What Works Hub for Global Education

A/B testing in education: rapid experimentation to optimise programme cost-effectiveness

Authors: Noam Angrist, Amanda Beatty, Claire Cullen and Moitshepi Matsheng

Youth Impact
Blavatnik School of Government, University of Oxford
UK International Development
Partnership | Progress | Prosperity

A/B testing has become a common approach for programme optimisation in the technology sector (Kohavi et al., 2020; Siroker and Koomen 2015). Companies like Google, Amazon, and Microsoft run thousands of micro experiments monthly, continuously optimising their products. In an A/B test, customers are randomly allocated to versions A or B of a product, for example, a website interface, and clicks or profits determine which version becomes the standard operating model. This culture of rapid, ongoing experimentation has produced striking results. One study found that firms that adopt A/B testing see a 30 to 100% improvement in performance after a year (Koning et. al., 2022). A/B testing helped Bing increase revenue by 10% to 25% each year (Kohavi and Thomke, 2017).

In the social sector, while the number of rigorous programme evaluations informing public policy has exponentially increased over the last two decades (Kaufman et al., 2022), still far too few programmes are evaluated rigorously or scaled successfully (List 2024; List 2022; Mobarak 2022). A review by the British National Audit Office found that only 8% of major government projects are robustly evaluated (National Audit Office, 2021). Moreover, less than 15% of studies measure cost – critical for evaluating social returns on investment (Brown and Tanner, 2019). These striking gaps reveal the need for more evidence, especially approaches that can inform cost-effectiveness and that can be embedded within policy and programme implementation. Implementers could use A/B testing to frequently assess cost-effectiveness and scalability of a programme, comparing the standard model, option A, to an optimised version, option B. [1]A/B testing shares multiple features with randomised controlled trial evaluations but is often more nimble, iterative, and embedded in government and nonprofit implementation processes.

We illustrate the benefits of A/B testing in the context of education programming in low- and middle-income countries, grounded in Youth Impact's experience implementing A/B testing over the last seven years.[2] We compare similarities and differences between A/B testing and Randomised Controlled Trials (RCTs), a complementary rigorous evaluation approach. We share how our organisation arrived at A/B testing and how we use it, with the aim of informing how other implementers can use A/B testing to deepen their impact, reduce costs, and scale programs.

[1] In the past, we used the term Rapid Impact Assessment (RIA) as synonymous with A/B testing. We have since adopted the term A/B testing for consistency. Approaches referred to as "adaptive testing" are also often considered in the family of A/B testing (e.g. Kasy and Sautmann, 2021; Athey et al., 2023).

[2] Multiple funders have been critical in supporting our A/B testing system, starting with the Mulago Foundation, and later including the Jacobs Foundation, Agency Fund, Prevail Fund, and most recently the What Works Hub for Global Education. Funding incentives to engage in long-run and iterative learning are critical to enable A/B testing to take off in the sector.

Key features of A/B testing: rigorous, rapid, regular

A/B testing and RCTs both fill an important gap in maximising social returns on investment, with a focus on generating causal evidence to evaluate whether programmes and policies had the desired impact. RCTs randomly assign individuals or groups to a programme (treatment) or no program (control). As an example, a treatment group of schools or classrooms might, through random assignment, receive a remedial education programme and the control group would receive business-as-usual schooling. By virtue of randomisation, the groups are equal on average, except for whether they received the remedial programme. As a result, any change in outcomes observed between the groups can be confidently attributed to the programme; that is, one can calculate the causal effect of the programme.

In A/B testing, there is also random assignment between groups, except rather than include a pure control group, multiple versions of a programme are compared: version A vs. B. For example, group A would receive the remedial education programme and group B would receive the same programme but with more-intensive mentoring for teachers, and success would be measured by comparing student learning outcomes for groups A vs. B.[3] Similar to RCTs, random assignment ensures equal groups, so any difference in learning outcome reflects the causal impact of the program optimisation.

Table 1 below illustrates key principles of A/B tests -– the "3Rs" – in comparison with RCTs, including both differences and similarities. A/B tests are rigorous. Like RCTs, they use randomisation to generate causal evidence. While RCTs typically aim to answer the question "does the programme work" with an external evaluator and a long-run lens, A/B tests are typically focused on internal and immediate programme decision-making and aim to answer the question "how does the programme work most effectively, cheaply, and scalably." A/B tests are rapid. Usually A/B tests last weeks or months and results feed back into immediate programme decision-making, while RCTs may last years. A/B testing is regular or routine. A/B testing often uses existing organisational monitoring and evaluation (M&E) data and is an integral part of an organisation's M&E system. For example, if an organisation implements foundational literacy and numeracy (FLN) programming and collects learning data every six to eight weeks over a school term, it conducts A/B tests over that same timeframe and could repeat testing each term.

