# Learning Curve: Progress in the Replication Crisis

*By* NOAM ANGRIST, CLAIRE CULLEN, MICHEAL AINOMUGISHA, SAI PRAMOD BATHENA, PETER BERGMAN, COLIN CROSSLEY, THATO LETSOMO, MOITSHEPI MATSHENG, RENE MARLON PANTI, SHWETLENA SABARWAL, TIM SULLIVAN*

The science of scaling is an emerging field. Growing evidence reveals that replicating results across contexts, at scale, and with government systems remains difficult. These challenges persist in multiple disciplines, from psychology to economics.

A common finding in the scale literature is that effectiveness tapers as programs are adapted to new settings or are scaled up (List 2022; Mobarak 2022). A "replication crisis" in psychology highlights prominent cases where results from one study failed to replicate when repeated (Schooler 2014). In economics, for example, a contract teacher program in Kenya that was effective when delivered by an NGO failed when delivered by the government (Bold et al. 2018). Scale-ups of early childhood development (ECD) programs saw diminishing returns as the program grew from 70 children in Jamaica to 700,000 students in Peru (Araujo, Rubio-Codina, and Schady, 2022).

A frequent rationale for lower effectiveness across replications is difficulty in maintaining implementation fidelity (Banerjee et al. 2017). Others include contextual differences (Pritchett and Sandefur 2015) and selection bias, such as non-representative initial study partners and study sites (Allcott 2015).

Is it necessarily the case that programs and policies experience diminishing returns as they are adapted across contexts, scaled up, and delivered through government systems?

In this paper, we investigate this question leveraging detailed monitoring data from a five-country randomized replication study of a phone tutoring program for disadvantaged students – one of the largest multi-country replication efforts in education to date. Randomized trials have proliferated, yet an analysis of top journals shows that fewer than 1 percent of recent randomized studies have been conducted across multiple countries, notwithstanding notable examples in education such as Teaching at the Right Level (Banerjee et al. 2017) and a few early grade reading interventions (Lucas et al. 2014).[1]

We report implementation fidelity results over time, across countries, and for both government and NGO implementation. These replication studies took place in India, Kenya, Nepal, Philippines, and Uganda and built on a proof-of-concept of phone-based tutoring in Botswana during covid-19, which improved learning outcomes by 0.12 standard deviations (Angrist, Bergman, and Matsheng, 2022).

*Angrist: Youth Impact, Oxford (e-mail: nangrist@youth-impact.org). We are grateful to a coalition of partners who enabled this multi-country response. Implementing and research partners include Youth Impact, J-PAL, Learning Collider, Oxford, World Bank, Ministry of Education, Science and Technology of Nepal, Teach for Nepal, Street Child, Department of Education in the Philippines, IPA, Building Tomorrow, NewGlobe, Alokit, and Global School Leaders. Funding partners include UBS Optimus Foundation, Mulago Foundation, Douglas B. Marshall Foundation, Echidna Giving, SNF Foundation, Jacobs Foundation, Peter Cundill Foundation, Northwestern's "economics of nonprofits" class, and J-PAL IGI. We thank Natasha Ahuja, Janica Magat, and Abraham Raju for great research assistance.

[1] Out of a set of 400 papers in development from 2019 and 2021 in a set of top economics journals, 19 percent were RCTs; of those that were RCTs about 1 percent of RCTs were multi-country studies. The set of journals includes the Top 5 economic journals (American Economic Review, Quarterly Journal of Economics, Econometrica, Journal of Political Economy, and Review of Economic Studies) and other top-tier general interest journals (Review of Economics and Statistics, Economic Journal, Journal of the European Economic Association, and all four American Economic Journal AEJ journals), and a top field journal (the Journal of Development Economics).

In contrast with the literature showing replications have diminishing returns, our results show consistently large and *improving* implementation fidelity. This holds true across replications and contexts, over time, and with governments as well as NGOs. This result reflects a well-known phenomenon rarely considered in the replication or scaling literatures: learning from experience. Learning from experience has been well documented at individual levels, such as human capital accumulation at work (Jedwab et al. 2021). At organizational levels, this insight has been modeled, but empirical study remains limited (Herriott, Levinthal, and March 1985).

