# Consolidated Benchmark

# for Efficacy and Effectiveness Frameworks in EdTech

**Natalia Ingebretsen Kucirkova, Anna Lindroos Cermakova and Petra Vackova**

University of Stavanger, 2024

ISBN: 978-82-8439-209-7

**Suggested citation:**

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

**Acknowledgments:** The authors would like to thank Dr Anna Åkerfeldt, Stockholm University, Dina Foster, University College London, and Beth Havinga, European EdTech Alliance for useful comments and feedback on an earlier version of this report.

**Declaration of interest:** Professor Natalia Kucirkova is co-founder and CEO of WiKIT, which is cited in this report. WiKIT is a spin-off from the University of Stavanger and the university remains a shareholder through its Technology Transfer Office company, Validè AS.

## Abstract

Post-COVID19 evaluation reports of Educational Technologies (EdTech) pointed to the imperative to scientifically evaluate EdTech's impact on learners. The absence of a shared definition for such evaluations, coupled with the availability of diverse frameworks and criteria in the field, poses a challenge. This paper is concerned with two critical dimensions of impact on learning outcomes — efficacy and effectiveness — with a focus on teaching and learning EdTech for the K12 age range. A systematic literature search identified 65 frameworks that target the efficacy or effectiveness of K12 EdTech products. The frameworks were analysed in relation to their rigour, using the science of learning principles embedded in the EdTech Evidence Evaluation Routine (Kucirkova, Brod & Gaab, 2023). The results were synthesised into a consolidated benchmark that categorises the available frameworks at three levels based on the rigour applied to their assessments. The consolidated Effectiveness/Efficacy benchmark serves as a valuable tool for evaluating any EdTech type with available frameworks, facilitating informed decision-making in the dynamic landscape of educational technology.

*Keywords:* EdTech; apps; evidence; efficacy; effectiveness

# Introduction

In 2023, the converging message from academic and sector reports was that the majority of Educational Technology (EdTech) products lacked evidence of positive impact on learning. However, the criteria for determining what qualifies as appropriate evidence of impact in EdTech is still in the process of development. The aim of this report is to provide a consolidated benchmark, or a shared international standard, for ranking the evidence of EdTech's impact on learning outcomes.

EdTech encompass three major categories: 1) EdTech products for teaching and learning (including learning content, activities, assessment, and learning devices); 2) education governance (involving data integration, analytics, and management systems); and 3) employability and entrepreneurship (covering upskilling, reskilling, skills qualification, and career planning, see the Asian Development Bank, 2023). The focus of this report is on EdTech products for teaching and learning, such as apps, online platforms, and software tools, which were developed to target the formative years of education, i.e., the K-12 age range that in most countries covers children aged 3 to 18 years.

A clear definition and measurement of impact for the K12 EdTech sector is essential given that EdTech products in this category play a crucial role in addressing fundamental learning outcomes such as literacy and maths skills, and thus play a key role in children's education. Additionally, the K12 EdTech sector is a significant target for investment from private, governmental, and philanthropic investors, with an anticipated total investment reaching USD 132.4 billion by 2032 (Global Newswire, 2023). As education undergoes swift global digitization, there is an urgent need for a collective understanding and evaluation of "what works" in terms of which type of EdTech use works for which type of children's learning outcomes.

## What is evidence?

Simply defined, evidence is "the facts, signs, or objects that make you believe that something is true evidence (of something)" (Oxford Dictionary, online). In the EdTech context, 'evidence' has diverse interpretations, encompassing teachers' views on a technology's impact on the learners in their classrooms, parental perspectives on the technology's usefulness for their children, or researchers' formal assessments of statistically significant effects. These variations in definition mirror the distinct objectives of various stakeholders involved in EdTech, and all stakeholders play a crucial role in contributing to the formulation of effective practices.  In research terms, 'evidence' is understood as scientific proof, defined by various metrics that indicate the strength of the proof.  For example, the ESSA's definition of 'evidence' in the US government's guiding documents to research partners and schools, states that: "…the term 'evidence based,' when used with respect to a State, local educational agency, or school activity, means an activity, strategy, or intervention that demonstrates a statistically significant effect on improving student outcomes or other relevant outcomes" (from section 8101(21)(A) of the ESEA, US Government, 2022). This report is concerned with evidence defined as a measurable proof of EdTech's impact on learning and teaching (education).

## The current evidence base of EdTech's impact

The widespread adoption of EdTech during the COVID-19 pandemic prompted a comprehensive examination of its impact on children's learning. This scrutiny involved synthesising various sources, including teachers' reviews, scholarly articles, industry reports, and independent reports from philanthropic or other sector organisations. An example of such a summative report is the Grunndig report (Munthe et al., 2022), commissioned by the Norwegian government in 2022. Munthe et al. (2022) analysed 262 systematic reviews published between 2019 and 2022, synthesising research on the integration of digital tools in education, mainly focusing on mathematics, science, and languages. The findings highlighted potential benefits for students' learning, fostering creativity, critical thinking, and self-regulated learning. However, the report also highlighted studies that showed no or negative impact of EdTech on students' learning or teachers' instructional practices.

The findings from the UNESCO GEM report, commissioned by UNESCO in 2023, mirror the findings, although criticise more directly the low impact EdTech has had on learners globally. Based on a comprehensive review of academic literature, the report identified a scarcity of reliable and impartial evidence regarding the impact of EdTech, emphasising the limited robust evidence on the added value of digital technology in education. This limitation was primarily attributed to the rapid evolution of technology outpacing the pace of evaluation. The majority of available evidence originates from affluent countries, with a notable absence of randomised controlled trials (RCT) and third-party certification in the education technology sector (UNESCO, 2023). Furthermore, the report highlights that most educators in highly digitised countries, such as the USA, do not seek peer-reviewed evidence before adopting EdTech in classrooms, and a significant proportion of evidence is being generated by companies seeking to promote their products (UNESCO, 2023).

Although the findings from the UNESCO GEM report were highly cited in national media and prompted several governments to conduct inquiries into the usefulness of technologies or screens more broadly (e.g., ‘Screen Time: Impacts on education and wellbeing’ commissioned by the Education Committee United Kingdom Parliament, November 2023), the findings were not new. Indeed, several researchers’ investigations of popular EdTech have highlighted the low quality of commercial EdTech products before. For example, a study in 2015 of the most popular apps used by US children showed that the apps are misaligned with the principles of learning sciences (Hirsh-Pasek et al., 2015). Additionally, an analysis of the most popular children’s digital books in four European countries reported that digital books, instead of teaching children to read, might hamper their native language development, as they contain content that is not culturally appropriate and only comes in US English rather than in local languages (Sari et al., 2019). It is not only the learning outcomes that have been evaluated by researchers: a recent study by Mallawaarachchi et al. (2023) investigated 132 apps designed for toddlers and preschoolers, with a particular focus on the apps’ persuasive design features. The study revealed that most of these apps, despite being widely popular among the target age group, incorporated persuasive features, referred to as “dark design,” characterised by deceptive practices aimed at manipulating users.