[3] Implementers can also run A/B tests with multiple treatment arms, such as A/B/C. A/B testing is often conducted at scale and with large sample sizes, to facilitate rapid learning and to ensure enough statistical power to detect differences between groups.

This regularity allows for real-time optimisation, enabling adoption of ever-more cost-effective programme implementation models each school term. Importantly, it often takes multiple tries to identify an effective optimisation, with multiple rounds of A/B tests yielding null results, and one of every few yielding substantial returns. This pattern follows the innovation literature where a few wins generate up to 17:1 social returns across a portfolio of innovation investments (Kremer et al., 2021).

Table 1: The 3Rs – Key principles and attributes of A/B tests

TYPICAL A/B TEST ATTRIBUTE	COMPARISON TO TYPICAL RCT ATTRIBUTE
Rigorous Randomised; results capture causal impacts. Multiple groups receive various optimised treatments to test "how the programme works most effectively, cheaply, and scalably".	Randomised; results capture causal impacts. Often the main comparison is a no-programme control group to test the overall question "does the programme work?"
Rapid Results reported in weeks or months using short and mid-term indicators to inform real-time decisions.	Results reported over years using longer-term outcomes.
Regular Built into regular and existing organisational M&E systems to directly inform programme implementation and operations; multiple related tests in rapid succession to optimise cost-effectiveness.	Often a once-off high-stakes study testing novel ideas and involving external data collection.

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

RCTs and A/B tests can be used in sequence to generate complementary evidence. A use case for A/B testing before an RCT is when a programme is in the pilot stages and an organisation wants to test out what programme version it would subject to a high-stakes RCT. For example, an A/B test could be used to determine how to best promote take-up or enrolment in a tutoring programme before subjecting the programme to an RCT to determine its impact. A/B testing can also be used after an RCT. For example, once there is proof of concept that the programme is effective, A/B testing is useful for testing whether the programme will work in a new context, with different implementers (e.g., government teachers vs. volunteers), or if the program can be made more cost-effective.

Youth Impact's arrival at A/B testing

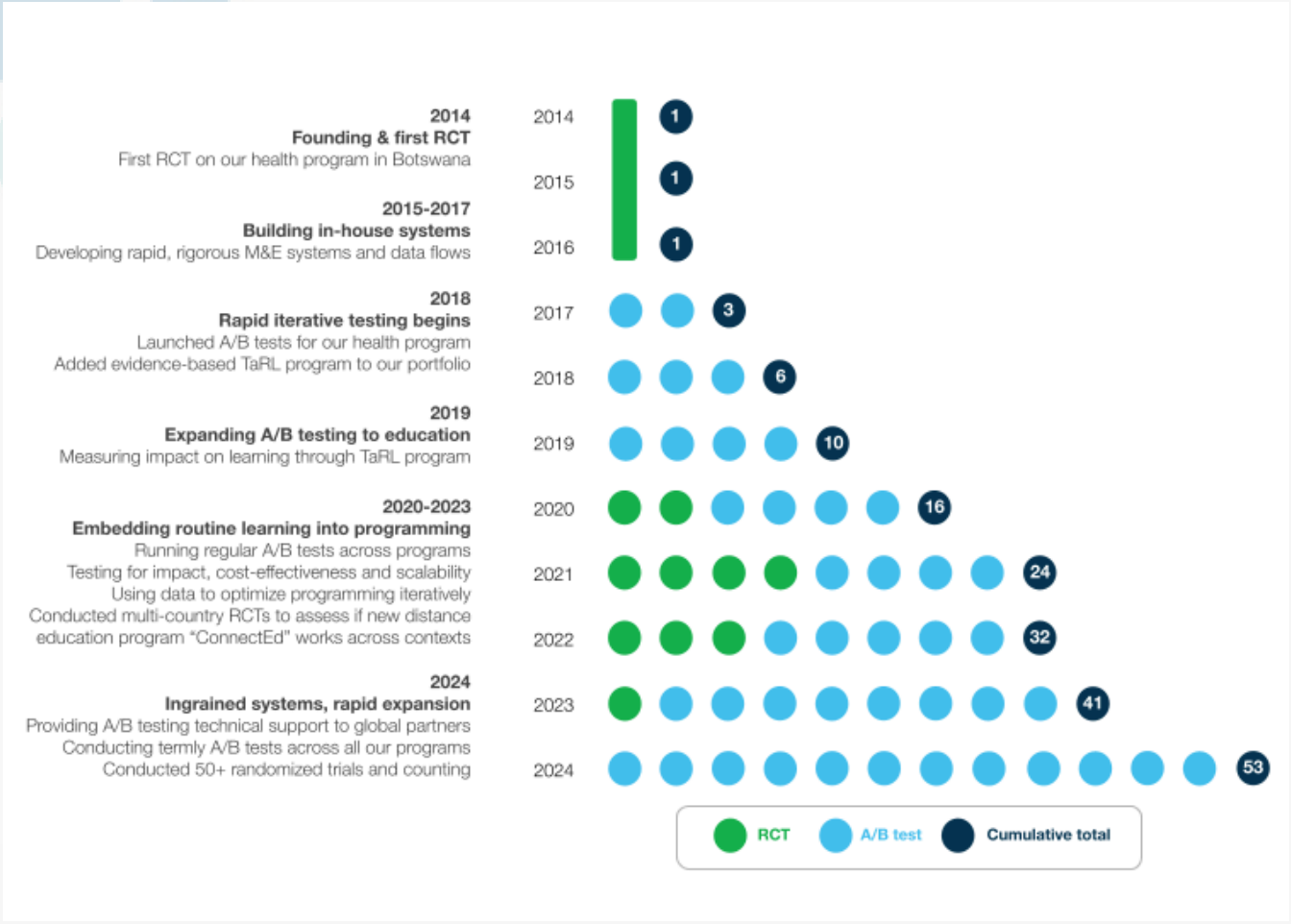
When Youth Impact was founded in 2014, we launched a large-scale RCT across a third of Botswana to test the effectiveness of a sex education programme that had demonstrated impacts in Kenya (Dupas, 2011). A few years later, results emerged, and we found mixed results. The messenger mattered: the programme worked when delivered by near-peer educators (young and aspirational figures), but not when delivered by teachers.