Our results suggest that replication and scale-up are not intractable. Rather, equipped with mechanisms for learning from experience at an organization level, such as high frequency monitoring data and sharing lessons across replications, high implementation fidelity can be achieved. Our results reveal a substantial "learning curve" – implementation fidelity is enhanced across settings, over time, and with governments, facilitating effective replication and scale-up.

### *A. Proof-of-Concept Study*

An initial proof-of-concept study tested the effectiveness of phone-based tutoring during covid-19 school closures in Botswana from April to July 2020. The study was a randomized controlled trial with 4,500 households. The intervention consisted of weekly 20-minute phone calls with parents and their primary school children in grades 3 to 5 and covered foundational math content, such as addition, subtraction, multiplication, and division. These phone calls complemented weekly SMS messages which provided content and practice problems. The intervention lasted 8 weeks and the total dosage was 3 hours. The intervention and study were led by Youth Impact, one of the largest NGOs in Botswana, in partnership with the government, J-PAL, Learning Collider, and the University of Oxford.

The program was highly effective and provided some of the first experimental evidence during covid-19 of successful remote education. Results showed learning gains of 0.12 standard deviations. The program was cheap since it leveraged existing devices rather than providing new ones and required relatively minimal tutoring time. As a result, learning gains were equivalent to over a full year of high-quality instruction gained per $100, ranking among one of the most cost-effective approaches to improving learning in low- and middle-income countries.

One reason the approach is so cost-effective is that it harnesses a highly scalable technology – mobile phones – to reach students *en masse* and at low cost. This "low tech" phone-based approach enables wide reach even in low resource settings whereas higher tech approaches, such as internet-based ones, might not. While less than 15 percent of households have access to the internet in low-income countries, over 70 percent have access to mobile phones.

Another major reason for the effectiveness of the approach is highly targeted instruction. This approach stands in sharp contrast with the status quo where instruction is rarely targeted to student levels. Growing evidence shows that in many education systems in low- and middle-income countries there is a learning crisis, with students learning very little in school (Angrist et al. 2021). Students often do not meet grade-level expectations, yet teachers teach grade-level curricula regardless. For example, a teacher might teach fractions when students cannot recognize numbers. This means many students start behind grade-level and stay behind (Pritchett and Beatty 2015).

The weekly phone calls targeted instruction along a few dimensions. First, data was collected to assess children's learning level with a problem-of-the-day; instructors used this information to then target instruction. For example, a child who answered addition questions correctly would be taught subtraction

in the next lesson. A student responding incorrectly to an addition question would repeat an addition lesson in the following week.

This approach built on similar targeted approaches, such as Teaching at the Right Level (Banerjee et al. 2017). In addition, phone calls were conducted one-on-one, akin to tutoring interventions, another effective and highly targeted educational approach. These features worked together to provide instruction that was particularly well-targeted.[2]

### *B. Replication Studies*

The initial results from Botswana were released in August 2020. Shortly after, efforts to conduct multi-country randomized trial replications commenced. This advance from a proof-of-concept study to trials across contexts and evaluating scalable delivery models reflects progress from "efficacy" to "effectiveness" studies.[3] In each replication, the core model included weekly 20-minute phone calls delivered to primary school children and their caregivers over an 8-week period with highly targeted instruction covering basic math content.

In each setting mobile phone access was high. Sensitization calls were conducted in most studies to ensure mobile numbers worked and to obtain consent to participate, with high consent rates ranging from 70 to 90 percent. Slight context adaptations included translating content to local languages, taking into account baseline learning across contexts to ensure instruction was targeted in each setting, and optimizing teacher-student ratios depending on available hours teachers had in each setting. For full-time instructors of the program, typical ratios were 1:20 – close to common class ratios. Rather than teach all children in a group in non-targeted fashion as per a typical class, each child had targeted 20-minute sessions weekly.

