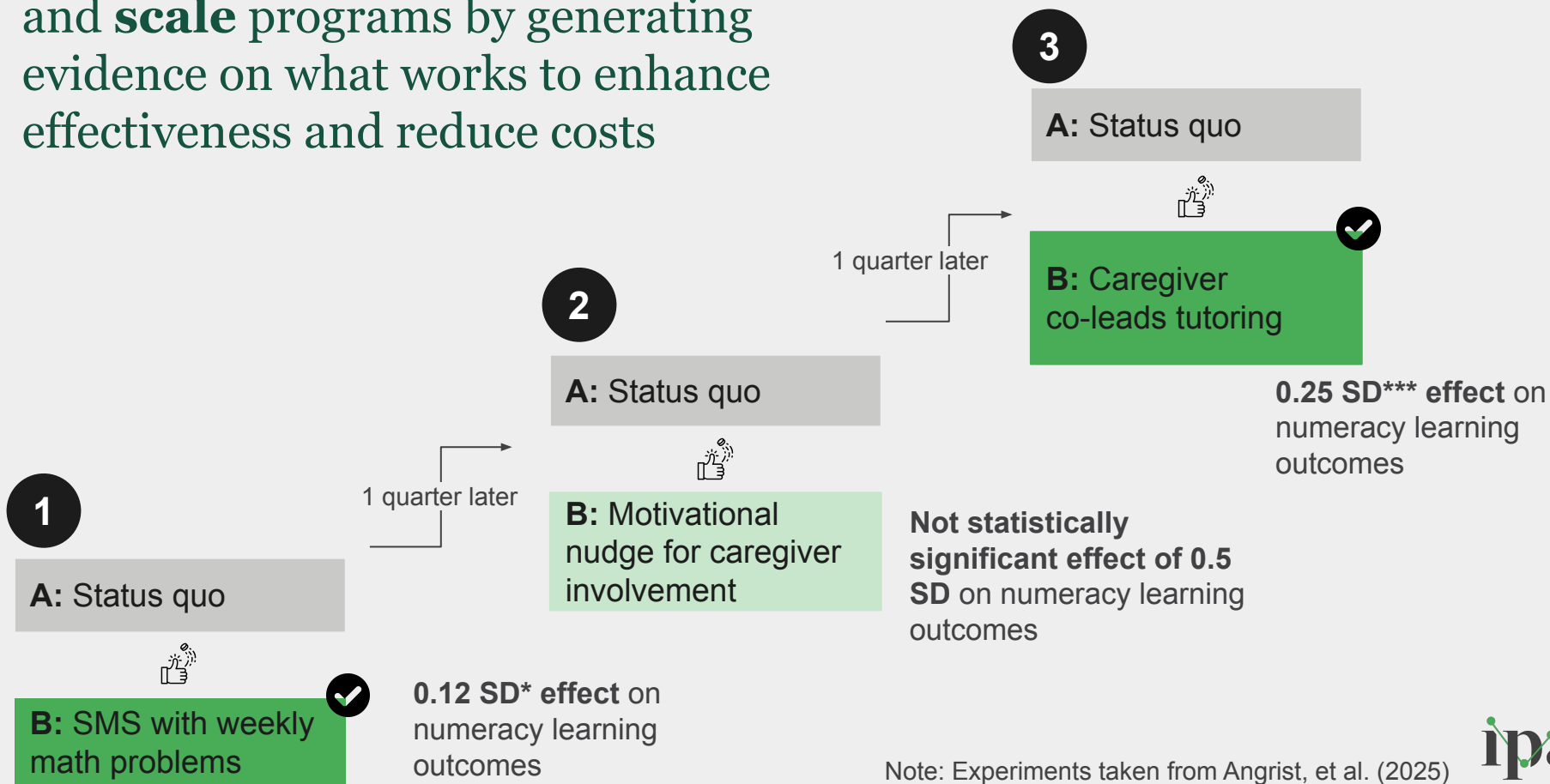


Costs increase at scale



Effectiveness decreases with scale

Iterative A/B testing can help **optimize** and **scale** programs by generating evidence on what works to enhance effectiveness and reduce costs



Note: Experiments taken from Angrist, et al. (2025)

Not every test will be a hit

JAMES BLUNT

GREATEST HITS

01. You're Beautiful
02. Heart To Heart
03. Stay The Night
04. 1973
05. Bonfire Heart
06. Wiseman
07. Carry You Home
08. SAME mISTAKE
09. Goodbye My Love
10. I'll Be Your Man
11. Bartender
12. Don't Give Me Those Eyes
13. Tears And Rain
14. Hight
15. Lose My Number
16. No Tears
17. Postcards
18. When I Find Love Again

??