The RCT generated important insights, but it had taken over a decade between the initial Kenya RCT to producing results from the Botswana RCT. As we considered next steps for the programme, we knew we wanted to tweak programme components to improve impact and scalability. We wanted to test these changes rigorously, but we wanted to do it quickly, at low cost, and to be able to iterate and optimise over time.

Over the next few years, we built and repurposed our M&E system to be A/B testing ready. While two RCTs had taken over ten years, by 2019 we had conducted 10 A/B tests in the span of just 10 months. We next expanded A/B testing to an evidence-based education programme we had started to scale up in Botswana, adapted from India, called Teaching at the Right Level (TaRL). When COVID-19 struck, powered by this rapid learning capability, we quickly adapted TaRL principles to create a distance education programme called ConnectEd. We launched the world's first RCT on distance education, and months later results showed sizable learning gains from ConnectEd (Angrist et al., 2022). These results catalysed five additional RCTs to test external validity and replicability of our original results across multiple settings and with governments (Angrist et al., 2023). With substantial RCT evidence generated on programme effectiveness across settings and delivery models, we next ran a series of A/B tests to continuously improve the program's cost-effectiveness on the path to scale.

Since 2019, we have run over 50 randomised evaluations including both RCTs and A/B tests; we now run an A/B test on each of our health and education programs every school term. In the first six months of 2024 alone, we ran 12 in-house A/B tests across all our programmes and countries of operation.

Figure 1: Timeline of Youth Impact A/B testing journey



An example: Optimising Teaching at the Right Level in Botswana

TaRL is a remedial educational programme that focuses on grouping children by learning level rather than by age or grade to improve foundational literacy and numeracy skills. Evidence from multiple randomised controlled trials conducted across countries have found the programme to be highly effective at improving student learning outcomes (Banerjee et al., 2007; Banerjee et al., 2017; Duflo et al., 2024). Since adopting and scaling TaRL in Botswana with the government, we have expanded delivery to multiple countries and employed A/B testing to optimise the programme's cost-effectiveness at scale. We are supporting the government in reaching all schools in Botswana by 2026.

A recent review showed that the returns to improving the implementation take-up and fidelity of proven programmes, such as TaRL, are 5-10x higher than identifying the next effective program (Angrist and Meager 2023). We ran an A/B test aimed at improving exactly this margin, program fidelity, defined as targeted instruction to each student's learning level. We randomised different approaches to most efficiently target instruction, grouping students by operation level or the ability to recognise digits. The standard implementation of TaRL (Option A) grouped students according to their understanding of operations. The new treatment (Option B) involved additionally subgrouping students according to their digit recognition level. This additional level of subgrouping in B represents even greater fidelity to the targeted instruction approach.

