OCTOBER 2026

Measuring What Matters in EdTech Innovations:

A Quality Standards Framework for Quasi-Experimental Evaluations in Constrained Evidence-base Systems

IN COLLABORATION WITH:

ProFuturo | Fundación Telefónica | "la Caixa" Foundation

SUMMA | Laboratorio de Investigación e Innovación en Educación para América Latina y el Caribe

A joint publication by SUMMA – Laboratory of Education Research and Innovation for Latin America and the Caribbean, and ProFuturo Foundation

AUTHORS

María José Sepúlveda Herane and Dante Castillo-Canales

COORDINATION

ProFuturo Foundation: Virginia Soto Sira, Valentina Ríos, Raquel Valencia

DESIGN AND LAYOUT

ProFuturo

EXTERNAL REVIEWERS

Faizaan Sami (Education Endowment Foundation), Daniel Plaut (EdTech Hub), Katherina Thomas (OECD) and Julie Swaenen (Jacobs Foundation)

INTERNAL REVIEWERS

Virginia Soto Sira (PF), Nuvia Rosas Zambrano y Nicolás Caglevic (SUMMA)

WORKSHOP PARTICIPANTS — BUILDING A CONSENSUS-BASED FRAMEWORK FOR MEASURING THE IMPACT OF EDTECH INTERVENTIONS, MADRID, 3–4 MARCH 2026

María Rodrigues (EEF) · Daniel Plaut and David Hollow (EdTech Hub) · Katherina Thomas (OECD) · María Espinet (EduCaixa / "la Caixa" Foundation) · Juan Bravo (UNESCO / LLECE) · Arnau Picón (ESADE ECSI) · Valentina Ríos, Isabelle Fries, Raquel Valencia, Lia Roitburd and Virginia Soto (ProFuturo) · Dante Castillo, María José Sepúlveda and Andreas Hein (SUMMA)

ACKNOWLEDGEMENTS

We thank the workshop participants and external reviewers for their generous contributions, which substantially shaped this framework.

Publication: October 2026

Use of artificial intelligence. This framework was developed under the direction and responsibility of the authors. Generative AI (Claude, Anthropic) was used to support the organisation and synthesis of source material and reviewer comments, the drafting of text proposals, and general editing. The AI tool did not define the framework's content, standards or conclusions. These were decided by the authors based on the Madrid workshop, expert review and their own methodological judgement. All content was reviewed, revised and validated by the authors, who verified references against original sources and took full responsibility for the final text.

SUGGESTED CITATION

Sepúlveda, M.J., Castillo, D. (2026). *Measuring what matters in EdTech innovations: A quality standards framework for quasi-experimental evaluations in Evidence-base Systems*. SUMMA; ProFuturo Foundation.

LICENCE

CC BY 4.0 This publication is intended for the widest possible dissemination as a public good.

Contents

Presentation 4

01 Why this framework, and for whom? 6

02 Key concepts and what this framework adds 10

03 Actionable Quality Standards for Quasi-Experimental Evaluations in EdTech 15

04 What these standards can and cannot solve 30

05 References 34

Annex 1. Quasi-experimental designs: key assumptions and how to check them

Presentation

EdTech programmes hold great promise for improving education. Yet this promise has still to be proven and harnessed in different contexts, and many initiatives do not always prioritise initiatives don't always prioritise conducting an impact evaluation to assess their merits and limitations. Implementers need evidence to decide whether to improve the programme, continue investing in it, scale it to new settings or discontinue it. They should not have to wait for ideal evaluation conditions before they begin learning from the work they are doing.

This framework is meant to serve as an evaluation lens for organisations that implement EdTech programmes and commission evaluations: foundations, non-governmental organisations and education ministries, among others. It helps them work with evaluators to ask better, more appropriate questions; ensure the production of the right information to support decision-making; and interpret evaluation findings responsibly, by understanding what an impact evaluation can and cannot do.

In particular, this framework advocates for well-designed quasi-experimental evaluations, because this is the reality many organisations face: randomised assignment is often unavailable or unsuitable for the decision at hand, yet a credible design is still needed to measure an intervention's effects. The value of such a design depends on a defensible comparison, meaningful outcomes and a clear account of its assumptions and limitations. Evidence about how a programme was implemented, for whom and under which conditions is equally important for deciding what to do next. Usage figures alone cannot establish an effect on learning, and an effect estimate alone cannot tell an organisation how to improve or sustain an intervention.

The eight standards in this framework provide a shared lens for commissioning, conducting and reading such evaluations. They help implementers and evaluators agree on what evidence is sensible to generate, identify risks to the conclusions and explain what remains uncertain. They also bring costs into view: a programme may need to achieve better outcomes with the limited available resources or maintain its outcomes at a lower cost. Evaluation should inform these choices throughout the programme's development, rather than arrive only as a final verdict.

The framework grew out of a technical workshop convened by Fundación ProFuturo and SUMMA in Madrid on 3–4 March 2026. The aim of the meeting was

Building a Framework for Measuring the Impact of EdTech Interventions. This report reflects the discussions, ideas and a shared effort with peers from EdTech practice, research and international organisations to make evaluation more useful for decision-makers.

The workshop itself builds on a collaboration that ProFuturo and SUMMA have sustained over several years, in which SUMMA has evaluated the ProFuturo programme in Latin America by combining quasi-experimental estimates of its effects on learning, based on national administrative data, with qualitative studies of how it is implemented in schools.

In this sense, we invite the local and international community to read this framework as an argument for why measuring impact matters, even where organisations are not yet able to act on it. Many implementers recognise the value of evaluation but lack the time, capacity or resources to prioritise it. Others assume that measuring impact requires a randomised controlled trial, and that such a design is out of reach in the real conditions in which their programmes operate. For them, the framework offers a clear map of what doing this well involves, so that the building blocks are ready when they decide to begin. The real test will be its practical use: only by applying it will we learn how much a tool like this can contribute.

Lola Martínez Bernabéu
Managing Director, ProFuturo Foundation

Javier González Díaz
Director, SUMMA Foundation
Laboratory of Education Research
and Innovation for Latin America
and the Caribbean

01

Why this framework, and for whom?

Educational technologies (EdTech) are firmly established in the world of education. Digital platforms, online resources, adaptive learning systems or hybrid models of instruction are promoted as ways to support the teaching and learning process and improve learning opportunities. Yet the potential of EdTech is neither automatic nor uniform. Technology can contribute to addressing educational challenges, but its benefits depend on the conditions of access, pedagogical use, governance, sustainability and inclusion under which it is implemented (UNESCO, 2023, Haßler et al., 2025, Castillo-Canales et al., 2026). Without these conditions, solutions designed to reduce learning gaps may reproduce or deepen existing inequalities.