While some EdTech tools, especially those emerging from research labs and spun out by scientists, have demonstrated impact (see, e.g., Wang & Tahir, 2020, on accumulative evidence on the quiz game Kahoot! on learners globally, or McTigue et al.’s, 2020, systematic review concerning the impact of the reading app GraphoGame), others either lack systematic evaluation by researchers before scaling to classrooms or, when evaluated, exhibit low or concerning quality.  And yet, there are 567,000 apps labelled as “educational” on the Apple App Store and Google Play Stores (as of November 2023) without any systematic scientific verification of their impact on students’ learning or educators’ teaching.

Furthermore, the global investment in EdTech, estimated at $123.40 billion in 2022 (Grand View Research, 2023), appears disproportionately high when compared to other pressing needs (e.g., the investment required for universal access to water, sanitation, and hygiene by 2030 is approximately USD 114 billion per year, see De Albuquerque, 2021). The situation raises the policy requirement to align the substantial investment and widespread adoption of EdTech with a thorough documentation of EdTech's impact on the educational system and in turn, children's learning.

Accordingly, major international organisations, such as the World Bank, political figures, including Ministers from several EU countries, philanthropic organisations, such as The Jacobs Foundation, have called for global efforts to address the mismatch between large EdTech's investments and EdTech's limited evidence-based support. Researchers' calls for thorough science-based evaluations of EdTech and application of evidence standards in public schools intensified in 2023, especially with the increased use of generative AI in K12 EdTech solutions (Kucirkova, 2023c). The calls of the sector and scientists are united in recognising the undisputable value effective technology use can have for education *if* the EdTech's design and implementation are grounded in research-based evidence of impact.

The response to the pressing demand to strengthen the evidence base for EdTech involves collaborations between industry and academia, wherein research partners assess EdTech solutions in classrooms or independently evaluate the reported impact by technology companies. Crucially, the success of these initiatives relies on establishing a common understanding of how to measure and evaluate evidence in the EdTech landscape. Here, the EdTech field, and education interventions more broadly, lack a consensus.

## Current frameworks for measuring EdTech evidence

Perhaps surprisingly, considering the low evidence of EdTech's impact, there is no shortage of frameworks designed to evaluate and assess various aspects of its impact. These frameworks have been developed by diverse groups, including experts, researchers, scientists, and independent clearinghouses. Wadhwa, Zheng, and Cook (2023) conducted a review of the evidence criteria used by 12 U.S. clearinghouses that rate the effectiveness of educational programs, and revealed significant inconsistencies across the 12 evaluation frameworks. Different clearinghouses reached varied conclusions regarding the effectiveness of the same educational program. Whether EdTech is adversely affected or benefited by the lack of standards (Kucirkova, 2023a) remains uncertain. What is certain is that the various frameworks targeting EdTech impact need to be consolidated.

Two recent efforts aimed to consolidate all available EdTech frameworks for evaluating teaching and learning tools. Vanbecelaere and colleagues (2023) reviewed the frameworks available for assessing the quality of EdTech suitable for use in classrooms and provided examples of frameworks that are designed to measure whether EdTech improves learning outcomes and learning experience or those that evaluate whether the EdTech is financially viable and scalable. In addition, the review identified various frameworks that evaluate whether EdTech leads to more effective teaching approaches or increases stakeholder collaboration (Vanbecelaere et al., 2023).

Foster et al. (2023) conducted a rapid review to identify frameworks referenced in academic journals and conference proceedings using the search string "EdTech" AND "frameworks" OR "standards" in the Google Search engine. In addition, frameworks identified for inclusion were discovered through additional reading, particularly of grey literature, and through colleague recommendation. This search produced 171 results, which were screened for relevance and resulted in the total number of 74 currently available EdTech evidence frameworks. Foster et al. (2023) outlined frameworks that specify standards and teacher competencies to align EdTech to pedagogy, as well as standards relevant for inclusive learning and interoperability of technologies. Several frameworks were designed to support professional development training, observation protocols, and principles for digital development in EdTech. In addition, frameworks that specify dimensions of personalisation in technology-enhanced learning are used by educators, as are frameworks that specify criteria for the digital competence of educators.

Foster et al. (2023) categorised the main characteristics of the 74 reviewed frameworks according to four dimensions:

a) frameworks that provide an analysis of quality components of an EdTech design;

b) frameworks that focus on how an EdTech meets users' needs;

c) frameworks that evaluate how digital pedagogy/digital competences are enabled;

d) frameworks that allow teachers to determine whether an EdTech product adopts an evidence-informed approach and assess the **quality of evidence** behind the EdTech product.

It is the latter - *Evidence Quality* Frameworks – that are at the forefront of current governments' priorities as they seek answers to questions posed by the UNESCO GEM report regarding "what works" in EdTech.

The Evidence Quality Frameworks tap into various types of impact that an EdTech product can have on users, including social or learning impact. Kucirkova (2023b) proposed that there are five key impact dimensions for understanding evidence quality in EdTech: Efficacy, Effectiveness, Ethics, Equity and Environmental outcomes, the so-called “5Es”. The five dimensions align with key metrics employed by investors, philanthropic organisations, and governments when assessing EdTech resources. Among the five Es dimensions, evidence of ‘Efficacy and Effectiveness’ directly address the extent to which technology impacts children's learning outcomes. The EdTech Evidence Quality Frameworks that are directly focused on the efficacy and effectiveness dimensions were not reviewed before and are the subject of the current review. The research questions (RQs) were:

RQ1: What existing frameworks assess the effectiveness and efficacy of EdTech for the K12 age range?

RQ2: In what way can the current frameworks evaluating EdTech effectiveness and efficacy be consolidated, considering their rigour?

The aim was to ascertain which evidence quality frameworks are available and to synthesise the frameworks with a systematic evaluation in order to develop a benchmark that can serve as a reference “umbrella framework” for the available evaluation tools.

## Efficacy and effectiveness: note on terminology

Although the terms ‘effectiveness’ and ‘efficacy’ are sometimes used interchangeably in educational research, a clear distinction exists in research, delineating the two types of evidence. Efficacy relates to “the performance of an intervention under ideal and controlled circumstances, whereas effectiveness refers to its performance under ‘real-world’ conditions” (Singal, Higgins, & Waljee, 2014, p. 45). Thus, both effectiveness and efficacy studies provide useful insights into how an EdTech works in relation to specified learning outcomes.

However, they need to be distinguished so that accurate and objective evaluations can be made. The points of differentiation in intervention studies have been around the eligibility criteria for enrolling participants, the participants' possibility to influence the intervention (degrees of control), as well as handling missing data and employment of specific statistical tests (Streiner, 2002). It is this distinction that was followed in the present systematic review.