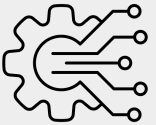
Maybe as few as 10% (cited in Angrist, et al. (2025))

A/B testing can help address scaling challenges


But, how to do it right?




Invest in tests that have a **high potential return** to enhance program cost-effectiveness



Develop the **capabilities and systems** for iterative A/B tests



How to design & prioritize high potential A/B tests?



1

Identify what do you want to enhance:
reach, efficacy, or cost

2

Identify the **metric** you will use to test

3

Identify and prioritize **evidence-based variations** to test out

① Identifying what to optimize: Look for opportunities to improve interventions' cost-effectiveness

$$\frac{\text{Program's reach} \times \text{Program's impact}}{\text{Program's costs}} = \text{Program's Cost-effectiveness}$$

The diagram illustrates the formula for Program's Cost-effectiveness. It features three grey boxes: 'Program's reach' and 'Program's impact' are positioned above a horizontal line, with a multiplication symbol 'X' between them. Below the line is a box labeled 'Program's costs'. To the right of the line is an equals sign '='. Further right is a dark green box containing the text 'Program's Cost-effectiveness' in white.

② Identify key metrics for A/B testing

Cost-effectiveness lever	Example metrics
Reach	<ul style="list-style-type: none">● # active users● # onboarded users
Impact	<u>Product KPIs</u> <ul style="list-style-type: none">● User engagement: Weekly frequency of usage
	<u>Theory of Change outcomes</u> <ul style="list-style-type: none">● Knowledge: Changes in foundational numeracy skills● Attitudes/perceptions: Changes in teacher's perception of the importance of addressing learning variability in the classroom● Behaviors: Changes in teacher practices
Cost	<ul style="list-style-type: none">● Unit cost of program delivery

3

Identify and prioritize variations to test

Collect data to inform what to test

Literature review

User research

Interviews with experts/field staff

Define variation options

1: Send nudging reminders

2: Call students daily

3: Create a “streak” feature

4: Written commitment

Prioritize variations and sequence them

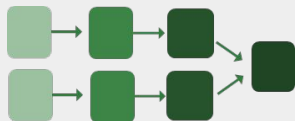
- **Relevance:** How likely is this variation to improve the key outcome metric?
- **Implementation feasibility:** How easy and/or costly will it be to implement this variation?

At the end you should have a road map of possible A/B tests

Illustrative example: Designing a set of A/B tests for Shaia (teacher support chatbot)

Mentimeter decides they want to test things that could enhance the:

Program's impact



Mentimeter identified the use of the learning variability feature as the key metric.

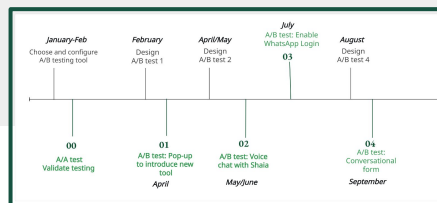
Shaia
FOR MENTIMETER

Short on time to plan your lessons?

Do it fast and easy with Shaia.

Use Shaia for free

Articulated a series of A/B tests into a roadmap



Superpoderes Shaia

Si quieres aprovechar al máximo mis superpoderes, **crea un perfil de grupo para tus estudiantes**. Así, recordaré sus características y te haré recomendaciones más profundas y personalizadas.

Crear grupo

Actividad
Crea una actividad. Mejora la participación y el aprendizaje de tus estudiantes.

Dinámicas de inicio
Crea dinámicas para romper el hielo y activar, calentar, enfocar y preparar a tus estudiantes.

Abp
Conecta el aprendizaje con el mundo real. Descubre el poder del Aprendizaje Basado en Proyectos.

What capabilities and systems are needed to run an A/B test?



1

Define sample size

2

Set-up the necessary data systems & tools

3

Analyze results and make decisions

Key enablers for organizations to adopt and implement A/B testing

Sample size considerations

- **A stable and sufficiently large user base**, to support rigorous comparisons.