In Kenya, the study launched in December 2020 through April 2021 with a large, low-cost private school network (NewGlobe). In Nepal, the study ran from January to July 2021 led by the World Bank, the government's Ministry of Education, Science and Technology, as well as two NGOs: Teach for Nepal and Street Child. In India, Alokit, a local NGO that is part of the Global School Leaders network, conducted a replication from April to June 2021. In the Philippines, the study was launched in August 2021, ran through January 2022, and was led by Innovations for Poverty Action (IPA) and the government's Department of Education. The final study in Uganda was launched by an NGO called Building Tomorrow from September to January 2022. Instructors were most often teachers and teacher aides; in a few cases, instructors were community volunteers.

Altogether, five replications were conducted within 18-24 months across five countries and 16,000 students. Notably, two countries, Nepal and the Philippines, included delivery by government in addition to delivery by NGOs. This experimental design tested replicability across contexts as well as implementation mode, including scalable government delivery models. In all settings, Youth Impact had no prior presence in the country, besides the original study site of Botswana. This reveals potential for replication and scale through new presence and partnerships rather than pre-existing ones. Youth Impact provided a common thread to enable learning from experience across trials, know-how of the mechanisms needed for program effectiveness, and technical support and training, including collecting monitoring data to effectively target instruction.

[2] Of note, in this paper, we focus on the combined phone call and SMS treatment arms, where we have substantial monitoring data, and which initially worked well, to explore if it could be implemented successfully across contexts, rather than SMS only treatment groups which did not find statistically significant effects in the initial proof-of-concept study (although we continued to evaluate it in replications).

[3] This framework and terminology of "efficacy" and "effectiveness" trials has been used in the field of implementations science in health (Bauer et al. 2015). The field of implementation science, which to date has been most prominent in health, gained traction in the early 1960s with Rogers et al. (1962) publication of "Diffusion of Innovations."

## II. Data and Methods

### A. Monitoring Data Collected

Monitoring data was collected across nearly all replication sites and partners, with an emphasis on implementation fidelity and program quality, in addition to more traditional input metrics such as household reach.

In terms of implementation fidelity and instruction quality, we measure the degree to which instruction was delivered and accurately targeted. To derive this measure, we collected data on the math operation taught and whether the child got the operation correct. Detailed monitoring data was collected in India, Nepal, Philippines, and Uganda, including for government delivery as well as NGO delivery in Nepal and the Philippines. In Kenya, the first replication study, monitoring data focused on reach rather than targeted instruction, since this was the first replication study and the targeted instruction monitoring system took effect in later replication trials. In addition to the replication studies, we include monitoring data on targeted instruction in Botswana, the proof-of-concept study, as a comparison point.

### B. Measuring Implementation Fidelity – The Degree of Targeted instruction

We construct a measure of accurately targeted instruction, a key mechanism for program effectiveness and implementation fidelity, using monitoring data as follows:

$$(1)\ T_s = \begin{cases} 1, \ if\ o_{s-1} - 1 = o_s \ \cap \ c_{s-1} = 0 \\ 1, \ if\ o_{s-1} \quad\ = o_s \ \cap \ c_{s-1} = 0 \\ 1, \ if\ o_{s-1} + 1 = o_s \ \cap \ c_{s-1} = 1 \\ 0, \qquad\qquad\qquad\ otherwise. \end{cases}$$

where $T_s$ denotes targeted instruction in a weekly session *s*, *o* is the operation taught in session *s* (with a designated order such that *o*: 0 = cannot do any operations, 1 = addition, 2 = subtraction, 3 = multiplication, 4 = division), and $c_{s-1}$ denotes whether the problem given in the prior session was answered correctly. If the prior sessions' problem was answered correctly ($c_{s-1} = 1$), then the operation in the current session is targeted if it is an operation more advanced than in the prior session: $o_{s-1} + 1 = o_s$. For example, if the student was given a subtraction question and got it correct, they would progress to multiplication. If the prior sessions' problem was answered incorrectly ($c_{s-1} = 0$) then the operation taught in the current session is targeted if it did not progress or is more basic: $o_{s-1} = o_s$ or $o_{s-1} - 1 = o_s$. If the student was given a subtraction question and got it incorrect, they would be taught subtraction again, or reinforce addition skills.

