

An agenda for implementation science in education

A persistent global learning crisis calls for studying the implementation of 'what works' at scale

Noam Angrist^{1,2}, Tahir Andrabi³, Stefan Dercon¹, Sharnic Djaker¹, James Habyarimana⁴, Michelle Kaffenberger^{1,5}, Clare Leaver¹, Akshay Mangla⁶, Nompumelelo Mohohlwane⁷, Abhijeet Singh⁸

While nearly all children around the globe now attain at least some schooling, learning achievement remains low worldwide. Although there is growing evidence for approaches that improve learning outcomes (1), less is known about implementing programs at scale effectively and equitably, contributing to persistently low learning outcomes. Research studies often remain small and tightly controlled. Yet, ideas that work in small pilots often fail to achieve similar impacts when implemented at large scale and by governments (2). Implementation fidelity and take-up are rarely measured or reported. This leaves important questions unanswered, such as how much of the variation in effects across contexts is due to implementation and whether an ineffective program failed because of poor design, poor implementation, or both. This article aims to advance the field by putting forward a research agenda and a consolidated conceptual framework to guide implementation science in global education.

Closing gaps between evidence and widespread implementation is challenging: in the area of human health, research highlights a 17-year lag between research discoveries and implementation in clinical practice (3). Scholars were motivated to close gaps that they observed between health research and practice, and as a result began studying how to effectively support the implementation of evidence and innovations in healthcare. Momentum is growing for a comparable focus on implementation in education (4), including by coalitions such as the Building Evidence in Education group (5). Yet a recent study found that quantitative implementation metrics were formally accounted for in only 12.1% of education randomized controlled trials published from 2019 to 2022 (6).

¹Blavatnik School of Government, University of Oxford, Oxford, UK ²Youth Impact, Gaborone, Botswana ³Economics Department, Pomona College, Claremont, CA, USA ⁴McCourt School of Public Policy Georgetown University, Washington DC USA. ⁵Social Science Research Council, Brooklyn, NY USA. ⁶Saïd Business School, University of Oxford, Oxford, UK ⁷Department of Basic Education, Pretoria, South Africa ⁸Department of Economics, Stockholm School of Economics, Stockholm, Sweden. Email: noam.angrist@bsg.ox.ac.uk

