

What Would Convince You a System Had Changed?

A primer on measuring systems change in education

Systems change has become one of the most confident claims in education and international development. Funders ask for it, governments announce it, and organizations use the term to signal that a successful program has become larger and more permanent. The instinct is sound: a single intervention, however well evidenced, rarely transforms an education system on its own. Furthermore, no one has settled what counts as evidence that a system has changed.

Health has spent multiple decades with this question. Grace Chee and colleagues¹ separated strengthening a system from adding capacity to it (e.g. medicines, equipment, vehicles or training). Added capacity may improve services while funding lasts, but it is weak evidence that the public system can finance, procure, deploy, supervise and replace it through its own routines. When the Global Fund reviewed sixteen requests in 2018, most spending labelled as systems strengthening was still buying delivery capacity. Education is reaching the same distinction: adopting something that works differs from having a system that can keep it working.

That harder question is asked less often: what evidence would convince us that a system has changed? Even major funders working explicitly on this concede that few established indicators exist for system-level outcomes; shifting from claiming attribution to the more modest standard of contribution.²

When the claim outgrows the evidence

Without a standard, different claims get bundled under a single label. Take a familiar story. A government adopts a promising approach, extends it to a hundred schools and folds it into standard teacher training. This often gets called systems change, though it contains three distinct claims: formal approval signals administrative intent, wider coverage signals reach, and standard training may signal whether the approach can survive without its original champions. None tells us whether classroom practice changed, whether the ministry can better notice and solve problems, or whether the incentives and relationships behind the original problem shifted. The last question is hardest to see directly, which may be why it is so often assumed.

¹ Grace Chee, Nancy Pielemeier, Ann Lion and Catherine Connor, “Why differentiating between health system support and health system strengthening is needed,” *International Journal of Health Planning and Management*, 28(1) (2013), 85–94; “Ambitions and realities: Are Global Fund investments designed to achieve resilient and sustainable systems for health? Findings from the Global Fund Prospective Country Evaluation,” *PLOS Global Public Health* (2024), reporting the Global Fund Technical Review Panel’s 2018 analysis of sixteen 2017–2019 funding requests.

² Varja Lipovsek and Krizna Gomez, with Co-Impact, *Learning, Measurement & Evaluation: A Guidebook* (December 2025).

Approval, the first of the three, can exist entirely on paper. Andrews, Pritchett and Woolcock³ describe how institutions can acquire the outward form of a functioning system, such as a signed policy or a redrawn organisational chart, without changing how it functions. They call it isomorphic mimicry. Their example is a national schooling programme that improved infrastructure and enrolment for years while teacher absence remained at 48 percent. The system looked reformed without yet behaving differently.

Cynthia Coburn identifies a related problem: a reform can spread to many more schools without reaching the depth that determines whether it lasts.⁴ Mimicry and shallow spread are distinct failures, and a status update reporting only the number of schools will surface neither. Mimicry and shallow spread survive scrutiny for the same reason: the evidence collected for a pilot is often promising, but it tends to carry a much bigger claim. A simple discipline catches this: what are we claiming, what evidence supports it, what can we infer, and what remains unproven?

Government approval for scale-up proves political intent to expand funding or grant a mandate. It says little about institutional capability, durability or outcomes. Pratham's Read India had the approval: several Indian state governments adopted it. Bano and Oberoi⁵ found that implementation depended on continuous support from Pratham, which tended to fade before new ways of working replaced old habits. Approval proved intent. The story's third claim, survival without the original champions, remained unproven.

A pilot can move literacy scores by a meaningful margin. That result belongs to the conditions of the trial, and effect sizes vary across contexts, so a result in one place carries little predictive value for another. Whether the gains hold under ordinary government delivery, without intensive support, needs its own standard of evidence.

Does the change survive?

Suppose practice shifts, incentives realign and an institution gets better at solving its own problems. Does the change persist once funding ends? Does it survive a change of government or the departure of a supportive leader? Does it continue when the organisation that supported it steps back? A reform with shallow roots can look durable until leadership changes or support evaporate, and evaluations rarely continue long enough to answer those questions.

Sobral, a municipality in Brazil's Ceará state, shows what durability can look like. In 2001, 40 percent of third-grade pupils could not read a single word. The municipality introduced regular assessment, clear learning goals and sustained support for teachers, and maintained the work through successive mayors

³ Matt Andrews, Lant Pritchett and Michael Woolcock, "Looking Like a State: Techniques of Persistent Failure in State Capability for Implementation," Harvard Center for International Development Working Paper No. 239 (2012); *Journal of Development Studies*, 49(1) (2013), 1–18. The teacher-absence figure is from the Public Report on Basic Education (PROBE), 1996; the revisit figure is described by the authors as preliminary.

⁴ Cynthia E. Coburn, "Rethinking Scale: Moving Beyond Numbers to Deep and Lasting Change," *Educational Researcher*, 32(6) (2003), 3–12.

⁵ Masooda Bano and Zeena Oberoi, "Embedding Innovation in State Systems: Lessons from Pratham in India," RISE Programme Working Paper 20/058 (2020).

and education secretaries. It rose from 1,366th among Brazilian municipalities on the national education index in 2005 to first in 2017.⁶ By then, the reform had endured for sixteen years, through several changes in political and technical leadership.

Repeating the cycle

To an extent, even Sobral is “one problem solved once”. The next test is whether the same public institutions can turn to another problem and run the cycle again: define the problem, decide what evidence a decision needs, test a response, choose whether to adopt, adapt or stop, then carry what works through public budgets and routines. Durable adoption shows that a system can hold one reform in place. Repetition shows that the capability to keep solving sits in the system, beyond any particular reform, leader or partner, and remains available long after the external partner leaves.

How would we see it? The obvious signal is learning outcomes: if scores rise, the system must have changed. But outcomes may improve before government absorbs a reform, and absorption can deepen before national outcomes move. Changes in teaching practice, incentives or institutional problem-solving may take years to register. Outcomes can also improve for unrelated reasons. Children's outcomes remain the reason the work matters, but they cannot by themselves establish whether the system changed.

A leading indicator is a second cycle, on a government-defined priority, run largely by government.

The discipline

The claim for systems change should be no larger than the supporting evidence. Announcing adoption as transformation, or counting coverage as depth, usually claims more than it should. The error runs the other way too: absence of evidence is not evidence of absence⁷. Both require some judgement.

The same discipline applies to us. Elimu-Soko tests an approach through a government-aligned pilot, then supports its transfer to public financing and ownership at scale. Rwanda and Zanzibar have shown adoption and early absorption. Whether those systems can run another adoption cycle will tell us whether we changed, strengthened the system or merely enabled the delivery of a programme.

In the coming weeks we will publish Elimu-Soko's draft framework for measuring systems change. **We want it to start a conversation about what systems change means and what evidence a claim to it needs.** Until then, we are taking the question into every conversation this season.

⁶ Andre Loureiro and Louise Cruz, *Achieving World-Class Education in Adverse Socioeconomic Conditions: The Case of Sobral in Brazil*, World Bank (2020); Harvard Building State Capability, “A Problem-Driven Approach to Education Reform: The Story of Sobral in Brazil” (2022). Sobral's continuity ran in part through one political family; the durability claim rests on the reform surviving successive mayors and education secretaries, and on its replication across Ceará.

⁷ Douglas G. Altman and J. Martin Bland, “Absence of evidence is not evidence of absence,” *BMJ* 311 (1995), 485.