Getting this right has become more urgent. The COVID-19 pandemic intensified an already existing learning crisis (West, 2023). Learning poverty in low- and middle-income countries increased substantially after the pandemic, reaching approximately 70% in 2022 (World Bank, 2022), and learning losses larger among disadvantaged students (Betthäuser et al., 2023). This context created an urgent demand for solutions capable of supporting learning recovery, reducing inequalities and strengthening system resilience.

Decision-makers, at the policy and school practice level in the global south, face the challenge of taking decisions with limited resources and in contexts with acute social and economic needs. In these types of contexts, knowing whether an intervention works is therefore not enough. There is also a need to know whether its potential effects justify the full costs of investing in infrastructure, connectivity, maintenance, teacher training, school adaptations, technical support, and whether it is affordable at scale and sustainable over time (Plaut, 2024). This calls for impact evaluation that is rigorous and fit for purpose (Foster et al., 2023; EdTech Hub, 2026).

In Latin America and the Caribbean, the impact of EdTech depends less on the mere introduction of devices and platforms, and more on the institutional, pedagogical and technological conditions that allow those tools to be used with pedagogical purpose. These include ICT infrastructure, institutional capacities, adequate governance, public-private coordination and programme sustainability are central conditions for maximizing the impact of EdTech interventions (Castillo-Canales et al., 2023; Cueto et al., 2023). From this perspective, the **relevant question is not only whether an EdTech works, but under what conditions, for whom, in which contexts it can effectively contribute to improving learning.**

EdTech innovation, investment and scaling often move faster than the capacity of education systems to generate, interpret and use robust evidence. In many low- and middle-income contexts, decisions to adopt or scale an intervention

are made with limited information, fragmented data and few independent evaluations. The evidence that exists is often weakly comparable, insufficiently transparent or dependent on provider-generated claims (West, 2023; UNESCO, 2023; Cueto et al., 2023; Plaut, 2024). These tensions are particularly acute in constrained evidence systems, where data systems, analytical capacity, evaluative independence and demand for evidence are still developing (Suazo-Galdames et al., 2025; Gough et al., 2020).

In this context, estimating the impact of an intervention on learning is essential. However, the reference standard for doing so, the randomised controlled trial (RCT), is often not feasible, because many programmes are already running at scale when the question of impact arises. The impossibility of randomising is not a reason not to evaluate. Well-designed quasi-experimental designs (QEDs) can generate rigorous evidence on programmes already in motion and under real-world conditions, provided their assumptions and limitations are made explicit. The challenge is therefore not only to produce more evidence, but to produce relevant evidence that can inform programme improvement, policy and school decisions.

This framework is primarily written for teams that implement EdTech programmes and commission evaluations: foundations, NGOs, social enterprises and ministry programme teams. Its contribution is to help them **commission and conduct better evaluations**. By better evaluations we mean evaluations that can credibly and reliably estimate effects on learning or other relevant education outcomes, but also to understand how and what specific mechanisms can help explain why those effects tend to happen or fail to occur.

Evaluators can use the same standards to justify their design choices, assess risks to the findings and report the limits of their conclusions. School leaders and teachers should also know why an evaluation is being conducted, what it involves and how its findings will be used (SUMMA, 2024).

The framework serves three moments:

01

Before an evaluation, use it to agree on the intervention being assessed, the outcomes that matter, the feasible comparison, the implementation information to collect and the costs to document.

02

During an evaluation, use it to check that the programme is delivered as planned, record changes to the platform, content or delivery model, document deviations from the analysis plan and detect attrition or missing data in time to address them.and consequences of that decision.

03

After an evaluation, use it to judge what the findings support and whether they are sufficient to improve, continue, adapt, scale or discontinue th programme. The strength of evidence required should reflect the importance and consequences of
that decision.

The framework proposes eight actionable quality standards for conducting and interpreting QEDs in EdTech. It does not present QEDs as a self-sufficient answer to the question of what works. It treats them as one piece of a shared and evolving evidence infrastructure, in which knowledge gaps are addressed progressively by triangulating impact estimates with evidence on implementation, qualitative insight into how change unfolds across contexts and subgroups, cost information and longitudinal monitoring. High-quality evidence of impact is possible when evaluations keep in mind the decision they must inform, acknowledge the limitations of context, data and methods, and combine different approaches to establish what can and cannot be attributed to the intervention.

02

Key concepts and what this framework adds

2.1

What do we mean by impact in EdTech interventions?

In many educational initiatives that rely on digital devices (EdTech) to improve learning outcomes, evidence of access, use or engagement is treated as an indication of effectiveness. However, devices distributed, platform usage, completion of digital activities or time spent online can inform adoption and implementation, but they do not demonstrate by themselves that an intervention improved learning or changed educational trajectories (Kucirkova et al., 2024; West, 2023; Plaut, 2024).

Impact in EdTech should be understood as a causal claim, this is, a change that can be *plausibly attributed* to an intervention, in relation to outcomes that matter for students, teachers, schools or systems. This requires distinguishing between access to technology, intensity of use, changes in pedagogical practice, student engagement, and learning outcomes, progression indicators and system-level effects. Each level can be valuable, but each point to different types of inferences: a programme may be widely used without improving learning; an initiative can improve engagement without changing academic outcomes; an innovation can generate effects only under specific implementation conditions (Kucirkova et al., 2024; Foster et al., 2023; Garg & Bakker, 2025).

From a methodological perspective, the aim is to isolate the effect of the intervention as far as possible. This framework is oriented towards strengthening attribution. Where a design cannot fully support it, the claim should be qualified and its limits made explicit. In such cases, it may be more appropriate to speak of a *plausible contribution* to change, provided that mechanisms, implementation conditions, process evidence and limitations are made explicit.

These distinctions matter because **EdTech initiatives generate large volumes of data that can make programmes appear more measurable than they are**. Platform analytics, dashboards and usage indicators can support monitoring and improvement, but they can also distort evaluation if they substitute independently measured outcomes. In EdTech, the risk is not only measuring irrelevant outcomes, but measuring what is easiest to collect rather than what is educationally meaningful (Kucirkova et al., 2024; UNESCO, 2023).

A stronger approach connects impact claims to a clear theory of change. The theory of change specifies what the intervention is expected to change, through which mechanisms (for example adaptive practice, additional learning time, teacher support or feedback loops), for whom and under what conditions (Sharples et al., 2024). Without this clarity, impact estimates may indicate whether outcomes changed but say little about why those changes occurred or whether they can be sustained, adapted or replicated elsewhere (Damschroder et al., 2009).

Impact is not as a single metric, but an evaluative claim that depends on alignment between intervention logic, outcome measurement, implementation conditions and causal design. This framing is consistent with EdTech quality frameworks that distinguish evidence of use from evidence of efficacy or effectiveness, and with implementation-oriented approaches that emphasise how programmes operate in context (Kucirkova et al., 2024; Foster et al., 2023).

2.2

Why quasi-experimental designs?