# Methodology

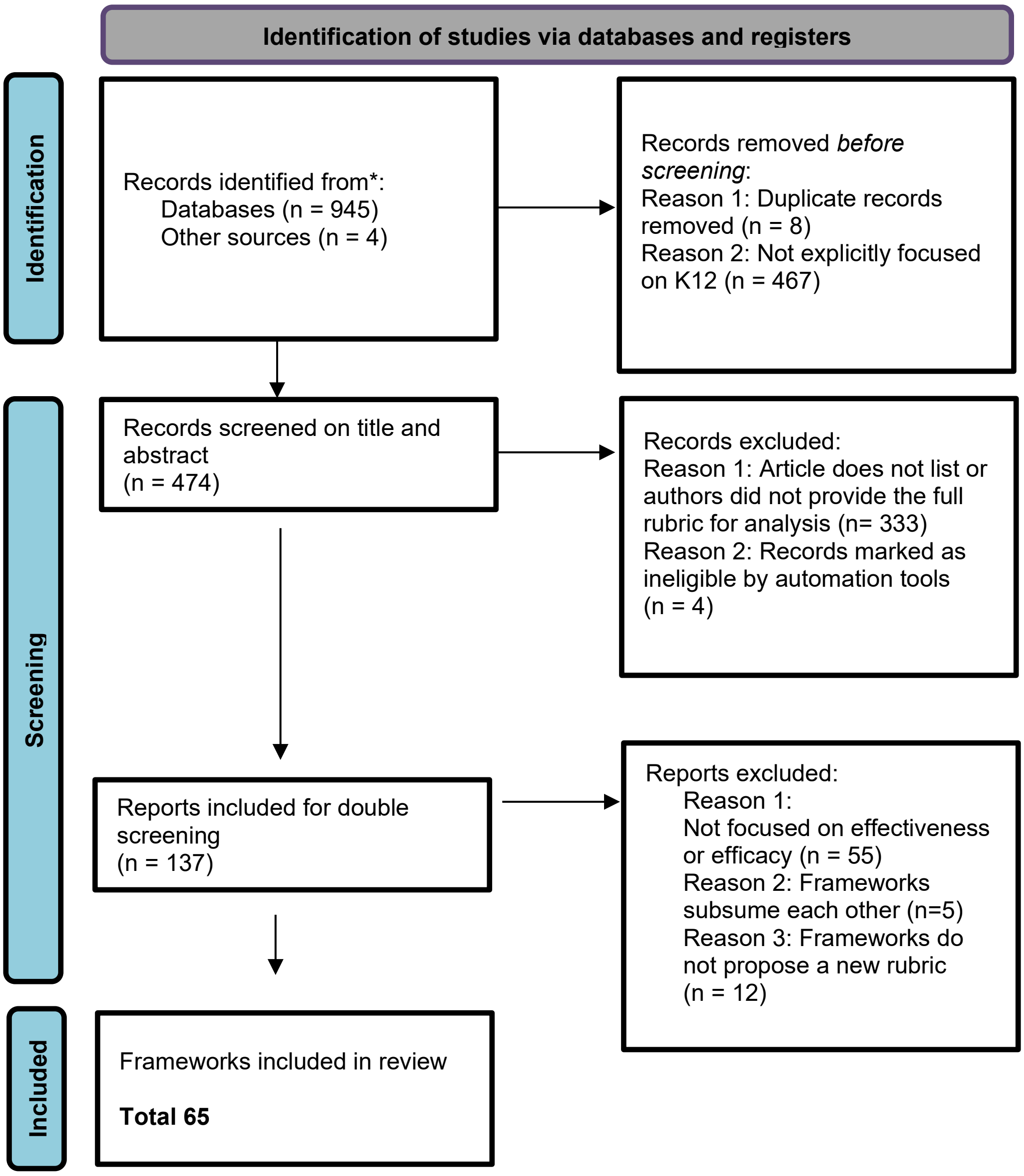
## Systematic review

The focus of our systematic review was on Evidence Quality EdTech frameworks as discussed above. Some are relevant for interventions that use EdTech products, some are pertinent to EdTech products themselves, and some do not make a distinction between the two. The Foster et al.'s (2023) review captures the main EdTech frameworks available up to date. Foster et al.'s (2023) search focused on "EdTech" as its main keyword, without including additional keywords for learning and teaching EdTech tools, such as apps or learning platforms. Therefore, to supplement the sample of frameworks, a rapid evidence review of available frameworks was conducted. To this end, a similar procedure to that by Foster et al. (2023) was followed with the keywords "frameworks" and "standards" used in the main search string while expanding from "EdTech" to "apps OR platforms OR learning technology OR educational media". Furthermore, the search was performed not only in Google Scholar but in major academic electronic databases such as Web of Science, PubMed, Ovid, Medline, APA PsycInfo, and Scopus and studies published anytime since 1940 (not limited to the last twenty years as in Foster et al., 2023). The search encompassed titles, abstracts, topics, tables of contents, and key terms, published in the English language. Additionally, a manual search of policy-level and commercial frameworks broadened the review ensuring a cross-sectoral scope of the literature review. The results in terms of the number of papers obtained and the reasons for inclusion/exclusion at each stage of analysis are presented in the PRISMA (Page et al., 2021) diagram in Figure 1.

Frameworks were excluded if the research articles lacked a focus on children within the K12 age range. Additionally, exclusion criteria applied if the framework's implementation was not conducted by expert reviewers, researchers, or evaluators trained on the framework's criteria (this was assessed on a Yes/No basis). Furthermore, studies were excluded if the framework lacked a clear description of its development process. The number of frameworks was also reduced if the frameworks have, over time, been subsumed under one joint framework. Such cases were analysed as one joint framework and included:

- 'EdTech Developer's Guide', Office of Educational Technology, 2015 and 'EdTech Evidence Toolkit Office of Educational Technology', 2023, subsumed as one ESSA framework.
- 2012, EdSurge Product Index & Decision Guide- Research and Evidence, maps on and was analysed as Digital Promise Research-Based Design Product Certification.

***Figure 1**: PRISMA diagram representing studies identified through systematic search*


Identification of studies via databases and registers
Identification
Screening
Included
Records identified from*:
Databases (n = 945)
Other sources (n = 4)
Records removed *before screening*:
Reason 1: Duplicate records removed (n = 8)
Reason 2: Not explicitly focused on K12 (n = 467)
Records screened on title and abstract (n = 474)
Records excluded:
Reason 1: Article does not list or authors did not provide the full rubric for analysis (n= 333)
Reason 2: Records marked as ineligible by automation tools (n = 4)
Reports included for double screening (n = 137)
Reports excluded:
Reason 1:
Not focused on effectiveness or efficacy (n = 55)
Reason 2: Frameworks subsume each other (n=5)
Reason 3: Frameworks do not propose a new rubric (n = 12)
Frameworks included in review
Total 65

*From:* Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

- Evidence rating system, Evidence 4 Impact (now part of EE), 2017 and EEF Padlock rating in ‘Toolkit Guide EEF’, 2023 map on the What Worked Framework, which was adapted by EdTech Impact for the 'Researched Impact' vertical (see https://interventions.whatworked.education/edtech).