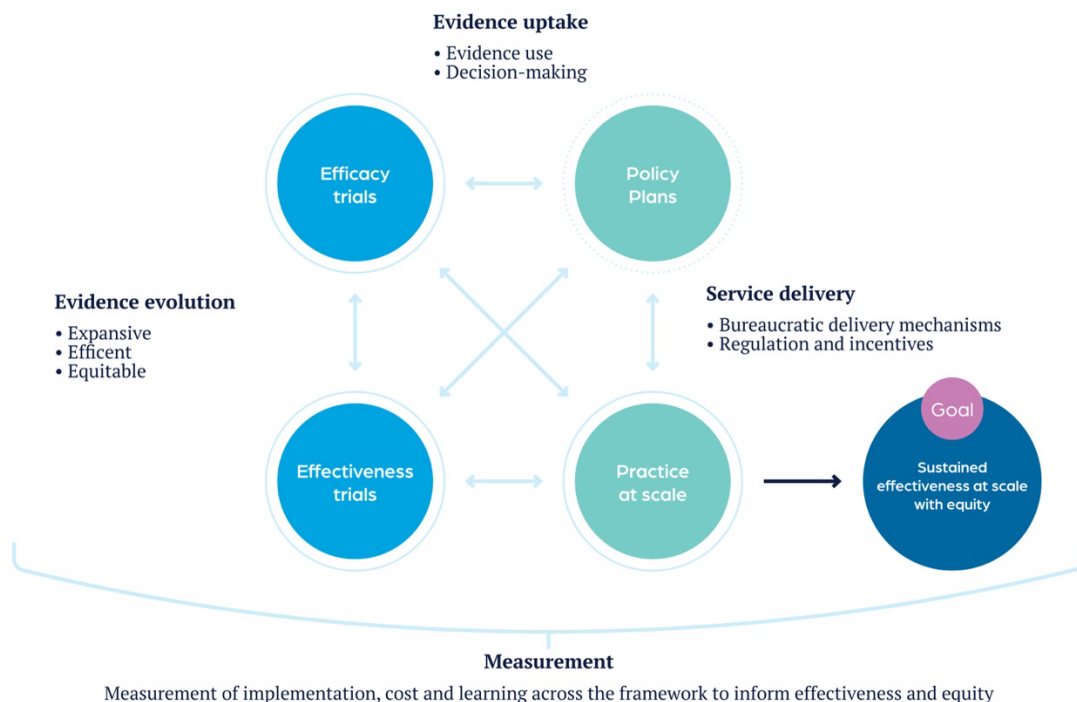
CONCEPTUAL FRAMEWORK

Building on previous implementation science efforts in health and growing interest in implementation in education, we put forward a consolidated framework for implementation science in global education. We build on existing definitions of implementation research (4, 5), and on related themes that have not historically been referred to as implementation science explicitly, but which can advance the field of implementation science, such as the science of scaling (2), state capacity (7), and political economy (8), among others.

The aim of the framework is to organize, prioritize, motivate, and inform the next frontier of research questions in global education and to provide a shared language for researchers as well as policymakers and practitioners. The framework is deliberately simple and includes four phases, two in an evidence strand (efficacy trials to effectiveness trials), two in a policy and practice strand (policy plans to practice at scale), and connections between them (see the figure).

Figure 1. A framework for studying implementation in global education

The aim of the framework is to organize, prioritize, motivate, and inform research questions. The framework includes four phases, two in an evidence strand (efficacy trials to effectiveness trials), two in a policy & practice strand (policy plans to practice at scale), and connections between them. Policies and programs might move through the framework in different ways, or require multiple iteration cycles.



The first phase, 'efficacy trials', includes evidence that is often based on impact evaluations of small, controlled proof-of-concept studies. Most research today consists of efficacy trials. This is in part because publication incentives reward answering narrowly defined questions definitively, which results in studies conducted in artificially controlled environments rather than in complex, ever-evolving, real-world settings. Moreover, a currency in research is novelty – trying new ideas – rather than replicating or expanding on existing ones at larger scale across contexts, and across implementation models. Encouraging leaps beyond efficacy trials is a central goal of implementation science.

The second phase, 'effectiveness trials', builds on efficacy trials to test whether effects are expansive, efficient, and equitable. In terms of expansive effects, these trials test interventions at large scale, when delivered by governments, and across multiple settings. Effectiveness trials also examine efficiency, testing multiple costed delivery models to achieve cost-effectiveness. Understanding how to reduce costs and achieve efficiency is crucial for scale-up by government systems. Effectiveness trials, crucially, also seek to understand equity, testing effectiveness across populations and contexts, exploring heterogeneity to understand not just 'what works', but 'what works for whom'. Effectiveness trials test at least one, if not more, of these dimensions. Far fewer effectiveness trials exist today than efficacy trials – a priority for implementation research is conducting more effectiveness trials.

Making the leap from efficacy to effectiveness trials is challenging but possible in several ways. Coordinated research agendas can establish new norms. Early adopter authors can conduct effectiveness trials, providing concrete examples and setting a new standard in the literature. Collective efforts to synthesize these trials can build a growing body of knowledge, generating momentum and bandwagon effects. Similar shifts evolved in medicine, with effectiveness trials, rather than efficacy trials, increasingly required to prove results. Just as medical scientists have agreed that scientifically reliable evidence must include studies that replicate and which are conducted at large-scale, across contexts, and across implementation models, so too can social scientists. Effectiveness trials can also become more common by advancing credible techniques that are as rigorous as those used in efficacy trials, but which are designed to answer more complex questions in fast-evolving real-world implementation settings, for example, adaptive testing approaches rather than static one-off trials.

The third phase, 'policy plans', focuses on policymaking by governments at senior levels. Plans include policies, memos, frameworks, regulations, and laws. These plans may or may not have translated into implementation at scale and may or may not be evidence-based. Facilitating evidence uptake into policy plans and ensuring that policy plans translate into effective and equitable practice at scale are objects of study for implementation research. Understanding the policymaking process requires careful attention both to

how politics and policy interact, and how technocratic evidence might or might not be integrated into policy formulation.