Once impact is understood as a causal claim, the evaluation challenge is to estimate what would likely have happened in the absence of the intervention. This counterfactual problem is especially difficult in education systems, where programmes are rarely implemented in neutral conditions. Schools, teachers or students who participate in an EdTech programme may differ systematically from those who do not, including differences in motivation, leadership, resources, connectivity, prior achievement or institutional capacity.

As the literature suggests, Randomised Controlled Trials (RCTs) can provide strong causal evidence when they are feasible, ethical and aligned with implementation conditions. However, many EdTech programmes are evaluated after they have already been implemented, expanded or embedded into policy decisions (Kucirkova et al., 2024). In these cases, random assignment may no longer be possible, politically acceptable or useful for the decision at hand, especially in large-scale programmes, public-private partnerships, regional initiatives or interventions already operating across schools and territories because educational and social needs are urgent to address.

Quasi-experimental designs (QEDs) can be a pragmatic and rigorous response to a common real-world situation. This solution offers an answer to the question of how to estimate effects when randomisation is not possible nor available. Designs such as difference-in-differences, regression discontinuity, interrupted time series, matching approaches or combinations of these methods can strengthen causal inference by constructing a credible comparison group or using existing implementation rules, time variation or eligibility thresholds.

The value of QED lies not in being easier alternatives to experiments, but in making explicit the assumptions under which causal claims can be made using observational, administrative or implementation data (What Works Clearinghouse, 2022; Scher et al., 2015; Waddington et al., 2014; Barker et al., 2024).

QEDs do not automatically produce credible evidence. A poorly designed model can overestimate or underestimate the effect of an intervention, especially when selection bias, baseline imbalance, differential attrition, weak outcome measurement or implementation heterogeneity are not adequately addressed (Sterne et al. 2016; What Works Clearinghouse, 2022).

For EdTech, these risks are particularly acute because participation may be shaped by digital infrastructure, school readiness, teacher motivation, prior technology use, provider engagement or differential access among students. In these cases, the comparison group may look similar on paper while still differing in ways that matter for outcomes (Waddington et al., 2014).

This is why QEDs to be conducted with rigour require explicit quality standards. A credible QED needs to define the intervention and comparison condition clearly, specify the estimand, justify the identification strategy, assess baseline comparability, document data limitations, measure outcomes independently where possible, report missing data and attrition, and interpret results in light of

implementation variation (What Works Clearinghouse, 2022; Sterne et al., 2016; Des Jarlais, et al., 2004).

Their credibility is further strengthened when they are combined with implementation analysis, qualitative evidence, heterogeneity analysis and costing. Read this way, the framework is better understood as a proposal for building an ecological perspective on the role and value of evaluation in educational improvement, rather than as a purely methodological guide.

2.3

What exists and what this framework adds

A growing number of frameworks guide how evaluations and EdTech interventions should be assessed. They fall into three families:

Family	Examples	Main contribution	Primary target audience
Methodological standards for impact evaluation	What Works Clearinghouse procedures and standards handbook, ROBINS-I, JBI Critical Appraisal Tools, Transparent Reporting of Evaluations with Nonrandomized Designs (TREND)	What non-randomised evaluations require to be credible: addressing selection, baseline imbalance, attrition, measurement and selective reporting.	Researchers and systematic reviewers: standards to appraise the risk of bias of impact studies and to report them transparently
Implementation frameworks	Education Endowment Foundation: A School's Guide to Implementation. Guidance Report; Consolidated Framework for Implementation Research (CFIR).	Impact depends on how programmes are introduced, adapted, supported and sustained in real settings. The importance of understanding the problem, preparing the system, supporting adoption, monitoring delivery and using evidence for improvement.	School leaders and implementation teams: guidance to plan, deliver and sustain change in schools
EdTech-specific frameworks	Consolidated Benchmark; EdTech quality frameworks; EdTech Hub toolkit	Evidence of use is not evidence of efficacy; fitness for purpose, pedagogical value and implementation context also matter (Kucirkova et al., 2024; Foster et al., 2023; EdTech Hub, 2026).	EdTech developers, investors, procurers and policymakers: criteria to judge the quality and evidence claims of EdTech products and programmes

Taken together, these frameworks provide a strong foundation for designing a guidance of quality standards for conducting impact evaluations. However, they do not offer a clear route for integrating causal credibility, implementation quality, outcome validity, cost and context into a single applied evaluation process. Nor are they designed primarily for those who commission evaluations of programmes already in motion.

EdTech evaluations rarely follow methodological textbooks: programmes may be implemented before the evaluation begins, comparison groups may be built from administrative data, platform versions may change during implementation, usage data may be proprietary, and implementation quality may vary across schools, teachers and territories. This framework does not replace existing standards. It translates their most relevant contributions into actionable criteria, organised around eight practical risks:

01
Theory of change and evaluative clarity

02
Selection, comparability and causal credibility

03
Outcome validity and meaningful measurement

04
Implementation and contextual transparency

05
Missing data, attrition and robustness

06
Cost, affordability and sustainability

07
Transparency, reporting and interpretability

08
Transferability and progressive strengthening of evidence

Its purpose is not to eliminate the trade-offs between rigour, feasibility and use, but to make them visible. That way, those who commission and conduct evaluations can agree on what can be concluded, what remains uncertain and what evidence should be strengthened next.

> An EdTech intervention may be well designed but poorly implemented; widely used but ineffective; promising in one context but not scalable.

03

Actionable Quality Standards for Quasi-Experimental Evaluations in EdTech

This section presents eight actionable standards for quasi-experimental designs in EdTech. Table 1 summarises each standard: its main safeguard, a key question for those who commission evaluations and the minimum actionable requirements. Each standard is then developed in more detail, with the rationale behind it, guiding questions and a "What can go wrong?" example: a hypothetical but frequent situation that shows the consequences of not applying the standard.

Each standard plays a distinct function: Standard 1 establishes the evaluative logic on which the others depend, and anchors the evaluation to the decision it must inform. Standards 2, 3 and 5 form the core of internal validity, addressing the three classic threats to causal credibility in quasi-experimental designs: selection and comparability, outcome measurement, and missing data or attrition. Standards 4 and 6 address the real-world conditions under which interventions operate: *how they are implemented, and whether they are affordable and sustainable*. Standards 7 and 8 operate transversally across all the others: Standard 7 ensures the traceability and transparency of the evaluative judgement, while Standard 8 places each evaluation within a cumulative process of learning and specifies the conditions under which its findings can travel.

Four principles underpin the standards:

Credibility

The evaluation must make clear how the impact claim is constructed and what risks of bias may affect it.

Interpretability

The evaluation must explain what was implemented, under what conditions, and how outcomes should be understood.

Usefulness

The evaluation must produce evidence that can inform decisions about improvement, adaptation, scaling, procurement or discontinuation.

Equity

The evaluation must examine whether observed effects are distributed fairly, since a positive average effect can still mask widening gaps between students.