- 'ISTE Standards for educators, students, school leaders and coaches', ISTE 2018 was subsumed by a new version ISTE 2023 and the ISTE Seal.

## Methodology for arriving at a consolidated benchmark

Several quality assessment criteria exist, including those by Gough (2017), which provide a three-level (high, medium, and low) framework of criteria for appraising the quality and relevance of evidence. The 'Edtech Evaluation Evidence Routine' (EVER) proposed by Kucirkova, Brod, and Gaab (2023) considers quality of evidence specifically within EdTech interventions and studies. EVER is an evaluation tool grounded in the Science of Learning, developed to gauge the evidence level of an EdTech design, use, or intervention. Considering the pragmatic and ideological challenges with a hierarchical perception of evidence, the EVER model advocates for methodological plurality for assessing EdTech products. While for some EdTech products/solutions, a randomised controlled trial (RCT) might be appropriate, for others, such as a screening tool, for example, a predictive validity study would be more suitable (Kucirkova, Brod & Gaab, 2023). EVER is described as to be applicable to all types of EdTech aiming to improve children's learning or to modify learners' behaviour, as well as those that integrate assessment and intervention. As such, the EVER criteria were perceived as an appropriate assessment tool for EdTech frameworks, encompassing both qualitative and quantitative studies from an international and interdisciplinary perspective.

EVER relies on five criteria: methodological quality, outcomes strength, predictive value, generalizability, and ethics & transparency. These criteria are relevant for assessing external and internal validity in both qualitative and quantitative studies and are succinctly defined as follows:

- **Methodological Quality** evaluates the appropriateness, execution, description, and justification of evaluation methods and their results. It addresses questions about the rationale, effectiveness of the methodology, and the size of the target population.
- **Outcome Strength** measures the impact or predictive value of an EdTech, quantifying effects through significance measures or effect sizes. It answers questions about the extent of impact and tool accuracy, including sensitivity/specificity, validity, and classification accuracy.
- **Quantifying Predictive Value** involves assessing sensitivity, specificity, validity, and classification accuracy. It quantifies how effectively a tool can accurately distinguish between different groups or categories, such as those with or without learning difficulties.
- **Generalizability** involves extending research findings from a specific sample to the larger population.
- **Ethics & Transparency** encompasses ethical questions related to the design and purpose of an EdTech intervention, with emphasis on culturally-responsive approaches and the transparent use of participants' data.

The strength of each of the five aspects was assessed on a 0–5 point scale for each eligible framework. The detailed rubric for this scoring was based on the EVER rubric and examples are provided in the Appendix.

The individual scores of 0-5 were categorised into three levels of low, medium, and high based on Gough's (2017) 'Weight of Evidence'. Gough (2007) proposed four main questions that should be asked when evaluating the quality of data. The answers to the questions should be rated as high, medium, or low. The questions are:

1. Regarding the overall coherence and integrity of an individual study, ask: Taking account of all quality assessment issues, can the individual study findings be trusted?

2. Regarding the appropriateness of the given form of evidence, ask: What is the appropriateness of the research design and analysis for addressing the aims of the individual study?

3. Regarding the relevance of the evidence, ask: What is the relevance of the particular focus of the individual study for addressing its aims?

4. Regarding the overall judgement of the evidence, ask: Taking into account the quality of execution, appropriateness of the design and relevance of focus, what is the overall weight of evidence this individual study provides to answer its research questions?

To judge the answers to these questions in terms of high, medium, and low in the context of EdTech, the aforementioned five EVER criteria (methodological quality, outcomes strength, predictive value, generalizability, and ethics & transparency) based on Kucirkova, Brod, and Gaab (2023), were used.

Two researchers independently rated each framework after thoroughly reviewing the framework's available description against the 5-points-scale of EVER criteria. The ratings provided by the two coders were then combined, and the average was calculated to determine a consolidated score ranging from 0 to 5. These consolidated scores were categorised into three levels:

- Level 1, corresponding to EVER scores of 0-2;
- Level 2, corresponding to EVER scores of 3-4;
- Level 3, corresponding to EVER scores of 4-5.

There was an intentional overlap between the highest scores at Levels 2 and 3 because of the difficulty to establish an exact cut-off point for higher-ranking studies. National frameworks reflect this degree of flexibility in scoring based on the final study's rigour. For example, in the US ESSA standards, the distinction between Tier II and Tier I varies depending on the level of randomization; and the exact level will be contingent upon a peer-reviewed consensus of the detailed study plan.

Given that the EVER criteria encompass conceptual studies (which evaluate a specific EdTech product or intervention based on systematic criteria or relevant research studies but do not involve direct testing of the product with children), Level 1 was assigned to frameworks that conducted reviews of EdTech products without empirical testing, while Level 2 and Level 3 denote frameworks that involved direct empirical testing.

Finally, for frameworks at Level 2 and 3, consideration was given to whether the frameworks evaluated EdTech's effectiveness or efficacy by directly testing the products with children. This aspect was determined based on the criteria proposed by Singal, Higgins, and Waljee (2014), encompassing "study design, patient populations, intervention design, data analysis, and result reporting." Given a considerable conceptual overlap between these criteria and those proposed by Kucirkova, Brod, and Gaab (2023) for assessing the internal and external validity of EdTech studies, the frameworks scoring 3-5 on quantitative dimensions of the EVER routine were considered as meeting the criteria for rigorous efficacy studies.

# Findings

## Efficacy frameworks

The total number of EdTech Efficacy frameworks identified through the search was 65 (see the Prisma diagram for details). This number includes frameworks where efficacy was one dimension among several criteria as well as frameworks focused exclusively on various aspects of efficacy. In some cases, several dimensions within eligible frameworks were counted as one: for example, the Efficacy Framework by Barber and Rizvi (2013) contains the dimension relevant to ‘Quality of evidence’, which assesses ‘Comprehensiveness of evidence’, ‘Quality of evidence’ and ‘Application of evidence’, and all three dimensions need to be scored in order to determine the overall level of efficacy of an EdTech product.

Some frameworks were more specific than others in their descriptions of how their individual criteria apply to different dimensions of EdTech. For example, the evaluation rubric by Lee and Cherner (2015) contains 24 dimensions with specific examples for each, while other rubrics had only 3-4 criteria. Similarly for quantitatively oriented rubrics, some contained detailed descriptions for the calculation of effect sizes while others do not mention power calculations. For example, the ‘Queensland Standards of Evidence’ provide information about the extent of measured improvement with detailed scores for five levels from “unknown to very high”. It considers time-based comparison, comparing pre-test scores with post-test scores, group-based comparison comparing Group A test scores with Group B test scores and desired effect size (see https://altqed.qed.qld.gov.au/publications/management-and-frameworks/evidence-framework/foundations-evidence/standards-evidence). This is different from, for example, the ‘Nesta Standards of Evidence’, which provide a narrative description of individual levels, without specifying how causal evidence is measured in terms of effect sizes or expected sample sizes.