The fourth phase, 'practice at scale', includes actual program delivery throughout an education system, translating to classrooms and students on the ground. It may or may not reflect government policy, such as when de facto practices differ from de jure policies. Ensuring effective and equitable delivery in practice is its own challenge and a key area of implementation research inquiry. Translating policy into practice at scale involves careful research on bureaucratic norms and incentives. Practice at scale may or may not be evidence-based – implementation research efforts aim to understand both how to translate policy into practice and how to best facilitate evidence-based practice.

The ultimate goal is that as policies and programs move through these phases, they result in sustained effectiveness at scale with equity. In the context of global education, this goal captures widespread and sustained learning for all children. Even the most effective approaches in education still have room to bring many more children up to threshold levels of acquiring all basic foundational skills (9, 10), making a focus on equity essential for achieving learning for all. Crucially, because effectiveness and equity are both central to the end goal, these facets should be considered throughout the entire framework. Moreover, the end goal captures sustainability, ensuring that effects last in the long-run, an aspirational goal that might require several iterations through the framework before sustainability is fully achieved.

Understanding how the phases connect is an object of study for implementation research (see the figure). This includes studying how tightly controlled proofs of concept can lead to widespread, real-world evidence; how de jure policies become de facto practices on the ground; and when efficacy and effectiveness trials are taken up into policy plans or lead directly to practice at scale in the absence of formal government policy and mandates.

Policies and programs might move through the framework in different ways. A more direct path might start with an efficacy trial and progress to an effectiveness trial, then to policy, and then to practice. A less linear path might first involve a program rolling out in practice, and only later being studied in an efficacy trial or informing policy, with challenges encountered in practice at scale informing policy plans or the design of efficacy and effectiveness trials. Interventions and reforms might require multiple iteration cycles through the framework before ultimately arriving at the end goal.

The framework does not aim to prescribe the exact progression through the framework but aims to characterize what progressions could happen. Implementation research – including both quantitative and qualitative studies – aims to bring evidence to bear on specific components in the framework,

while the implementation science agenda overall aims to assess how policies and programs move through the framework, both linearly and nonlinearly, and how this might or might not drive the ultimate goal of sustained effectiveness at scale with equity. Over time, as our understanding evolves, the framework could motivate more normative stances, as evidence accumulates on which productive pathways to follow.

Research along any stage in the framework beyond efficacy trials – where most evidence sits today – as well as on connections across stages, represents a major advance.

PRIORITIES MOTIVATED BY THE FRAMEWORK

The framework motivates several priorities for advancing implementation research in global education (see the figure), with each priority capturing an advance beyond the typical efficacy trial. Of note, the framework does not focus on reproducibility (e.g., code and data transparency and p-hacking). Reproducibility is an adjacent agenda to validate efficacy trials, rather than making the leap beyond them.

Advances beyond efficacy trials increase the likelihood of reaching the ultimate goal of sustained effectiveness at scale with equity. We aim both to encourage such advances – translating evidence to policy, for example – and also to encourage systematic study of each phase and connection in the framework. Such systematic study across the framework requires a combination of quantitative and qualitative methods. This research agenda remains nascent, and below we highlight emerging work in four priority areas to provide a picture of possible research paths forward, identify gaps to be filled, and motivate further work going forward.

Evidence evolution

Advancing from efficacy to effectiveness trials means that the type of evidence we produce should evolve, going beyond initial proof-of-concepts to generating evidence that tests whether effects are expansive, efficient, and equitable. More research is needed to test models that are scalable and delivered through government delivery channels from the outset (2). Rare and prominent examples include the national scale up of the Tusome structured pedagogy program in Kenya and the scale up of Teaching at the Right Level in India (9,10). Understanding system spillovers, or ‘indirect effects’, rather than only focusing on ‘direct effects’, is also essential as interventions are scaled (10).

Programs and policies also need to be continually tested and adapted across settings and delivery models. This is a departure from the typical efficacy trial, which is often one-off, rather than an ongoing learning endeavor. Iterative approaches for evidence generation can support the adaptation process and

enable testing of multiple delivery models, achieving greater cost-effectiveness. Examples include adaptive testing or A/B testing, which involve multiple cycles of fast, affordable, ongoing experiments to generate policy-responsive evidence. In a recent example, 7 of 12 A/B tests yielded efficiency gains for a targeted educational instruction program in Botswana, a rate of implementation improvement on par with the technology sector. These tests addressed both sides of the cost-effectiveness equation, conducting cost-reducing tests, such as redistributing program dosage more efficiently as well as effectiveness-enhancing tests, such as engaging caregivers which more than doubled learning impacts at low marginal cost (11). Often, these approaches are embedded within an implementing organization's infrastructure allowing tight links between evidence and action. Use of these methods in research and in practice remains nascent and should be advanced.