Together, these principles recognise that quality in EdTech evaluations is not only a matter of statistical technique, but it also depends on the alignment between the intervention logic, causal design, outcome measurement, implementation evidence, cost information and transparency of reporting.

Table 1: Eight actionable standards for quasi-experimental impact evaluation in EdTech

The standards below do not replace existing methodological or implementation frameworks. Instead, they indicate when those frameworks should be used and translate their most relevant requirements into minimum actionable steps for QEDs in EdTech. Each standard identifies the main safeguard it provides, and the minimum actions needed to make impact claims more credible, interpretable and useful for decision-making.

Standard	Main safeguard	Key question for commissioners	Minimum actionable requirements
01 **Theory of change and evaluative clarity**	Avoid evaluating "technology", "the platform" or "the devices", when what schools receive is a model (devices, content, teacher training, coaching), and state how each mechanism will be observed.	What exactly are we evaluating, and through which mechanisms do we expect it to improve learning?	→ Define the intervention model and core components. → State expected mechanisms as testable "if… then…" statements. → Define target population and relevant subgroups. → define primary and secondary outcomes aligned with the theory of change. → state contextual and implementation assumptions required for impact (connectivity, devices, teacher time, leadership). → Plan how mechanisms will be measured during implementation.
02 **Selection, comparability and causal credibility**	Assess and signal the risk that effects reflect *which* schools joined (better connectivity, stronger leadership, higher prior results) rather than the programme itself.	Why did these schools join, and why are the comparison schools a credible picture of what would have happened without the programme?	→ Describe how schools, teachers or students entered the programme. → Define and justify the comparison group. → Assess baseline equivalence on key pre-intervention variables; including teacher-level variables when assignment is by school. → State the design's key assumption and how it was checked*. → report remaining risks of selection bias and unobserved confounding. → conduct robustness or sensitivity checks*.
03 **Outcome validity and meaningful measurement**	Prevent platform usage, engagement or easily available metrics (logins, minutes on the platform, completed activities or in-app scores) from being reported as learning impact	Is learning measured with an instrument the provider does not control, aligned with the theory of change and applied at the right time?	→ Define primary and secondary outcomes before analysis. → Separate usage, engagement, implementation and learning outcomes. → Justify why each outcome is meaningful and aligned with the theory of change. → Use independent or externally validated measures*. → Triangulate platform data with other sources. → Report limitations of platform-generated or provider-owned metrics. → Avoid treating usage as direct evidence of learning impact.

Standard	Main safeguard	Key question for commissioners	Minimum actionable requirements
04 **Implementation and contextual transparency**	Ensure effects are read in light of what actually happened in classrooms: dose received versus planned, device sharing, connectivity, teacher mediation and platform changes.	Was the programme implemented as intended, and what explains differences across schools and teachers?	→ Describe the programme as implemented, not only as designed. → Report fidelity, dosage, intensity, adaptation and reach; including usage trajectories over time. → Record key contextual conditions affecting implementation; and changes in platform, content, algorithms or delivery model. → Collect stakeholder perspectives to explain the "why". → Analyse variation across schools and teachers; Report effects for pre-specified disadvantaged subgroups* → Interpret effects in light of implementation.
05 **Missing data, attrition and robustness**	Avoid basing impact claims only on the most visible, connected or engaged participants, especially when missingness is related to access, use or implementation quality.	Who is missing from the data, why, and does it differ between groups?	→ Report total and differential attrition (reduction in numbers from the original sample) and the reason for it; → Test whether missingness relates to baseline characteristics or usage; → Explain how missing data were handled; → Run sensitivity checks using alternative samples, models or assumptions*; → Avoid presenting complete-case results as definitive when missingness is systematic.
06 **Cost, affordability and sustainability**	Prevent effectiveness claims from being interpreted without considering its full cost: devices and licences are often a fraction of total costs, which include connectivity, replacement, coaching and teacher time.	What does it cost per student to achieve this effect, who pays, and what would it cost at scale?	→ Report costs in enough detail to recalculate them under different scenarios. → Distinguish start-up, recurrent, maintenance and replacement costs. → Identify who bears each cost (provider, ministry, school, families). → Relate costs to observed outcomes*. → Discuss affordability and sustainability at scale. → Compare with alternative interventions when evidence allows.
07 **Transparency, reporting and interpretability**	Avoid overstated, or selective claims, especially when data come from the provider's platform or when those who analyse have a stake in the programme.	Could an independent reader trace how the impact claim was built, what was planned versus found, and who funded and analysed it?	→ Document the evaluation protocol and analysis plan before analysis. → Report data sources, eligibility criteria and sample construction. → Distinguish confirmatory from exploratory analyses. → Make assumptions, deviations and limitations explicit. → Disclose conflicts of interest and provider involvement. → Report results in a way that supports comparison, synthesis and decision-making.

Standard	Main safeguard	Key question for commissioners	Minimum actionable requirements
08 **Transferability and progressive strengthening of evidence**	Map the conditions under which an effect was found (context, platform version, dose, population, time) so others can judge whether it transfers, and plan what evidence comes next.	Under what conditions did this work, and would they hold in another context?	→ State the level of certainty the evidence supports; do not treat a single study as conclusive. → Describe the conditions and assumptions under which effects were observed → Identify what remains uncertain and why; specify what evidence should be strengthened next; → use findings to improve implementation and future evaluation design; → update evidence when the intervention, platform or context changes; → Share data and documentation to future synthesis*.

(*) Required when data or context allow. If not done, the evaluation report must state why.

01

Standard 1: Theory of change and evaluative clarity [1]

An evaluation should begin with a clear theory of change that explains how the EdTech intervention is expected to generate outcomes, for whom and under what conditions, and which decision the evaluation is meant to inform. This matters in EdTech because digital interventions often combine multiple components: technology provision, platform use, teacher training, coaching, connectivity, data systems, content alignment and institutional support.

Evaluating "the platform" when the model includes all of these makes the results impossible to attribute or replicate. Current evaluation standards accept that a combined intervention can produce a valid impact estimate for the package, provided the package is clearly defined (What Works Clearinghouse, 2022).

A credible theory of change should specify the problem addresses, the target population, the core components of the programme, the expected mechanisms of change, the outcomes to be measured and the contextual assumptions required for the intervention to work (Education Endowment Foundation, 2026).

→ Evaluations of EdTech often report whether an intervention worked without explaining how (Haβler et al., 2025). Examples for mechanisms of change may include grounding technology in pedagogical processes, creating dynamic feedback, leveraging technology for adaptive teaching and planning for implementation.

1 For intervention logic and theory of change, evaluation teams should draw on implementation-oriented frameworks such as the EEF Guidance on Theory of Change and logic model and broader implementation research frameworks, including CFIR, to specify the problem, intervention components, mechanisms, actors, implementation process and contextual assumptions.

