Identifying Scalable Models: Aggregating Evidence Across 8 Rounds of a Numeracy Intervention

Noam Angrist, University of Oxford
Claire Cullen, Youth Impact
Janica Magat, Youth Impact

Abstract

We present results from a meta-analysis of 8 successive A/B tests conducted on a phone-based math tutoring program on the path to scale. After multiple rounds of testing, a cheap program modification improved cost-effectiveness by 30%, revealing the returns to iterative testing. Meta-analysis across rounds can help to synthesize the evidence, identifying typologies of innovations that yield the biggest cost-effectiveness gains, and contributing new evidence on programs' cost-effectiveness at large scale.

Extended Abstract

Introduction

There is ample randomized controlled trial (RCT) evidence demonstrating the efficacy of education programs when they are tested at small scale. However, as these programs are expanded and taken to scale, there are gains to iteratively and rigorously testing program modifications using A/B testing. This iterative learning process can prevent a 'voltage drop' as programs are taken from small-scale pilots to large-scale policies; they can also generate substantial improvements in cost-effectiveness.

Given A/B testing is typically conducted rapidly for iterative optimisation, it can generate many insights over a short period of time. Program modifications may also span a wide array of innovations, offering potential improvements in cost-effectiveness from cutting expensive budget line-items to adding new, potentially high-impact features. As such, evidence aggregation methods such as meta-analysis can be a useful tool to synthesize A/B testing evidence. Meta-aggregation can help policymakers identify typologies of innovations that yield the biggest cost-effectiveness gains, provide insight on the 'innovation trajectory', as well as contribute new evidence on programs' cost-effectiveness at large scale.

Background

In 2020-2022, researchers ran efficacy-focused RCTs with 18,000+ students across six countries to test the impact of a phone-based math tutoring program in Botswana, India, Nepal, Philippines, Kenya and Uganda (Angrist et al. 2022; Angrist et al., 2023). The program produced

large positive effects on learning across all countries, with average effects of 0.30-0.35 standard deviations. These effects were highly cost-effective, delivering up to four years of high-quality instruction per $100 spent, ranking in the top percentile of education programs and policies. With growing efforts to scale up the successful pilots, there were open questions about the program's cost-effectiveness as it grew beyond the initial RCT pilot conditions.

Design

To address efficacy-to-effectiveness questions and to improve scalability and cost-effectiveness during the program's scale-up, we employed the approach of A/B testing. We conducted rapid, iterative randomized tests each school term to modify and assess different adaptations of the phone-based tutoring program.

A/B testing is an experimental methodology traditionally used by the technology sector to compare the impact of different variants of a product (Kohavi et al., 2020; King et al., 2017; Siroker and Koomen 2015). A/B testing is procedurally similar to a RCT in that participants are randomly assigned to different groups. While many RCTs compare a "no program" pure control group to an intervention group, A/B tests typically allocate two treatment groups – A and B – to the same main base program, and introduce a small program variation in one group. Additional distinctions between RCTs and A/B tests include running multiple tests in rapid succession rather than one high-stakes study every several years. These rapid tests help answer the question "what works most optimally" – often the most relevant question to scale – not just "does the program work." This approach relates to the economic literature on adaptive experimentation in the social sciences (Kasy & Sautmann, 2021; Athey et al., 2021), and evidence-based decision-making (Abadie et al., 2023).

During each school term, the A/B tests for the phone-based tutoring program involved randomly assigning students to one of two program versions: the status quo model, or the modified version with the potentially more cost-effective or scalable adaptation. For example, in one test, both groups received math tutoring via phone calls, while only group B received additional support through complementary WhatsApp videos. After implementation, students in each arm underwent phone-based assessments, allowing for a comparison of learning outcomes between the two program versions.

Results

Given the rapid generation of evidence through A/B tests conducted every school term, we employed the aggregation method of meta-analysis to synthesize the results. Figure 1 shows results from a meta-regression of 8 A/B tests conducted successively on the program between 2021 and 2023. Learning effects are shown in standard deviations.

As may be expected in the innovation and learning process, many tested modifications do not generate a difference in learning outcomes between groups A and B. However, every few tests (roughly 1 in 4) yield substantial cost-effectiveness gains compared to the status quo program. These stand-out A/B tests included one where the modification arm encouraged caregivers to lead the second half of the phone tutoring call (which substantially boosted impact, raising cost-effectiveness by thirty percent). Another successful improvement in cost-effectiveness came from a test that found no difference in learning when implementing a much cheaper version of the program that was delivered for twice as long but half as frequently (reducing the frequency of costly call-scheduling). Overall, interventions that yielded the largest cost-effectiveness gains came from caregiver involvement and reducing costs.

Meta-analysis of these results reveal three key insights: Firstly, the program's initial large impact from the pilot RCTs can still be improved upon further as it is taken to scale. Secondly, meta-analyzing results by innovation type (e.g. dosage, caregiver engagement, etc.) can be useful for identifying patterns in successful education innovations from A/B testing. Lastly, it underscores that learning is not linear progression – while not every test yields substantial cost-effectiveness gains, those that do can lead to significant program improvements.

Figure 1: Meta-regression results of 8 successive A/B tests on student learning