There were also considerable differences among the frameworks in relation to the accessibility of their language to educators or policy-makers. For example, the ‘Evaluation Taxonomy, Learning Assembly, 2017’ framework was explicitly designed for non-researchers and is highly non-technical in language, with efficacy described with the use of example questions that educators can ask EdTech providers. On the other hand, the ‘EdTech Standards of Evidence’ framework developed by What Worked contains scientific terminology and expected numbers for attrition, effect sizes, and sample sizes, to determine the EdTech’s causal impact on learners.

In terms of their types, the frameworks were categorised according to their primary audiences: 1) frameworks designed by, and intended for, researchers reported in peer-reviewed articles (academic frameworks); 2) frameworks designed for policy-makers and procurement teams available as guiding national or international frameworks; and 3) frameworks and rubrics developed by commercial or independent organisations for public-facing ratings/certifications of EdTech’s quality.

## 1. Research frameworks reported in peer-reviewed articles

The category with the highest number of frameworks was those developed by researchers. Two comprehensive articles offer summative reviews of all these frameworks: Mustaffa et al.

(2016) conducted a literature review of educational app evaluation rubrics in 2016, while Papadakis (2021) carried out a systematic review. Collectively, these and the present analysis show that researchers' frameworks have been developed to assess EdTech in relation to specific outcomes, such as language or phonemic awareness (e.g. Rosell-Aguilar, 2017) or particular interests of researchers (e.g., persuasive features examined by Mallawaarachchi et al., 2023). The researchers used the frameworks to assess the evidence of a selection of EdTechs, often the most popular apps from the App Store (e.g., Meyer et al., 2021), or to compare free and paid apps (e.g., Kolak et al., 2021). Some frameworks validated their scores against media ratings, while others relied on teachers' ratings, but only a few frameworks validated their assessments against children's actual use of the apps. For instance, 'MAD Learn' (Herodotou, 2021, May) was based on visualising the learning design and learning components of a given app, combined with an analysis of children's actual interactions with a selection of apps.

Most researchers' frameworks were developed with ideas and observations pertinent to children based in the Global North. In contrast, the framework by Huntington, Goulding, and Pitchford (2023) was developed by comparing apps that were used by children in Tanzanian villages in terms of their impact on the children's learning outcomes over a 15-month-long intervention. Huntington, Goulding, and Pitchford's (2023) framework thus specifies which features and combination thereof, are effective at supporting learning of children's literacy and mathematics skills for out-of-school children in low- and middle-income countries.

## 2. National and international frameworks

The mapping exercise showed that several pertinent frameworks are accessible from official national bodies or philanthropic organisations. Some of these frameworks are specifically designed for evaluating causal evidence of impact, for example the ‘Standards of Evidence’, Australian Education Research Organisation (AERO) 2021; ‘ESSA Tiers of Evidence’, US Gov, 2015; ‘Queensland Standards of Evidence’, Queensland Department of education, 2023 and ‘Nesta Standards of Evidence’, 2022, while others focus on the review of products by experts. For example, the ‘Tulna standards’, developed by academics and collaborators at the Central Square Foundation, assess quality along three dimensions: ‘Content Quality’, ‘Pedagogical Alignment’, and ‘Technology & Design’, with each dimension listed with a set of detailed criteria: https://www.edtechtulna.org/standards.

## 3. Efficacy frameworks developed by/for the EdTech industry

The ‘EdTech Impact Quality Framework’ by EdTech Impact Ltd. incorporates four assessment verticals, emphasising not only the impact of utilising a technology but also pedagogical criteria. The latter rely on the evaluation criteria of the Education Alliance Finland, now part of EdTech Impact. Other frameworks use calibrated criteria applied to various EdTech products by trained experts (e.g., Common Sense Media Review), or they focus on teachers’ and experts’ reviews of products based on proprietary rubrics. However, most certifying organisations do not publicly disclose their exact criteria for determining effectiveness or efficacy. This hindered the present EVER analysis and the inclusion of these frameworks in the consolidated benchmark. For example, given the lack of publicly available information on possible effectiveness/efficacy aspects of the ISTE Seal framework, the ISTE Seal was not included in the final set of frameworks.

## Consolidated Benchmark for the Efficacy Frameworks

Table 1 offers an overview of eligible frameworks along with their corresponding scores on EVER, divided by three levels characterised by increasing rigour from Level 1 to Level 3. Level 1 frameworks primarily focus on conceptual evidence and, to some extent, effectiveness studies. On the other hand, Level 2 and Level 3 frameworks are more oriented towards effectiveness and efficacy studies. High ratings on frameworks in the Level 3 category indicate the most rigorous quality of evidence provided on an EdTech product or intervention.

**Table 1.** Consolidated Benchmark of available EdTech Efficacy and Effectiveness Frameworks.

| **Framework** | **EVER Score 1-2** | **EVER Score 3-4** | **EVER Score 5** |
|---|---|---|---|
| Consolidated benchmark | **Level 1** | **Level 2** | **Level 3** |
| | Conceptual studies<br>Effectiveness | Effectiveness --- > Efficacy evidence | |
| **Certification and commercial frameworks** | | | |
| Digital Promise: ‘Research-Based Design Product Certification’ | Achieved Certification | N/A | N/A |
| WiKIT: ‘Evidence-ready’ | Achieved Impact | N/A | N/A |
| LearnPlatform: ‘IMPACT-Ready’ | Certification/<br>Badge | N/A | N/A |
| LeanLab: ‘Pilot Readiness Audit’ | Report/design/<br>recommendation | N/A | N/A |

| | | | |
|---|---|---|---|
| What Worked Education: 'EdTech Standards of Evidence' | Very limited and weak evidence | Moderate evidence | Strong evidence |
| Barber, M. and Rizvi, S (2013). *Efficacy Framework (Pearson)* Section 2 'Evidence': Comprehensiveness of evidence<br>Quality/Application of evidence | Amber/Green<br>Amber | Green<br>Amber/Green | N/A<br>Green |
| Learning Assembly: Evaluation Taxonomy, 'Efficacy' dimension | Positive rating on the framework<br>Note: the taxonomy is not detailed enough to distinguish between levels | | |
| **National and international standards of evidence** | | | |
| ESSA'Tiers of Evidence'<br>US Gov (2015) | Tier IV | Tier III and Tier II | Tier II and Tier I |
| AERO 'Standards of Evidence' | Level 2 | Level 2 and 3 | Level 4 |
| Queensland department of education: 'Queensland Standards of Evidence', dimension 'Impact' (2023) | Low | Moderate | High and Very high |
| NESTA Standards of Evidence (2020) | Level 1 and 2 | Level 2 and 3 | Level 3, 4 and 5 |
| **Academic frameworks** | | | |
| 47 peer-reviewed frameworks | Rating above a median score | N/A | N/A |
| Wang et al. (2019) | Level 1 study | N/A | N/A |
| Tahir, R., & Arif, F. (2014) | Level 1 study | N/A | N/A |
| Herodotou (2021) | Level 1 study | N/A | N/A |

| | | | |
|---|---|---|---|
| Huntington et al. (2021 & 2023) & XPrize (2019) | Huntington et al. ratings | XPrize ratings | N/A |
| Outhwaite et al. (2023) (for maths only) | Level 1 study | N/A | N/A |