- EdTech assumptions may include access to devices, minimum connectivity, teacher readiness, curriculum alignment, student digital literacy, school leadership, technical support or sufficient time for use.

Where possible, the theory of change should be tested and refined before a formal impact evaluation, through iterative or design-based research. Even for programmes already running, reviewing what is and is not working before the evaluation begins can strengthen both the programme and the study.

Guiding questions

- What exactly are we evaluating: the platform, or the model that surrounds it?
- Through which mechanisms do we expect learning to improve, and how will we observe them?
- Which conditions must hold for the programme to work, and do they hold in our schools?
- Which decision will this evaluation inform?

What can go wrong?

A funder commissions an evaluation of "the tablet programme". The programme also includes teacher training and monthly coaching, but neither is documented. The evaluation finds a positive effect. No one can say whether it came from the devices, the training or the coaching, or whether a cheaper version would work.

This standard addresses a common risk in EdTech evaluation: treating technology as the intervention, rather than specifying the model of change that makes technology educationally meaningful. The relevant unit of evaluation is not simply a platform, device or software package, but the whole intervention model through which technology is expected to affect teaching, learning or system outcomes.

02

Standard 2: Selection, comparability and causal credibility [2]

A QED must explain why the comparison group provides a credible estimate of what would likely have happened in the absence of the intervention. This is the central condition for causal inference. In EdTech, this challenge is especially important because schools, teachers or students who participate in a programme may differ systematically from non-participants in ways that affect outcomes. These differences may include prior achievement, resources, motivation, connectivity, digital readiness, institutional capacity or previous exposure to technology.

2 For selection bias, comparability and causal credibility, teams should consult WWC v5.0 for education-specific standards on baseline equivalence, attrition and evidence ratings; ROBINS-I for domains of bias in non-randomised studies; and QED-specific guidance such as Scher et al. and Waddington et al. for design and risk-of-bias considerations.

When evaluations are conducted after implementation has begun, the task is not to eliminate this risk but to assess and signal it transparently. The evaluation should make explicit how participants entered the programme, how the comparison group was built, which variables were used to check baseline equivalence and what risks of unobserved confounding remain.

- When assignment is at school level, but implementation depends on teachers, baseline equivalence should also be checked on teacher-level variables, such as prior EdTech use or digital competence, where these could shape effects. This is particularly relevant for teacher-facing tools, including generative AI.

Each design rests on a key assumption that should be stated and, where the data allow, checked. If a check is not feasible, the report should say why. Statistical techniques such as matching, difference-in-differences, regression discontinuity or interrupted time series can strengthen causal credibility, but only if their assumptions are clearly justified and tested *(see Annex 1)*.

Guiding questions

- Why did these schools join the programme, and why did others not?
- What makes the comparison schools a credible picture of what would have happened otherwise?
- What is the design's key assumption, and how was it checked?

What can go wrong?

A provider compares test scores of schools that use its platform with those that do not, and reports the difference as impact. Participating schools had volunteered and already had better connectivity and higher prior scores. Part of the difference existed before the programme started.

This standard addresses the risk that impact claims are driven by pre-existing differences rather than by the intervention itself. A comparison group may appear similar on observable characteristics but still differ in ways that are difficult to measure, especially in contexts where participation is shaped by school readiness, provider engagement or digital infrastructure.

03

Standard 3: Outcome validity and meaningful measurement [3]

A QED should measure outcomes that are valid, meaningful and aligned with

3 For outcome validity and measurement, teams should use EdTech evidence frameworks such as the Consolidated Benchmark/EVER and EdTech quality frameworks to distinguish use, engagement, efficacy and effectiveness claims. WWC standards should be consulted when assessing whether outcome measures are valid, reliable, independent and aligned with the causal claim.

the theory of change. In EdTech, this standard is essential because digital programmes often generate large amounts of data that can be mistakenly used as evidence of impact. Platform logins, time on task, number of completed activities or engagement scores may be useful for monitoring implementation, but they do not by themselves demonstrate improvements in learning, teaching quality or educational trajectories.

Outcome measures should distinguish between implementation indicators, intermediate outcomes and final outcomes. For example, frequency of platform use may indicate exposure; student engagement may indicate a possible mechanism; but learning outcomes, progression, attendance or teacher practice require valid and independently interpretable measures. Primary outcomes should be measured with instruments that the provider does not control, that are aligned with the intended construct and that are administered at a time consistent with the intervention.

Measures developed by intervention developers can produce meaningfully different effect sizes from independent ones, which is why major evidence clearinghouses base their effectiveness ratings on independent measures where these exist (What Works Clearinghouse, 2022). Platform data remain valuable when they are triangulated with other sources, for example by linking usage records to independent assessments, to trace causal pathways.

This standard addresses the risk of overclaiming impact based on metrics that are easy to collect but weakly aligned with educational value. It also helps prevent Goodhart-type problems, where indicators used for monitoring become targets in themselves and distort the interpretation of programme success.

This standard should also flag the risks associated with proprietary data and provider power. In EdTech, much of the evaluative risk comes from the fact that data are decided, produced and selected by providers themselves. Addressing this requires attention to the independence of the data, access to raw or auditable records, visibility over changes in algorithms or platform versions, potential bias embedded in indicators, and conflicts of interest between those who analyse and those who report results.

Guiding questions

- Is learning measured with an instrument the provider does not control?
- Is the measure aligned with what the programme is designed to change?
- Is it applied at the right time relative to the intervention?
- How will platform data be used: as evidence of exposure, of mechanisms or of impact?

What can go wrong?

An evaluation report concludes that “students who used the platform more achieved higher in-app mastery scores”, and presents this as evidence of learning impact. The measure was designed by the provider, rewards familiarity with the platform itself and has no comparison group.

04

Standard 4: Implementation and contextual transparency [4]

An evaluation should document not only whether an intervention was implemented, but how it was implemented, with what level of fidelity, adaptation, intensity and support. This is particularly important for EdTech because the same digital solution can produce different results depending on connectivity, teacher mediation, leadership, technical support, platform version, curriculum alignment, student access and school routines.

> “
>
> Teacher involvement is especially important: a systematic review of EdTech for disadvantaged pupils concludes that “learning gains may be limited – or even negative – without the active involvement of teachers” (Haßler et al., 2025).

This is important because a platform requiring stable connectivity may function well in one school and fail intermittently in another with unreliable electricity or internet access; a design that assumes daily teacher facilitation may falter where staffing or training is inconsistent. These are not peripheral details: they illustrate the broader principle that context conditions technique, and that the same intervention can behave differently depending on the material and institutional conditions in which it is implemented.

Implementation evidence helps interpret impact estimates. A null effect may reflect a weak intervention, poor implementation, insufficient exposure or a mismatch between the programme and the context. A positive effect may depend on specific enabling conditions that are not automatically transferable to other settings. Positive final outcomes depend on a combination of implementation outcomes (adoption, fidelity, feasibility, acceptability and sustainability), and each should be monitored (Sharples et al., 2024).