We construct this measure using weekly data in the Philippines and Uganda for all eight weeks. In Nepal and Botswana, we construct this measure using a similar but slightly adjusted variable: rather than have two independent variables for correct problem and operation taught, we coded a single variable indicating the highest correct problem that was solved. We collected these data in all weeks besides in Botswana where we collected data in the last four weeks. We omit comparative analysis of India, since different targeting rules were used such as progressing after two weeks of instruction, and we omit Kenya, since while reach data was collected, targeted instruction data was not. We include estimates of status quo instruction where instruction is rarely targeted. To construct this benchmark, we use monitoring data in non-targeted comparison groups of prominent studies in India (Banerjee et al. 2017); rates range from 0.0 to 3.8 percent.

We conduct two primary analyses where we have comparable targeted instruction data. First, we compare average rates of how accurately targeted instruction was across replication studies in sequential order of when they were conducted: benchmark rates, Botswana, Nepal, Philippines, and Uganda. Second, we analyze accurate targeting each week in the Philippines and Uganda, where we have the most consistent weekly data.

## III. Results

Figure 1 shows results of average rates of targeted instruction across replication studies in the chronological order in which they were conducted. The benchmark rate of targeted instruction is, on average, less than 1 percent, indicating very little targeting in the absence of explicit efforts to do so. In the proof-of-concept study, Botswana, we find targeted instruction is 41.1 percent, on average. This demonstrates high degrees of targeting relative to benchmark rates, while also revealing significant room for further targeting. The Botswana study showed 0.12 standard deviation gains in learning (Angrist, Bergman, and Matsheng 2022), demonstrating that some targeting, even if imperfect, has impacts on learning outcomes.

While implementation fidelity is typically hard to sustain, we find the replication studies have steadily improved fidelity and quality, measured by rates of targeted instruction. Nepal has approximately 50.9 percent targeted instruction, followed by the Philippines with 64.9 percent targeted instruction, and finally by Uganda with 81.5 percent targeted instruction. These results showcase a clear "learning curve" – lessons on how to effectively target instruction are increasingly operationalized as the replication trials progress.

Lessons to better target instruction include collecting data to target instruction every week (e.g., in Nepal, Philippines, and Uganda) rather than for only four weeks (in Botswana). Another lesson includes modules and practice sessions during training to emphasize targeted instruction, which was done intensively in the Philippines and Uganda. Other lessons include minimizing complex data flows: for example, in the Philippines teachers would input data and rely on pre-programmed recommendations of what operation to teach the following week based on the child's performance. In Uganda, where targeting was highest, instructors would record their student's level as well as the next level to target on a piece of paper on hand, requiring almost no complex data flow.

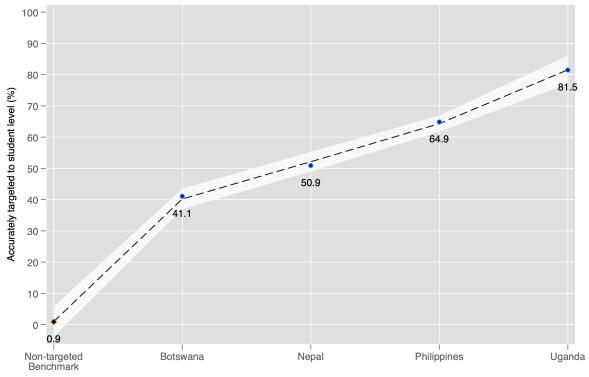


FIGURE 1. LEARNING CURVE – TARGETED INSTRUCTION ACROSS STUDIES

*Notes:* Targeted instruction rates are averaged across all weeks with available data per country. Countries are organized from left to right by the order in which the trial took place starting with status quo benchmarks, the proof-of-concept study in Botswana, and then replications in the following order: Nepal, Philippines, and Uganda. Benchmark rates in the first data point are derived from studies in Bihar and Uttar Pradesh, India from the non-targeted treatment groups. Rates are the weighted average of the percentage of time classes were grouped by ability level in Table 2 in Banerjee et al. (2017).