Systems and tools

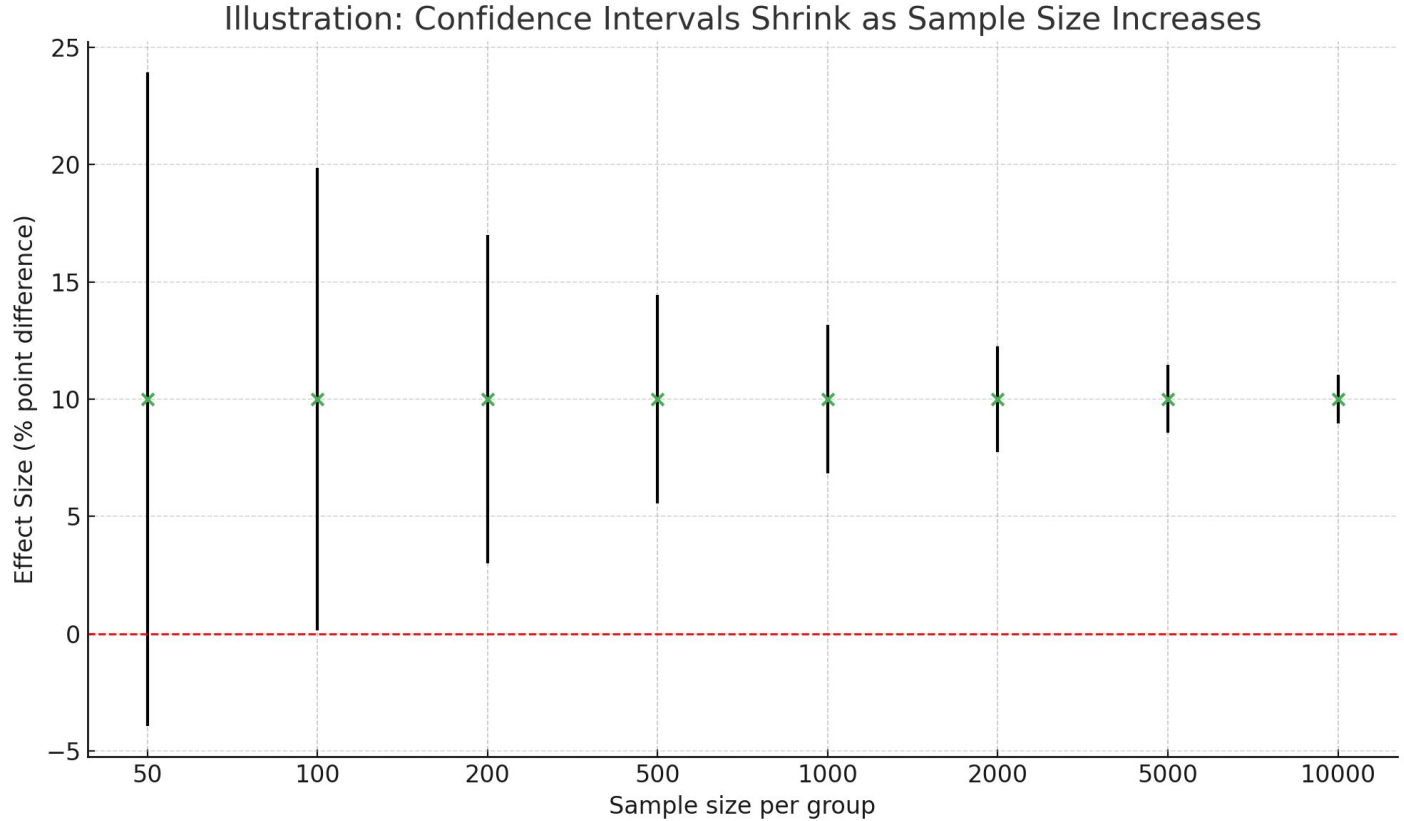
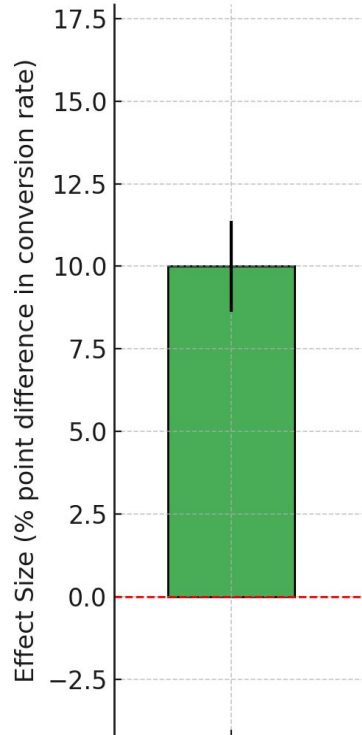
- **Reliable data systems** to track relevant outcomes.
- **Experimentation tools**, that allow for random assignment and multiple test management.

Analytical capacity

- **Strong data management capacity**, to connect data infrastructure with experimentation tools.
- **Statistical and analytical capacity** to design valid experiments and accurately interpret results.

1

Key concept: Larger sample size reduce uncertainty of results



2

What systems & tools are needed for A/B testing?