In EdTech, platform analytics can show usage trajectories over time: high initial engagement often declines, leaving a smaller group of sustained users, which also matters for statistical power. Beyond process indicators, qualitative evidence from teachers, school leaders and students helps explain why outcomes occurred and how the model could be adjusted.

4 For implementation quality and contextual interpretation, teams should use implementation frameworks such as EEF implementation guidance and CFIR to document fidelity, adaptation, dose, reach, actors, setting and implementation process. TREND can also guide transparent reporting of how the intervention was delivered.

Guiding questions

- Did students receive the planned dose, and if not, why not?
- What conditions in schools (connectivity, devices, timetables, teacher support) shaped implementation?
- How did use evolve over time, and did it vary across schools and teachers?
- What do teachers and students say about how the programme works in practice?

What can go wrong?

An evaluation reports a null effect without implementation data. The funder discontinues the programme. Later it emerges that most schools received their devices four months late and teachers were never trained: the model that was evaluated was not the model that was designed.

This standard addresses the risk of producing impact claims that are statistically plausible but operationally uninterpretable. Without implementation evidence, decision-makers may know whether an effect was observed, but not what conditions made it possible or what would be required to replicate, adapt or scale the intervention.

05

Standard 5: Missing data, attrition and robustness [5]

A QED should make visible how missing data, attrition and incomplete implementation records affect the credibility of the findings. In EdTech, missingness is often related to the intervention itself. Students with limited connectivity, low attendance, weaker digital access or lower engagement may be less likely to generate platform data or appear in follow-up measures. Schools with weaker implementation capacity may also produce less complete records.

If not properly addressed, missing data and attrition can distort impact estimates. The evaluation should report the amount and pattern of missing data, assess whether attrition differs between treatment and comparison groups, and explain how missingness was handled. Two kinds of attrition should be distinguished: attrition from the *study* (students not assessed at follow-up) and attrition from the *intervention* (students or teachers who stop using the programme).

Where possible, robustness checks should assess whether findings are sensitive to alternative assumptions, model specifications, samples or outcome definitions.

5 For missing data, attrition and robustness, teams should use WWC standards for attrition and baseline equivalence, ROBINS-I for missing data and risk of bias domains, and JBI tools for critical appraisal of quasi-experimental studies.

Guiding questions

- Who is missing from the data, and why?
- Does attrition differ between groups or relate to prior achievement or usage?
- Would the conclusions change under different assumptions about those who are missing?

What can go wrong?

The endline assessment is administered online. Students without connectivity at home or at school do not take it. The analysis, based only on complete cases, estimates the effect for the best-connected students and presents it as the programme's effect.

This standard addresses the risk that impact estimates are based only on the most visible, connected or engaged participants, while excluding those for whom the intervention may have worked differently or not at all. This risk is also linked to equity, since those who are systematically absent from the data because of limited connectivity, low attendance, or restricted access are often the same students for whom the intervention might work differently. Evaluators should therefore ask directly: who is missing from the data, and could their absence be masking a different pattern of effects?

06

Standard 6: Cost, affordability and sustainability [6]

An evaluation should support not only conclusions about whether an intervention generated effects, but also whether those effects are meaningful in relation to cost, affordability and sustainability. This is central for EdTech because costs often extend beyond the visible monetary value of devices or licences. They may include infrastructure, connectivity, maintenance, technical support, teacher training, coaching, data systems, platform updates, replacement cycles and opportunity costs.

Decision-makers need to know whether the observed effects justify the resources required to produce them, who bears each cost (provider, ministry, school or families), whether the programme can be sustained once external support ends, and whether scaling would change either the cost structure or the expected effect. A statistically significant effect may still be impractical if the cost per learner is too high, the implementation model is too dependent on external support, or the technology requires constant updating that schools cannot afford.

6 For cost, affordability and sustainability, teams should use EdTech Hub guidance on cost-effectiveness and decision-making, alongside EdTech quality frameworks that consider scalability, sustainability and fitness-for-purpose. Where possible, cost information should distinguish start-up, recurrent and system-level costs.

Guiding questions

- What does it cost per student to achieve this effect?
- Who pays for each component, and what is transferred to schools or families?
- What would it cost at scale, or once external support ends?
- What do comparison schools spend on similar activities?

What can go wrong?

An evaluation reports a licence cost of USD 8 per student per year and concludes that the programme is highly cost-effective. It omits connectivity, device replacement every three years and the coaching visits that schools say were essential. The real cost is several times higher.

This standard addresses the risk of interpreting effectiveness without considering whether the intervention is financially viable, scalable or preferable to alternative uses of resources. The question also asks for the relative value (opportunity cost): does the intervention produce more value than other pedagogical, technological or hybrid alternatives available? Asking whether an intervention is affordable is a critical question; evaluators should also ask what costs are transferred to schools, teachers or families, and how those costs change as the intervention scales. The standard's role is to keep this question open and visible, not to settle it in advance.

07

Standard 7: Transparency, reporting and interpretability [7]

A QED should report methods, assumptions, deviations, limitations and findings in a way that allows others to understand and scrutinise the basis of the impact claim. Transparency is especially important in EdTech because evaluations may rely on proprietary data, provider-generated metrics, complex algorithms or evolving platforms. Without clear reporting, decision-makers cannot judge whether findings are credible, comparable or relevant to their context.

Transparent reporting should include the intervention model, eligibility criteria, data sources, comparison strategy, baseline equivalence, outcomes, missing data, analytical decisions, implementation variation, limitations and conflicts of interest. It should also distinguish clearly between confirmatory and exploratory analyses, and between causal findings, associations and implementation insights. Transparency here should also extend to data independence and provider power: visibility over changes in algorithms or platform versions, awareness of

7 For transparency and reporting, teams should use TREND for reporting non-randomised evaluations, WWC and ROBINS-I for reporting causal assumptions and risks of bias, and EdTech-specific frameworks when reporting provider involvement, proprietary data, platform-generated metrics or product changes.

bias embedded in indicators, and disclosure of conflicts of interest between those who analyse and those who report results.

This standard addresses the risk that positive findings are overstated, selectively reported or difficult to interpret. It also supports evidence accumulation by making future synthesis, replication or adaptation more feasible.

Guiding questions

- Could an independent reader trace how the impact claim was built?
- Was the analysis plan documented before the results were known?
- Are exploratory findings clearly labelled as such?
- Who funded, implemented, analysed and reported the evaluation?

What can go wrong?

An evaluation measures the programme's effects on mathematics, reading and attendance. Only mathematics improves. The executive summary states that "the programme improves learning" and does not mention reading or attendance. Decision-makers conclude that the programme works across the board.