Adaptive testing might not always involve running new trials. For example, for a program like targeted instruction, which groups students by learning levels rather than grade (10), the same principles that underpin this evidence-based approach will look different in different contexts. For example, in India, there is substantial variation in students' ability to recognize digits, so targeted instruction approaches group students on the basis of digit recognition skills. In Botswana, there is substantial variation in the ability to perform operations, so targeted instruction is better off grouping students on the basis of operation skill levels (6). The same evidence-backed principle is at play in both settings – targeting instruction where students are furthest behind and where there is most heterogeneity – but is adapted and operationalized differently depending on the context. This approach to adaptation and generalization involves clarity on the evidence-based principle, and adaptation of that principle using local data in each setting of interest. This example also embodies several principles from problem driven iterative adaptation, providing flexibility to iterate and adapt throughout the policy design and implementation process on the basis of local realities (7).

Understanding program effectiveness at scale beyond initial efficacy trials also necessitates careful consideration of equity across heterogeneous populations and contexts. Scientific progress and evidence-based policy turn on the ability to generalize evidence from one study to the ultimate population of interest. Yet often a program that worked well in one context or for one population might not generalize well to the next, yielding high internal validity but low external validity. Evidence aggregation as well as careful understanding of heterogeneous treatment effects can advance understanding of the generalizability of interventions and their effectiveness across key equity margins. Iterative testing across populations and contexts is another important tool to understand not only 'what works' but 'what works for whom and where'.

Evidence uptake

To facilitate the incorporation of evidence into policy plans two areas emerge as priorities: systematically enabling and understanding the evidence uptake process, and studying policymaker decision-making itself, including its political and institutional environment. Policymakers often seek evidence not only on what works in general, but what will work in their context and for their political priorities. Moreover, policy plans are often not evidence-based, with policymakers having many priorities to balance for a given policy decision. Sometimes targeting second-best rather than first-best policy plans may be necessary to match the evidence to the political incentives of those in power.

Moreover, policymakers are often drawn to educational inputs over outcomes, such as constructing schools and hiring teachers, while showing less motivation to improve downstream outcomes such as learning which are less immediately salient. In many cases, learning outcomes are not incentivized politically because they are often not visible and are difficult to change in short-run political cycles (7).

Although politicians' preferences might not always align with achieving long-run learning outcomes – and evidence to improve them – policy can be better aligned toward outcomes and evidence in several ways. First, measuring learning more quickly, clearly, and visibly could make it politically salient. Second, bureaucrats, rather than politicians, have longer tenures and may prioritize longer-run outcomes. Education evidence labs, units that sit within or just outside governments and work regularly with bureaucrats, are being tested as mechanisms for improving implementation with better use of evidence to achieve downstream outcomes. Third, policymaking can often be improved by setting broad priorities, such as prioritizing foundational learning, not only by focusing on specific interventions. Evidence can and should aim to influence these types of decisions.

A growing body of rigorous research is generating new insights on evidence uptake by decision-makers. This strand of work explores policymaker decision-making and prioritization across various types of evidence, such as how policymakers trade off local versus global evidence or if they are more persuaded by words or numbers. There is a need for more evidence on the evidence uptake process. There is also a need for greater systematic codification of policymaker and practitioner priorities to shape which types of evidence are most relevant for policy.

Service delivery

To ensure that policy plans translate into effective practice in classrooms, understanding service delivery by governments is essential. To close service delivery gaps, two areas emerge as priorities: bureaucratic delivery mechanisms, and regulation and incentives. Emerging evidence aims to

better characterize the service delivery challenge, with new data documenting large gaps between education policies and practice across 50 countries (12).

Delivering a coherent package of interventions rather than a single input may improve service delivery by addressing multiple system constraints at once. Yet, calls for simplification might better enable higher quality service delivery by prioritizing and focusing on delivering the essentials. These debates refer both to government-led service delivery and to donor-funded efforts and private service provision. Emerging work aims to find a balance, identifying a cheap and scalable subset of program 'core components' needed for effectiveness (13). This type of guidance could facilitate and smoothen service delivery, ensuring that key components are in place, so that simplification doesn't result in oversimplification, while focusing on a feasible set of essential elements that can be delivered at scale.