It is particularly striking that results improve beyond Botswana. It is typically assumed that the proof-of-concept site has the highest fidelity. Yet, in these studies, fidelity improves with each replication, showcasing the ability to learn from experience and translate experience elsewhere, including to entirely new countries, implementation partners, and governments.

Of note, there could be multiple reasons why targeting instruction improved across sites. A likely explanation is learning from experience. However, additional explanations include country-fixed effects. For example, school closures were longest and most severe in the Philippines and Uganda, which might have created enabling conditions to pivot from grade-level curriculum instruction and further targeted instruction. To this end, we explore within-country targeting progress. We analyze trends in the Philippines and Uganda where we have the most detailed weekly data.

Figure 2 shows how targeted instruction improved every week in Uganda and the Philippines. Rates increase from 81 percent to 95 percent in Uganda over time, and from 58 percent to 81 percent in the Philippines. These data showcase another example of learning

from experience, this time within a given county context.

When we analyze trends for government delivery, we find similarly high rates of targeted instruction, as well as steady progress over time. While multiple examples exist of governments struggling to deliver similar implementation fidelity as NGOs (Bold et al. 2018), these results show high and improving implementation fidelity over time, including when effective programs are delivered through scalable government delivery models. This reveals that the "learning curve" applies to governments as well, reinforcing the potential for governments to collect frequent monitoring data and use it to iterate more repeatedly and rapidly (Andrews, Pritchett, Woolcock 2013).

Finally, we explore how accurate targeting translates into descriptive patterns in weekly learning progression captured in monitoring data. We measure progress in operation levels – that is, moving from beginner (those who cannot do any operations) to division. Table 1 shows the number of operations learned per week in Uganda and in the Philippines, where we show both government and NGO delivery.

We observe that in the Philippines NGO delivery yields roughly 0.56 operations per week. Government delivery in the Philippines is similar and slightly higher at 0.58 levels. In Uganda, the average levels gained per week is highest at 0.61. These gains translate into nearly all four basic operations (addition, subtraction, multiplication, and division) learned over all weeks of the program. These results mimic patterns expected from targeted instruction data, with substantial progress occurring in Uganda and the Philippines and the most progression in Uganda, where instruction was most targeted. Moreover, we see high rates of operation progression with government and NGO delivery, corresponding to similar rates of targeted instruction in both delivery models. These results further reinforce the potential to achieve high implementation fidelity through scalable government models.[4]

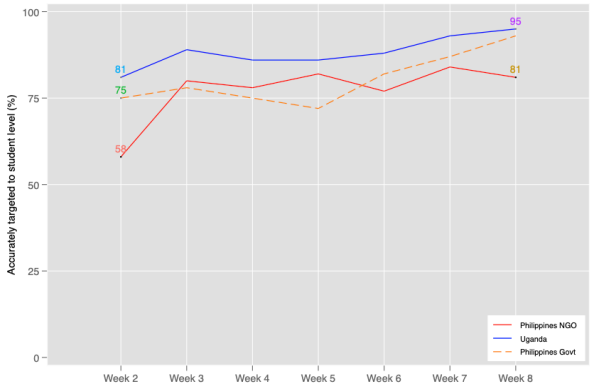


FIGURE 2. LEARNING CURVE – TARGETED INSTRUCTION OVER TIME (WITHIN COUNTRIES)

*Note:* Targeted instruction rates are calculated on average across all students per week in each country by treatment arm, separated by government delivery and NGO delivery in the Philippines.

TABLE 1 – WEEKLY LEARNING PROGRESSION

| | (1) | (2) | (3) |
|---|---|---|---|
| | Operation Taught Progress | Operation Taught Progress | Operation Taught Progress |
| Implementation Week | 0.558 | 0.581 | 0.608 |
| | (0.003) | (0.003) | (0.004) |
| | [0.000] | [0.000] | [0.000] |
| Observations | 4998 | 3488 | 4690 |
| Countries included | Philippines | Philippines | Uganda |
| Delivery Model | NGO | Government | NGO |
| R-squared | 0.863 | 0.906 | 0.850 |

*Notes:* This table reports results from an ordinary least squares (OLS) regression with no constant to capture weekly progress in operations taught. Beginner level (cannot do any operations) is coded 0, addition is coded 1, subtraction is coded 2, multiplication is coded 3, and division is coded 4. Coefficients are reported followed by standard errors in parentheses and p-values in square brackets. Observation counts are at the household-week level.