08

Standard 8: Transferability and progressive strengthening of evidence [8]

A single evaluation rarely settles whether an EdTech intervention works. More importantly for decision-makers, an effect estimated in one setting is not a permanent property of a programme: it holds under the conditions in which it was produced. This standard therefore asks evaluations to map those conditions (context, platform and content version, dose, population, implementation model and period) so that others can judge whether the findings are likely to hold in their own setting, and to specify what evidence should come next.

QED evidence should be interpreted as part of a broader process of evidence strengthening. In many EdTech contexts, especially where evidence ecosystems are still developing, the first evaluation may not provide definitive answers. However, it can identify promising effects, relevant implementation conditions, data gaps and priorities for future evaluation.

Evidence can be strengthened over time. A first QED may use available administrative data and provide preliminary causal evidence. A second study may improve measurement, strengthen comparison conditions or include implementation and cost evidence. Later evaluations may test mechanisms, subgroup

8 For progressive evidence strengthening, teams should use EdTech evidence maturity frameworks such as the Consolidated Benchmark/EVER (Kucirkova et al., 2024), 3E-Framework and WiKIT to position current evidence, identify what remains uncertain and define the next stage of evidence generation or synthesis.

effects, sustainability or scalability. Because EdTech changes rapidly, evidence also needs updating when the platform, content or delivery model changes. The goal is not to wait for perfect evidence, but to build a cumulative and transparent learning process.

Guiding questions

- Under what conditions was this effect observed?
- Do those conditions hold in my context?
- How confident can we be, and what remains uncertain?
- What evidence should be produced next, and when should it be updated?

What can go wrong?

A ministry scales a programme nationally on the basis of a positive pilot in one municipality, without documenting that the pilot had smaller classes, a dedicated coordinator and reliable connectivity. None of these conditions hold at national scale.

These eight standards are intended to support better evaluation decisions before, during and after a QED. Before the evaluation, they help define what must be specified and documented. During the evaluation, they help identify risks that may weaken interpretation. After the evaluation, they help decision-makers judge what can be concluded, what remains uncertain and what evidence should be strengthened next.

04

What these standards can and cannot solve

4.1

What these standards strengthen

The standards proposed in this framework are intended to improve the quality of QED evidence in EdTech by making impact claims more credible, transparent and useful for decision-making. Their main contribution is to make explicit the conditions under which a QED can support a responsible impact claim. This is particularly important in EdTech, where interventions often combine multiple components, evolve during implementation and generate data that may be useful for monitoring but insufficient for causal inference. By requiring evaluators to document intervention logic, comparability, measurement, implementation, missing data, cost and reporting decisions, the standards help reduce the risk of overclaiming results or interpreting weak evidence as definitive proof.

The standards also support better communication between who commission, conduct and use evaluations. Instead of presenting results only as "positive", "negative" or "not significant", they encourage a more precise interpretation of what can be concluded, what remains uncertain and what evidence should be strengthened before scaling, procurement or disinvestment decisions are made.

4.2

What these standards do not replace

These standards do not replace the need for strong evaluation design, independent data, technical expertise or institutional capacity. They also do not turn weak data into strong evidence. A QED remains vulnerable to bias when comparison groups are poorly constructed, baseline data are limited, attrition is high, outcomes are weakly measured or relevant confounders cannot be observed.

Nor do these standards replace experimental designs when randomisation is feasible, ethical and useful for the decision at hand. Randomised controlled trials remain an important source of causal evidence. The purpose of this report is not to present QEDs as equivalent to RCTs in all circumstances, but to clarify how quasi-experimental evidence can be strengthened when randomisation is not available or when programmes are already operating at scale.

The standards also cannot resolve all challenges associated with EdTech markets and governance. They can make provider involvement, proprietary data, cost assumptions and platform changes more visible, but they cannot by themselves eliminate conflicts of interest, guarantee data access or ensure that evidence will be used in decision-making. These issues require institutional arrangements, procurement rules, data governance agreements and independent review mechanisms.

4.3

Using standards wisely

The standards should be applied proportionally to the maturity of the intervention, the decision at stake and the quality of available data. A pilot evaluation, a national procurement decision and a regional scaling strategy do not require

the same level of evidence, but all require clarity about what can and cannot be concluded.

Table 2. Matching evidence to the decision at stake:

Decision at stake	Evidence typically needed
Improve the intervention	Implementation evidence and intermediate outcomes
Continue funding it	Evidence of plausibility, use and early outcomes
Scale it	Robust evidence on effects, costs, heterogeneity and sustainability
Discontinue it	Clarity on whether the problem lies in design, implementation or context

Illustrative; the evidence required should be judged case by case.

The priorities also shift with programme maturity:

For early-stage interventions, the priority is to clarify the theory of change, test feasibility, document implementation and identify plausible outcomes.

For programmes already operating at scale, it is to strengthen causal inference, examine heterogeneity, estimate cost-effectiveness and assess sustainability.

For mature interventions, it is to update evidence, test adaptation across contexts and check whether effects remain stable over time.

This proportional approach avoids two risks:

Lowering standards simply because contexts are resource-constrained.

Applying standards so rigidly that useful evidence is dismissed because it does not meet ideal conditions.

The purpose of actionable quality standards is to support better judgement: to identify what is credible enough for a given decision, what should be interpreted cautiously and what requires further evidence before stronger claims can be made.

4.4

Implications for evidence ecosystems

The tensions described in this framework are particularly acute in constrained evidence systems: systems in which the organisational and technical infrastructure for generating, synthesising and using evidence remains underdeveloped.

These ecosystems are characterised by fragmented administrative data, limited technical capacity for data analysis, weak evaluative independence, low demand for evidence and the absence of institutional mechanisms to synthesise and use it (Suazo-Galdames et al., 2025; Gough et al., 2020).

In contexts like these, the problem is not only how to produce more evidence, but how to produce rigorous and relevant evidence: evidence that informs policy and school decisions, supports programme improvement and, ultimately, contributes to improving learning.

Improving the quality of EdTech QEDs is not only a technical task. It is also part of strengthening evidence ecosystems. Better evaluations require better data systems, clearer intervention documentation, transparent reporting, independent review, synthesis capacity and stronger links between evidence producers and decision-makers.

The goal, therefore, should not be to produce isolated evaluations, but to build cumulative learning infrastructure. Each evaluation should leave behind clearer data, better documentation, stronger assumptions, more transparent limitations and a more precise agenda for future evidence generation. Over time, this can help education systems move from fragmented claims about what works to a more systematic understanding of what works, for whom, under what conditions, at what cost and with what level of certainty.

In this sense, the standards proposed in this report should be understood as a practical contribution to responsible EdTech evaluation. They do not eliminate uncertainty, but they help make uncertainty visible. They do not guarantee impact, but they strengthen the conditions for interpreting impact claims. They support a more disciplined use of evidence in decisions about whether to continue, adapt, scale or stop EdTech interventions. Ultimately, the aim is not only to improve individual evaluations, but to raise the quality of evidence available to sustain decision-making in EdTech. This means focusing on what evaluation can realistically offer, rather than dwelling on the methodological limitations of the contexts in which EdTech interventions unfold. It also means recognising a shared commitment: to build evidence that is public, cumulative, and oriented towards equity, learning and sustainability.