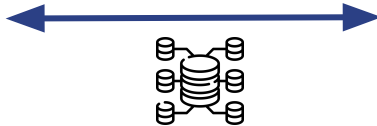
The tool selection depends on the nature of the program/ product



Experimentation tool

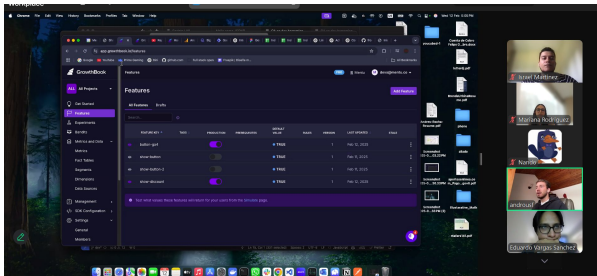
- Experiment setup: segment target sample and randomly assign users
- Deploy variations

Data Infrastructure should allow for the integration between these systems



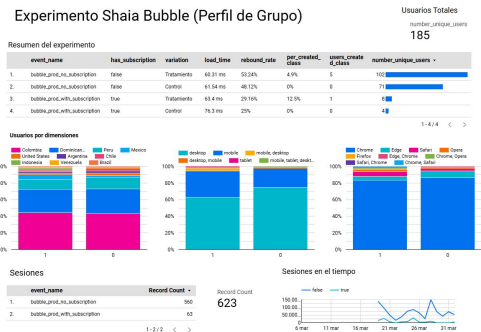
Data Analytics tool

- Capture and visualize key metrics
- Statistically analyze results



GrowthBook (open source)

Other tools specifically for the social sector are **Evidential** and **UpGrade** (education)



Looker Studio

3

Analyzing A/B tests results and making decisions

Assess the credibility of the experiment:

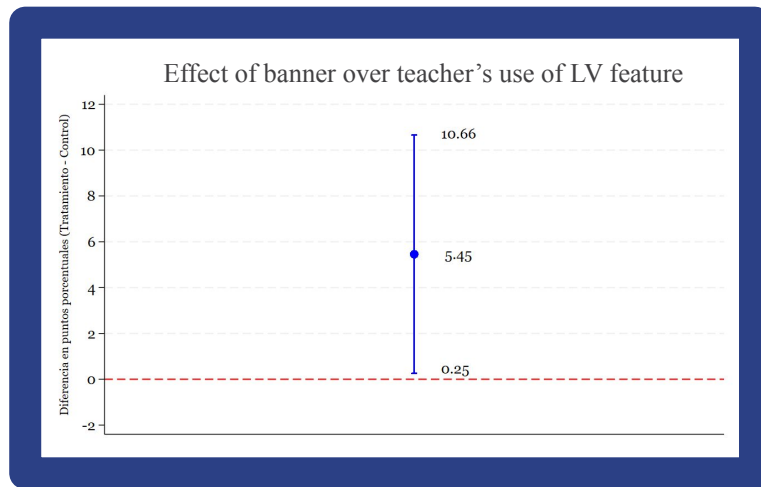
- **Sample size** necessary to detect a meaningful effect
- **Groups are comparable** on key characteristics

Assess the effectiveness of the variation:

- Is the difference between groups **statistically significant**?
- Is the difference large enough to matter?

Large effect with non statistically significant results?

Assess the level of uncertainty (CI) in the results



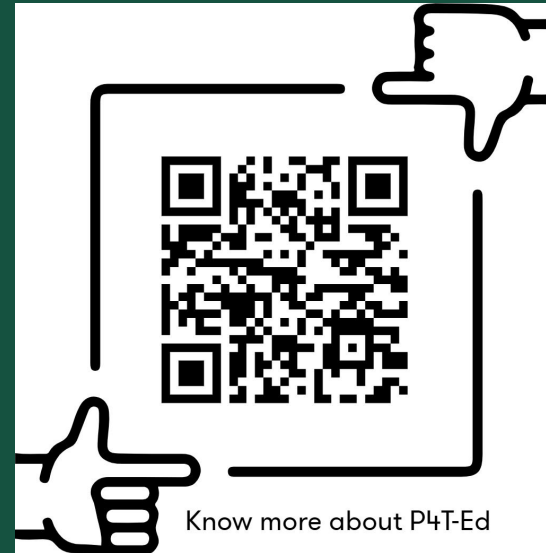
Interpret the results and make a decision

Inconclusive or null: repeat the experiment or discard variation

Positive effect: scale the variation to all users or continue testing to find a more effective variation

Negative effect: discard the variation and use findings to inform future iterations

Happy testing!



Thank you!

Contact us:

**Andrés Parrado, Associate Director
Right-fit Evidence Unit
aparrado@poverty-action.org**

poverty-action.org