# Discussion

Educators' collective effort to augment the evidence base of EdTech is progressing as an interdisciplinary field with various frameworks and rubrics used to establish "what works". As emphasised in a recent Special Issue focused on evidence of educational apps (Outhwaite & Van Herwegen, 2023), it is vital that the educational field advances the EdTech evidence base by exploring how apps can be designed to facilitate learning and how their implementation can support educational outcomes and educational endeavours. Recent survey results from the EEA European EdTech Map showed that 1,480 European startups indicated a large willingness on behalf of the edtech community to test - but their key responses, when asked about hurdles for testing, were access and funding (European EdTech Alliance, 2023). The industry and policy endeavours to test and evaluate EdTech can be further bolstered with a shared benchmark for EdTech effectiveness and efficacy.

This report provides a comprehensive mapping of available frameworks concerned with EdTech efficacy and effectiveness, showcasing their points of synergy across three levels of rigour. The consolidated benchmark is based on the latest insights from learning sciences to grade the rigour of evidence. A methodology similar to that followed in this report can be adopted for establishing consolidated benchmarks in other areas of evidence of impact, including ethics, equity, and environmental impact. Over time, new evaluation frameworks are likely to emerge, addressing specific outcomes targeted by EdTech products, such as mental wellbeing or the impact that students' use of generative AI might have on climate change. Simultaneously, there is a trend toward further consolidation in the EdTech field. This is evident as some frameworks are already being positioned as overarching others; for example, the EdSurge framework, under research and evidence, incorporates the 'Digital Promise Research-Based Design Product Certification'. Both future frameworks and ongoing consolidation in the field, will require a shared benchmark for gauging their rigour, as proposed in this paper.

Some ratings in the consolidated benchmark had to be approximate due to the current lack of sufficiently detailed descriptions in individual frameworks (e.g., the ISTE Seal) or the criteria of the frameworks not being publicly available. To enhance the accuracy of measurement and the positioning of individual frameworks on the benchmark, including their trustworthiness, it is crucial that both current and potential future frameworks transparently disclose the criteria they employ for different levels and scores within their frameworks.

Academic frameworks assessed in Table 1 deserve a further comment. 47 academic frameworks (see the full list in the Appendix) were concerned with either direct testing of children's engagement with a selection of apps or with a review of the apps' key features. While the academic frameworks are mostly placed at Level 1 in the consolidated benchmark, this is not to indicate that their methodological design is not rigorous, but that only the studies that achieve EVER score 3 and above for the 'Quantitative studies criteria' were placed at Level 2 and 3.

Some of these studies are conceptual (e.g. Outhwaite et al. 2023) and achieve high EVER scores in this respect (see Appendix 2). Some have a robust methodological design but only a small number of participants (e.g. Tahir & Arif, 2014; Herodotou, 2021). Others were concerned with adult participants (e.g. Wang et al., 2019; Huntigton et al., 2023). Other studies conceptualised efficacy through the eyes of teachers, although primarily relying on teachers' views and attitudes rather than their combination with learning outcomes data (e.g., Lubniewski et al., 2017; Papadakis et al., 2020; Vazquez-Camo et al., 2023). A group of studies had parents as the main respondent group (e.g., Urquahart et al., 2023; Vaiopoulou et al., 2021) or used students' self-reported assessments of learning (e.g., Lee & Kim 2015), which we considered problematic and did not count towards efficacy frameworks.

Some frameworks were excluded because they were out of scope, for example some academic frameworks provided robust ratings but they were not directly concerned with efficacy (e.g. Mallawaarachchi et al., 2023; Vaiopoulous et al., 2022). Missing sufficiently detailed information was another issue, for instance, Kay, Lesage and Tepylo (2019) and Huntington et al. (2021), were available only as short abstracts to us. The latter built on a large-scale X-Prize (2019) study, which likely could be placed at Level 3; however, the Executive summary available online did not provide enough information to make this judgement.

It is also important to underscore that although a consolidated benchmark can alleviate confusion in assessing effectiveness and efficacy, it doesn't entirely address the problem of limited evidence of impact within the EdTech sector. As for example the academic frameworks suggest, with most being placed at Level 1, large-scale testing needed as efficacy/effectiveness evidence at higher level is often out of reach for smaller research teams. For this, the systematic impact dimensions that influence the broader EdTech ecosystem remain the need for considerable investments into evidence (through governmental but also philanthropic and private-public partnerships) and regulation. Without a doubt, investments and national regulations must offer meaningful incentives for motivating the EdTech community to actively engage in producing evidence that can be evaluated across various impact categories.

Overall, the consolidated benchmark makes a valuable contribution to the EdTech field by charting frameworks that enable the assessment of whether the use of specific EdTech products and/or design features has a measurable and objectively verifiable impact on students' outcomes. These outcomes encompass not only learning outcomes, but also other dimensions targeted by EdTech, such as engagement, learning, and wellbeing. The identified frameworks can be utilised to formulate a straightforward rubric for EdTech companies to construct their impact metrics, which can then be more widely contextualised and transparently communicated to their funders and users.

# References

Asian Development Bank (2023). EdTech Product categorisation from Asian Development Bank 2022, adapted for Reimagine Tech-Inclusive Education: Evidence, Practices, and Road Map, Asian Development Bank, 2023.

Baloh, M., Zupanc, K., Kosir, D., Bosnić, Z., & Scepanović, S. (2015, June). A quality evaluation framework for mobile learning applications. *Proceedings of the 4th Mediterranean Conference on Embedded Computing, Budva*, Montenegro, 280–283. https://doi.org/10.1109/MECO.2015.7181923

Baran, E., Uygun, E., & Altan, T. (2017). Examining preservice teachers' criteria for evaluating educational mobile apps. *Journal of Educational Computing Research*, *54*(8), 1117–1141. https://doi.org/10.1177/0735633116649376

Barber, M. & Rizvi, S (2013). *Efficacy Framework: A Practical Approach to Improving Learner Outcomes*. Pearson.

Biesta, G., Wainwright, E., & Aldridge, D. (2022). A case for diversity in educational research and educational practice. *British Educational Research Journal*, *48*(1), 1–4. https://doi.org/10.1002/berj.3777

Barr, R., & Kirkorian, H. (2023). Reexamining models of early learning in the digital age: Applications for learning in the wild. *Journal of Applied Research in Memory and Cognition, 12*(4), 457–472. https://doi.org/10.1037/mac0000132

Bradley, J. (2012). Language-Learning Apps. (Product/service evaluation), *Wired* (San Francisco, Calif.), 20(10), p. 66.

Booton, S. A., Kolancali, P., & Murphy, V. A. (2023). Touchscreen apps for child creativity: An evaluation of creativity apps designed for young children. *Computers & Education*, 201, 104811.