We found over one school term that this additional targeted instruction 'tweak' to our standard model (group B) caused a 0.2 standard deviation increase in learning (Angrist and Meager 2023). The marginal cost of this optimisation is small, estimated at just a few cents. This example reinforces the potential of A/B testing to optimise programme implementation in the context of a scale-up, with substantial improvements to fidelity and cost-effectiveness.

We have similar examples of improvements in cost-effectiveness across our programmes, with many of the best A/B testing ideas coming from field implementation teams who can often best identify implementation improvement opportunities.

Conclusion: a growing A/B testing movement

Our experience with A/B testing in education has shown that rapid, iterative testing can significantly enhance programme cost-effectiveness and scalability. A few key A/B testing principles can be captured in the "3Rs" — rigorous evaluation through randomisation, rapid results to inform real-time decisions, and regular iteration to optimise programmes over time. In our TaRL example, we demonstrated lessons from a single A/B test; in future work we will showcase how multiple A/B tests can come together to inform a broader programme optimisation process and learning agenda.

A/B testing has been integral to our growth trajectory at Youth Impact and has become central to our everyday M&E practice. Every school term brings a new opportunity for innovation and improvement. Program modification ideas come from implementation teams who use field observation to elevate ideas that have promise and we then test these innovations rigorously. We test strategies to deepen impact, lower cost, and scale across contexts and delivery models.

While there are only a handful of organisations that implement A/B testing regularly at present, the movement is growing. We have started to directly support nearly a dozen organisations to integrate A/B testing into their M&E practice; in further work we will share lessons learned and examples, with the aim of facilitating greater uptake of A/B testing approaches in the social sector. Many international development organisations, such as USAID's Office of the Chief Economist, the Foreign, Commonwealth & Development Office (FCDO), the Global Education Evidence Advisory Panel, the What Works Hub for Global Education, and the effective altruism movement, among others, are promoting evidence-based decision-making and a greater focus on cost-effective and scalable interventions. Integrating A/B testing into the programme evaluation toolkit of researchers, policymakers, and implementers offers a promising path forward.

References

Angrist, N., Bergman, P., & Matsheng, M. (2022). Experimental evidence on learning using low-tech when school is out. Nature human behaviour, 6(7), 941-950.

Angrist, N., & Meager, R. (2023). Implementation matters: Generalizing treatment effects in education. Blavatnik School of Government, University of Oxford.

Angrist, Noam, Micheal Ainomugisha, Sai Pramod Bathena, Peter Bergman, Colin Crossley, Claire Cullen, Thato Letsomo et al. Building resilient education systems: Evidence from large-scale randomized trials in five countries. No. w31208. National Bureau of Economic Research, 2023.

Athey, S., Bergstrom, K., Hadad, V., Jamison, J. C., Özler, B., Parisotto, L., & Sama, J. D. (2023). Can personalized digital counseling improve consumer search for modern contraceptive methods?. Science Advances, 9(40), eadg4420.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. (2007). "Remedying education: Evidence from two randomized experiments in India." The Quarterly Journal of Economics 122, no. 3: 1235-1264.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. Journal of Economic Perspectives, 31(4), 73-102.

Brown, E. D., & Tanner, J. C. (2019). Integrating value for money and impact evaluations: Issues, institutions, and opportunities.

Duflo, Annie, Jessica Kiessel, and Adrienne M. Lucas. "Experimental Evidence on Four Policies to Increase Learning at Scale." The Economic Journal 134, no. 661 (2024): 1985-2008.

Dupas, P. (2011). Do teenagers respond to HIV risk information? Evidence from a field experiment in Kenya. American Economic Journal: Applied Economics, 3(1), 1-34.

References

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

Kohavi, R., & Thomke, S. (2017). The surprising power of online experiments. Harvard business review, 95(5), 74-82.

Koning, Rembrand, Sharique Hasan, and Aaron Chatterji. (2022). "Experimentation and start-up performance: Evidence from A/B testing." Management Science 68, no. 9 : 6434-6453.

Kremer, M., Gallant, S., Rostapshova, O., & Thomas, M. (2021). Is Development Economics a Good Investment? Evidence on scaling rate and social returns from USAID's innovation fund. Harvard University.

List, J. A. (2022). The voltage effect: How to make good ideas great and great ideas scale. Crown Currency.

List, J. A. (2024). Optimally generate policy-based evidence before scaling. Nature, 626(7999), 491-499.

Mobarak, A. M. (2022). Assessing social aid: the scale-up process needs evidence, too. Nature, 609(7929), 892-894.

National Audit Office. (2021, December 2). Evaluating government spending.

Siroker, D., & Koomen, P. (2015). A/B testing: The most powerful way to turn clicks into customers. John Wiley & Sons.