[4] Our results which highlight the importance of targeted instruction might also help reconcile results with other phone-based tutoring studies. For example, in Sierra Leone, where phone-based instruction was less effective, instruction was linked to national uniform radio programs, limiting scope to target instruction (Crawfurd et al. 2021).

## IV. Conclusion

We present detailed monitoring data across a five-country replication study of phone-based tutoring for disadvantaged students. While much of the literature finds diminishing returns as proof-of-concept studies are replicated and scaled, we find the opposite: implementation fidelity *improves* across replications and over time. A likely explanation is that organizations learn from experience, improving how the program was implemented across contexts and over time, focusing on a crucial mechanism for program effectiveness: targeted instruction. Similar "learning curves" might be achieved in more domains through careful monitoring data use and coordinated lesson sharing, enabling more effective replication and scale-up.

## REFERENCES


**Allcott, Hunt.** 2015. "Site selection bias in program evaluation." *The Quarterly Journal of Economics* 130, no. 3: 1117-1165.

**Andrews, Matt, Lant Pritchett, and Michael Woolcock**. 2013. "Escaping capability traps through problem driven iterative adaptation (PDIA)." *World Development* 51: 234-244.

**Bauer, Mark, Laura Damschroder, Hildi Hagedorn, Jeffrey Smith, and Amy Kilbourne.** 2015. "An introduction to implementation science for the non-specialist." *BMC psychology* 3, no. 1: 1-12.

**Angrist, Noam, Peter Bergman, and Moitshepi Matsheng**. 2022. "Experimental evidence on learning using low-tech when school is out." *Nature Human Behaviour* 6, no. 7: 941-950.

**Angrist, Noam, Simeon Djankov, Pinelopi Goldberg, and Harry Patrinos.** 2021. "Measuring human capital using global learning data." *Nature* 592, no. 7854.

**Araujo, Caridad, Marta Rubio-Codina, and Norbert Schady**. 2021. "70 to 700 to 70,000: Lessons from the Jamaica Experiment." In The Scale-Up Effect in Early Childhood and Public Policy, pp. 211-232. Routledge.

**Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton.** 2017. "From proof of concept to scalable policies: Challenges and solutions, with an application." *Journal of Economic Perspectives* 31, no. 4: 73-102.

**Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, and Justin Sandefur**. 2018. "Experimental evidence on scaling up education reforms in Kenya." *Journal of Public Economics* 168: 1-20.

**Crawfurd, Lee, David Evans, Susannah Hares, and Justin Sandefur**. 2021. "Teaching and testing by phone in a pandemic." Center for Global Development.

**Herriott, Scott, Daniel Levinthal, and James March.** 1985. "Learning from experience in organizations." *The American Economic Review* 75, no. 2: 298-302.

**List, John**. 2022. "The voltage effect: How to make good ideas great and great ideas scale."

**Lucas, Adrienne, Patrick McEwan, Moses Ngware, and Moses Oketch.** 2014. "Improving early-grade literacy in East Africa: Experimental evidence from Kenya and Uganda." *Journal of Policy Analysis and Management* 33, no. 4: 950-976.

**Jedwab, Remi, Paul Romer, Asif Islam, and Roberto Samaniego.** 2021. "Human Capital Accumulation at Work." The World Bank.

**Mobarak, Ahmed Mushfiq**. 2022. "Assessing social aid: the scale-up process needs evidence, too." *Nature*: 892-894.

**Pritchett, Lant, and Amanda Beatty**. 2015. "Slow down, you're going too fast: Matching curricula to student skill levels." *International Journal of Educational Development* 40: 276-288.

**Pritchett, Lant, and Justin Sandefur**. 2015. "Learning from experiments when context matters." 2015. *American Economic Review* 105, no. 5: 471-75.

**Schooler, Jonathan**. 2014."Metascience could rescue the 'replication crisis'" *Nature* 515, no. 7525: 9-9.