Campbell, L. O., Gunter, G., & Braga, J. (2015). Utilizing the Retain Model to evaluate mobile learning applications. In D. Rutledge & D. Slykhuis (Eds.), *Proceedings of the Society for Information Technology & Teacher Education International Conference,* 732–736. Association for the Advancement of Computing in Education.

Chatzopoulos, A., Karaflis, A., Kalogiannakis, M., Tzerachoglou, A., Cheirchanteri, G., Sfyroera, E., & Sklavounou, E. O. (2023). Evaluation of Google Play educational apps for early childhood education. Advances in Mobile Learning Educational Research, 3(2), 770-778.

Chen, X. (2016). Evaluating language-learning mobile apps for second-language learners. *Journal of Educational Technology Development and Exchange, 9*(2). https://doi.org/10.18785/jetde.0902.03

Cherner, T., Dix, J., & Lee, C. (2014). Cleaning up that mess: A framework for classifying educational apps. *Contemporary Issues in Technology and Teacher Education, 14*(2), 158–193. https://citejournal.org/volume-14/issue-2-14/general/cleaning-up-that-mess-a-framework-for-classifying-educational-apps/

Cherner, T., Fegely, A., Lee, C. Y., & Santaniello, L. (2016). A detailed rubric for assessing the quality of teacher resource apps. *Journal of Information Technology Education: Innovations in Practice, 15*(1), 117–143. https://doi.org/10.28945/3527

De Albuquerque, C. (2021). Bridging the Financial Gap: Investing in SDG 6, SDG Knowledge Hub Blog, https://sdg.iisd.org/commentary/guest-articles/bridging-the-financial-gap-investing-in-sdg-6/

Dekker, I., & Meeter, M. (2022). Evidence-based education: Objections and future directions. *Frontiers in Education*, 7, p. 941410. https://doi.org/10.3389/feduc.2022.941410

Dore, R. A., Shirilla, M., Verdine, B. N., Zimmermann, L., Golinkoff, R. M., & Hirsh-Pasek, K. (2018). Developer meets developmentalist: improving industry–research partnerships in children's educational technology. *Journal of Children and Media*, *12*(2), 227-235.

Evidence for ESSA: Standards and Procedures, Version 2.1, November 2023, Evidence for ESSA, Available online from: www.evidenceforessa.org

Foster, D., McLemore, C., Olszewski, B., Chaudhry, A., Cooper, E., Forcier, L., & Luckin, R. (2023). EdTech Quality Frameworks and Standards Review. Department for Education, Published 14. December 2023, https://www.gov.uk/government/publications/edtech-quality-characteristics-frameworks-and-standards-review

Goodwin, K., & Kucirkova, N. (2012, March). iTouch and iLearn: An examination of educational apps. Paper presented at the Early Education and Technology for Children Conference, Salt Lake City, Utah, USA.

Global Newswire (2023). K-12 Education Technology Spend Market to hit USD 132.4 billion by 2032; Amid the growing demands for personalized and online learning, https://www.globenewswire.com/en/news-release/2023/10/17/2761545/0/en/K-12-Education-Technology-Spend-Market-to-hit-USD-132-4-billion-by-2032-Amid-the-growing-demands-for-personalized-and-online-learning.html

Gough. D. (2007). Weight of Evidence: a framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education, 22* (2), 213–228. https://doi.org/10.1080/02671520701296189

Grand View Research. (2023). Global EdTech Market Size, Share & Trends Analysis Report. https://www.grandviewresearch.com/industry-analysis/edtech-market

GrandView Research (2023). Education Technology Market Size, Share & Trends Analysis Report By Sector (Preschool, K-12, Higher Education), By End-user (Business, Consumer), By Type, By Deployment, By Region, And Segment Forecasts, 2023 – 2030. Report ID: GVR-4-68038-878-7 https://www.grandviewresearch.com/industry-analysis/education-technology-market

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal, 26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015). Putting education in "educational" apps: Lessons from the science of learning. *Psychological Science in the Public Interest, 16*(1), 3-34.

Herodotou, C. (2021, May). MAD learn: an evidence-based affordance framework to assessing learning apps. In 2021 7th International Conference of the Immersive Learning Research Network (iLRN), 1-8. IEEE.

Huntington, B., Goulding, J., & Pitchford, N. (2021). Transforming global learning with digital technologies: A qualitative exploration of the use of educational technology with marginalised, out-of-school children living in remote settings. In *EDULEARN21 proceedings,* p. 9201. IATED.

Huntington, B., Goulding, J., & Pitchford, N. J. (2023). Pedagogical features of interactive apps for effective learning of foundational skills. *British Journal of Educational Technology.*

Hussain, A., Mkpojiogu, E. O. C., & Hassan, F. (2018). Dimensions and sub-dimensions for the evaluation of m-learning apps for children: A review. *International Journal of Engineering and Technology, 7*(3.20), 291–295. https://doi.org/10.14419/ijet.v7i3.20.19168

Ibrahim, N.K. et al. (2019). Multi-Criteria Evaluation and Benchmarking for Young Learners' English Language Mobile Applications in Terms of LSRW Skills, IEEE access, 7, 146620–146651. Available at: https://doi.org/10.1109/ACCESS.2019.2941640.

Israelson, M. H. (2015). The app map: A tool for systematic evaluation of apps for early literacy learning. *Reading Teacher, 69*(3), 339–349. https://doi.org/10.1002/trtr.1414

Kalogiannakis, M., & Papadakis, S. (2017, August). An evaluation of Greek educational android apps for preschoolers. *Proceedings of the 12th Conference of the European Science Education Research Association*, Dublin Ireland, 593–603.

Kay, R. (2018a, March). Creating a framework for selecting and evaluating educational apps. *Proceedings of the 12th International Technology, Education and Development Conference*, Valencia, Spain, 374–382. https://doi.org/10.21125/inted.2018.0106

Kay, R. (2018b, October). Developing a framework to help educators select and use mobile apps in the classroom. *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, Las Vegas, NV, USA, 1315–1320.

Kay, R., Lesage, A., & Tepylo, D. (2019, November). Evaluating the learning, design and engagement value of mobile applications: The mobile app evaluation scale. *Proceedings of the 12th International Conference of Education, Research and Innovation*, Seville, Spain, 1103–1107. https://doi.org/10.21125/iceri.2019.0336

Khan, A. I., Al-Khanjari, Z., & Sarrab, M. (2017, April). Crowd sourced evaluation process for mobile learning application quality. *Proceedings of the 2nd International Conference on Information Systems Engineering,* Charleston, SC, USA. https://doi.org/10.1109/ICISE.2017.17

Kolak, J., Norgate, S. H., Monaghan, P., & Taylor, G. (2021). Developing evaluation tools for assessing the educational potential of apps for preschool children in the UK. *Journal of Children and Media, 15*(3), 410–430. https://doi.org/10.1080/17482798.2020.1844776

Konca, A. S., Izci, B., & Simsar, A. (2023). Evaluating popular STEM applications for young children. *European early childhood education research journal*, 1-17.