More research is also needed on how to orient the middle tier of the bureaucracy toward implementation of policies to improve learning. A comparative study of implementation across Indian states (8) finds that bureaucratic norms that facilitate deliberation and flexible problem-solving result in better education service delivery than those focused more narrowly on legalistic compliance, by enabling policy adaptation to local needs. This includes research that combines field experiments with in-depth qualitative approaches at scale, which can illuminate key features of institutional contexts. This is consistent with growing calls to move away from one-size-fits-all policies to more locally adaptive and 'best fit' policies, enabling the middle tiers of government not just to adopt policy, but also to adapt it to local implementation realities.

Well-designed regulation and incentives could also improve service delivery. For example, outcomes-based funding models have received growing attention, which incentivizes delivery on outcomes over inputs. There are a growing number of impact bonds in the education sector. However, evaluation and careful documentation of how implementation evolves over time in the context of outcomes funding models remain rare. More rigorous evaluation of these approaches is needed, as is more work to understand models for governments to effectively scale and sustain these approaches.

Measurement

A priority that cuts across the framework is the measurement of implementation, learning, and cost to inform both effectiveness and equity. More consistent measurement is crucial across all education ecosystem actors, including by researchers, policymakers, and practitioners. When government data systems in particular are strengthened, this opens the door for understanding and improving implementation at scale.

It is only possible to understand and improve implementation if we measure it. Yet, measuring and reporting implementation metrics remains rare. Moreover, measurement often remains aggregated; data should be disaggregated along key equity margins. Although often not measured, when implementation is accounted for, it can make-or-break the difference in generalizing effects across contexts (6). Implementation measures are also crucial for understanding outcomes in a single setting. For instance, a null result could be the result of a poorly designed intervention or of a well-designed intervention that simply wasn't implemented as intended. The difference is critical for understanding whether an intervention design is promising and simply needs refinement to improve implementation or whether an entirely new design is needed. Measuring implementation can also enable real-time course-correction and improvement.

Implementation is a broad concept and includes multiple dimensions, including both fidelity and take-up. New work motivated by this framework aims to establish more consistent definitions and standards to measure implementation in education (14). For example, it is important to distinguish between the quantity of implementation (e.g., dosage, inputs) from the quality of implementation (e.g., changes in teacher or classroom practice). Strong implementation can take various forms: it can include high fidelity to a known best practice; or it can also include deliberately shifting from the original implementation plan if realities on the ground reveal that adaptations are essential to achieving the outcome.

Implementation measurement is still in a nascent stage of development. Future work should provide a suite of concrete tools to capture implementation metrics. Although implementation measurement is more advanced in the health sector than in education, even in health there is not yet widespread agreement and standardization of reporting (15). Achieving common reporting standards and metrics can contribute substantially to advancing the field of implementation science.

In addition to measuring implementation, measuring learning in more comparable, meaningful, and transparent ways is important for advancing implementation research in education. Despite there have been recent attempts to report outcomes in education on a common scale, bespoke assessments for each new study are still the norm. Moreover, results are often reported in standard deviations, which vary depending on the underlying distribution, among other factors. This makes it difficult to compare results across studies and contexts. Measures of learning that have common test items and align to global standards could help bridge this gap. Moreover, reporting on specific proficiencies can provide more transparent and actionable information.

Finally, cost collection is essential for understanding the total welfare implications of programs for governments and to facilitate feasible adoption,

adaptation, and scale up by governments. Yet costs remain underreported and understudied. Moreover, cost-effectiveness should consider equity to assess the returns to reaching marginalized populations. Though limited, a few studies aim to provide a review of education programs comparing cost-effectiveness using a unified metric in education (1). More of these types of efforts will be possible when more cost data are collected, analyzed, and reported consistently and comparably.

MAKING IMPLEMENTATION SCIENCE MAINSTREAM

The frontier of education research and practice is to improve implementation. There is growing evidence on what works – now we need to translate that evidence into action and systematically study the process, so that we can do it repeatedly and at increasing scale to achieve effective, equitable, and enduring learning for all.