05. References

Barker, T. H., Habibi, N., Aromataris, E., Stone, J. C., Leonardi-Bee, J., Sears, K., Hasanoff, S., Klugar, M., Tufanaru, C., Moola, S., & Munn, Z. (2024). The revised JBI critical appraisal tool for the assessment of risk of bias for quasi-experimental studies. JBI Evidence Synthesis, 22(3), 378–388. https://doi.org/10.11124/JBIES-23-00268

Betthäuser, B. A., Bach-Mortensen, A. M., & Engzell, P. (2023). A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemic. Nature Human Behaviour, 7, 375–385. https://doi.org/10.1038/s41562-022-01506-4

Castillo-Canales, D., Mejías, L., Roque-Gutierrez, E., Valentini, A., & Rüebcke, J. (2023). Ed-Tech landscape and challenges in Latin America and the Caribbean (Occasional Paper No. 90). Southern Voice.

Castillo-Canales, D. Sepúlveda, M.J, Rosas Zambrano, N. & Caglevic, N. (2026). Educational technology for equity: Impact of ProFuturo in Latin America. ProFuturo Foundation & SUMMA Foundation.

Cueto, S., Balarin, M., Saavedra, M., & Sugimaru, C. (2023). Ed-tech in the Global South: Research gaps and opportunities (Occasional Paper Series No. 91). Southern Voice.

Damschroder et al. (2009): Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. Implementation Science, 4, Article 50. https://doi.org/10.1186/1748-5908-4-50

Des Jarlais, D. C., Lyles, C., Crepaz, N., & Trend Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. American journal of public health, 94(3), 361–366.

EdTech (2026). A Strategic Toolkit for EdTech Policy and Decision-Making. Available on the EdTech Hub website at: https://edtechhub.org/case-study/strategic-toolkit-for-edtech-policy-and-decision-making/

Education Endowment Foundation (EEF). (2026). Guidance on Theory of Change and logic model development for EEF-funded evaluations.

Foster, D., McLemore, C., Olszewski, B., Chaudhry, A., Cooper, E., Forcier, L., & Luckin, R. (2023). EdTech Quality Frameworks and Standards Review. Department for Education, UK.

Garg, M., Bakker, T.C. (2025). Evidence-informed Evaluation of EdTech (3E): A Dutch Framework for Continuous Assessment of EdTech Effectiveness. (pp. 1–15). Npuls. doi.org/10.5281/zenodo.15070789

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch. 2016. Impact Evaluation in Practice, second edition. Washington, DC: Inter-American Development Bank and World Bank. doi:10.1596/978-1-4648-0779-4.

Gough D, Davies P, Jamtvedt G, Langlois E, Littell J, Lotfi T, Masset E, Merlin T, Pullin AS, Ritskes-Hoitinga M, Røttingen JA, Sena E, Stewart R, Tovey D, White H, Yost J, Lund H, Grimshaw J. Evidence Synthesis International (ESI): Position Statement. Syst Rev. 2020 Jul 10;9(1):155. doi: 10.1186/s13643-020-01415-5. PMID: 32650823; PMCID: PMC7353688.

Haβler, B., Huntington, B., Klune, C., Lester, J., Bhutoria, A., & Mansour, H. (2025). Understanding Quality Characteristics of EdTech Interventions and Implementation for Disadvantaged Pupils - Systematic Review with Meta-Analysis. Open Development & Education, Education Endowment Foundation. https://doi.org/10.53832/opendeved.1170

Kucirkova, N. I., Cermakova, A. L., & Vackova, P. (2024). Consolidated Benchmark for Efficacy and Effectiveness Frameworks in EdTech. University of Stavanger. https://doi.org/10.31265/usps.270

Moore, D., Proctor, R., Benham-Clarke, S., Gains, H., Melendez-Torres, G. J., Axford, N., ... & Lloyd, J. (2024). Review of Evidence on Implementation in Education. Evidence Review. Education Endowment Foundation.

Plaut, D. (2024). How can decision-makers assess EdTech interventions for cost-effectiveness to enable better investments? EdTech Hub. https://doi.org/10.53832/edtechhub.1018

Scher, L., Kisker, E., & Dynarski, M. (2015). Designing and Conducting Strong Quasi-Experiments in Education. Version 2. Decision Information Resources, Inc.

Sharples, J., Eaton, J., & Boughelaf, J. (2024). A School's Guide to Implementation. Guidance Report. Education Endowment Foundation.

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., Ramsay, C. R., … Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ, 355, i4919. https://doi.org/10.1136/bmj.i4919

Suazo-Galdames IC, Saracostti M and Chaple-Gil AM (2025) Scientific evidence and public policy: a systematic review of barriers and enablers for evidence-informed decision-making. Front. Commun. 10:1632305. doi: 10.3389/fcomm.2025.1632305

SUMMA (2024). Evaluar para Mejorar: Examinando el Impacto de Innovaciones Educativas en América Latina. www.summaedu.org

UNESCO. (2023). Global Education Monitoring Report 2023: Technology in education: A tool on whose terms? UNESCO. https://doi.org/10.54676/UZQV8501

Waddington, H., Aloe, A., Becker, B., Reeves, B., Tugwell, P., & Wells, G. A. (2014). Risk of bias assessment in credible quasi-experimental studies. JCE series.

West, M. (2023). An ed-tech tragedy? Educational technologies and school closures in the time of COVID-19. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386701

What Works Clearinghouse. (2022). What Works Clearinghouse procedures and standards handbook, version 5.0. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). This report is available on the What Works Clearinghouse website at https://ies.ed.gov/ncee/wwc/Handbooks.

World Bank. (2022). The state of global learning poverty: 2022 update. World Bank.

Annex 1. Quasi-experimental designs: key assumptions and how to check them

Design	Key assumption	How to check it	Data needed
Matching	No unobserved differences that affect both participation and outcomes (cannot be tested directly)	Balance on baseline covariates; overlap in propensity scores; robustness to alternative matching methods	Rich pre-intervention data for both groups
Difference-in-differences	Without the programme, both groups would have followed parallel trends	Compare pre-intervention trends; placebo tests with a "false" treatment group or outcome	At least two pre-intervention measurements plus follow-up
Regression discontinuity	Units just above and below an eligibility threshold are comparable; the threshold is not manipulated	Density of the eligibility index around the threshold; no jump in baseline outcomes; sensitivity to bandwidth	A continuous eligibility index with a clear cut-off; effects are local to the threshold

Source: adapted from Gertler et al. (2016)

ProFuturo | Fundación Telefónica | "la Caixa" Foundation