To make implementation science mainstream in global education, several institutional efforts are needed: first, a concrete research agenda and clear examples of what this research agenda can look like, as we set forth in this article; second, standards of reporting and measurement, including by funding bodies and academic journals; third, a set of institutions dedicated to making these advances. The What Works Hub for Global Education is one such institution, which will support research using the implementation research agenda set out in this article to advance implementation science. More such efforts are needed.

We hope the implementation science for global education agenda motivates a shift toward research that informs the decisions and challenges that confront implementers and government decision makers. By ensuring that more science makes it into implementation, and that more implementation is made scientific, implementation science can support the ultimate goal of all children learning.

**This is the authors' version of the work at the time of acceptance at Science. It is posted here by permission of the AAAS for personal use, not for redistribution. The definitive version was published in Science, 393 (6816): pp. 1084-1087, DOI: [10.1126/science.aee7868](https://doi.org/10.1126/science.aee7868).*

REFERENCES AND NOTES

1. Global Education Evidence Advisory Panel (GEEAP). (2023). *Cost-effective approaches to improve global learning: What does recent evidence tell us?*
<https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099420106132331608>
2. List, J. A. (2024). Optimally generate policy-based evidence before scaling. *Nature*, 626, 491–499. <https://doi.org/10.1038/s41586-023-06972-y>
3. Bauer, M. S. & Kirchner, J. (2020). Implementation science: What is it and why should I care? *Psychiatry Research*, 283, 112376.
<https://doi.org/10.1016/j.psychres.2019.04.025>.
4. Dowd, A. J. (2024). "Everyone is talking about it, but no one is doing it": How implementation research is – or isn't – informing adaptation towards scale in international education. *International Journal of Educational Development*, 108, 103068.
<https://doi.org/10.1016/j.ijedudev.2024.103068>
5. Allison, C. (2023). *Guidance note on using implementation research in education*. Foreign, Commonwealth & Development Office, prepared for Building Evidence in Education (BE²).
<https://www.unicef.org/innocenti/reports/guidance-note-using-implementation-research-education?>
6. Angrist, N. & Meager, R. (2023). *Implementation matters: Generalizing treatment effects in education* (EdWorking Paper No. 23-802). Annenberg Institute for School Reform at Brown University.
<https://doi.org/10.26300/ysyd-t435>
7. Andrews, M., Pritchett, L. & Woolcock, M. (2017). *Building state capability: Evidence, analysis, action*. Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780198747482.001.0001>
8. Mangla, A. (2022). *Making bureaucracy work: Norms, education and public service delivery in rural India*. Cambridge University Press.
<https://doi.org/10.1017/9781009258050>
9. Piper, B., DeStefano, J., Kinyanjui, E. M. & Ong'ele, S. (2018). Scaling up successfully: Lessons from Kenya's Tusome national literacy program. *Journal of Educational Change*, 19, 293–321.
<https://doi.org/10.1007/s10833-018-9325-4>
10. 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.
<https://doi.org/10.1257/jep.31.4.73>
11. Angrist, N., Cullen, C. & Magat, J. (2025). *Cheaper (and more effective) by the dozen: Evidence from 12 randomised A/B tests optimising tutoring for scale*. What Works Hub for Global Education.
https://doi.org/10.35489/BSG-WhatWorksHubforGlobalEducation-WP_2025/001

12. Angrist, N. & Dercon, S. (2024). Mind the gap between education policy and practice. *Nature Human Behaviour*, 8, 2261–2263. <https://doi.org/10.1038/s41562-024-02013-4>
13. Kaffenberger, M., Angrist, N., Hwa, Y.Y., Kayton, H.L., Jukes, M. & Stern, J. (2026). *Core components of teaching at the right level*. What Works Hub for Global Education. https://doi.org/10.35489/BSG-WhatWorksHubforGlobalEducation-WP_2026/001
14. Angrist, N., Djaker, S., Ho, S., Kaffenberger, M., Romero, G. & Sturla, K. (2026). *Measuring implementation in education*. What Works Hub for Global Education. https://doi.org/10.35489/BSG-WhatWorksHubforGlobalEducation-RI_2026/005
15. Lewis, C. C. & Dorsey, C. (2020). “Advancing implementation science measurement” in *Implementation Science 3.0* (pp. 227-251). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-03874-8_9