Kucirkova, N. (2018). A taxonomy and research framework for personalization in children's literacy apps. *Educational Media International, 55*(3), 255-272.

Kucirkova, N. (2023a). Are *EdTech* companies the *casualties* or winners of educational evidence wars? BERA Blogs, https://www.bera.ac.uk/blog/are-edtech-companies-the-casualties-or-winners-of-educational-evidence-wars

Kucirkova, N. (2023b). How can philanthropy catalyse a system-wide change in EdTech? *Alliance magazine*, https://www.alliancemagazine.org/blog/how-can-philanthropy-catalyse-a-system-wide-change-in-edtech/

Kucirkova, N. I. (2023c). Ethics: fund an independent system to verify EdTech. *Nature*, *618*(7966), 675-675.

Kucirkova, N., Brod, G., & Gaab, N. (2023). Applying the science of learning to EdTech evidence evaluations using the EdTech Evidence Evaluation Routine (EVER). *npj Science of Learning*, *8*(1), 35-42.

Lee, C. Y., & Cherner, T. S. (2015). A comprehensive evaluation rubric for assessing instructional apps. *Journal of Information Technology Education, 14*(1), 21–53. https://doi.org/10.28945/2097

Lee, J. S., & Kim, S. W. (2015). Validation of a tool evaluating educational apps for smart education. *Journal of Educational Computing Research, 52*(3), 435–450. https://doi.org/10.1177/0735633115571923

Lisenbee, P. S. (2018). Literacy app evaluation tool for teachers: Phonemic awareness and phonics apps rubric. https://plisenbee.wixsite.com/website/blank-page

Lubniewski, K. L., McArthur, C. L., & Harriott, W. A. (2017). Evaluating instructional apps using the app checklist for educators (ACE). *International Electronic Journal of Elementary Education, 10*(3), 323–329. https://doi.org/10.26822/iejee.2018336190

Lytras, M.D. et al. (2019). Evaluation of Mobile Apps for Chinese Language Learning. In *Cognitive Computing in Technology-Enhanced Learning*. IGI Global, 191–205. https://doi.org/10.4018/978-1-5225-9031-6.ch009.

Mallawaarachchi, S.R. et al. (2023). Persuasive design-related motivators, ability factors and prompts in early childhood apps: A content analysis, *Computers in Human Behavior*, 139, 107492–. https://doi.org/10.1016/j.chb.2022.107492.

Martín-Monje, E., Arús, J., Rodríguez-Arancón, P., & Calle-Martínez, C. (2014). REALL: Rubric for the evaluation of apps in language learning. *Proceedings of Jornadas Internacionales Tecnología Móvil e Innovación en el Aula*: *Nuevos Retos y Realidades Educativas.* https://www.researchgate.com/publication/255702557_REALL_Rubric_for_the_evaluation_of_apps_in_language_learning

McManis, L. D., & Parks, J. (2011). Evaluating technology for early learners. Hatch, Inc. https://www.eschoolnews.com/files/2012/01/EvaluatingTechnology_ebook_toolkit.pdf

McQuiggan, S., McQuiggan, J., Sabourin, J., & Kosturko, L. (2015). Mobile Learning: A Handbook for Developers, Educators, and Learners. Somerset: Wiley. https://doi.org/10.1002/9781118938942.

McQuiggan, S., McQuiggan, J., Sabourin, J., & Kosturko, L. (2015). The business of educational apps. In S. McQuiggan, J. McQuiggan, J. Sabourin, & L. Kosturko, *Mobile Learning: A Handbook for Developers, Educators, and Learners,* 215-235. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/10.1002/9781118938942.ch11

Meyer, M., Zosh, J. M., McLaren, C., Robb, M., McCaffery, H., Golinkoff, R. M., Hirsh-Pasek, K., & Radesky, J. (2021). How educational are "educational" apps for young children? App store content analysis using the Four Pillars of Learning framework. *Journal of Children and Media, 15*(4), 526-548. https://doi.org/10.1080/17482798.2021.1882516

Montazami, A. et al. (2022). Why this app? How educators choose a good educational app. *Computers and Education*, 184, p. 104513– https://doi.org/10.1016/j.compedu.2022.104513.

Montazami, A. et al. (2022). Why this app? How parents choose good educational apps from app stores. *British Journal of Educational Technology, 53*(6), 1766–1792. https://doi.org/10.1111/bjet.13213.

Munthe, E., Erstad, O., Njå, M.B., Forsström, S., Gilje, Ø., Amdam, S., Moltudal, S., & Hagen, S.B. (2022). Digitalisering i grunnopplæring; kunnskap, trender og framtidig forskningsbehov. Kunnskapssenter for utdanning: Universitetet i Stavanger.

Mustaffa, F. Y., Salam, A. R., Muhammad, F., Bunari, G., & Asary, L. H. (2016). Literature review of educational app evaluation rubrics. *Intervention in School and Clinic, 51*(4), 244–252.

Neumann, M., Wang, Y., Qi, G. Y., & Neumann, D. L. (2019). An evaluation of Mandarin learning apps designed for English speaking preschoolers. *Journal of Interactive Learning Research, 30*(2), 167–193.

Outhwaite, L. A., Early, E., Herodotou, C., & Van Herwegen, J. (2023). Understanding how educational maths apps can enhance learning: A content analysis and qualitative comparative analysis. *British Journal of Educational Technology*.

Outhwaite, L. A., & Van Herwegen, J. (2023). Educational apps and learning: Current evidence on design and evaluation. *British Journal of Educational Technology*, *54*(5), 1268-1272.

Papadakis, S. (2021). Tools for evaluating educational apps for young children: A systematic review of the literature. *Interactive Technology and Smart Education, 18*(1), 18-49. https://doi.org/10.1108/ITSE-08-2020-0127

Papadakis, S., Kalogiannakis, M., & Zaranis, N. (2017). Designing and creating an educational app rubric for preschool teachers. *Education and Information Technologies, 22*(6), 3147–3165. https://doi.org/10.1007/s10639-017-9579-0

Papadakis, S., Vaiopoulou, J., Kalogiannakis, M., & Stamovlasis, D. (2020). Developing and exploring an evaluation tool for educational apps (E.T.E.A.) targeting kindergarten children. *Sustainability, 12*(10), 4201. https://doi.org/10.3390/su12104201

Pearson, H.A., Montazami, A. and Dubé, A.K. (2022). Why this app: Can a video-based intervention help parents identify quality educational apps? *British Journal of Educational Technology* [Preprint]. https://doi.org/10.1111/bjet.13284.

Privitera, A. J., Ng, S. H. S., & Chen, S. H. A. (2023). Defining the Science of Learning: A Scoping Review. *Trends in Neuroscience and Education*, 100206.

Rosell-Aguilar, F. (2017). State of the app: A taxonomy and framework for evaluating language learning mobile applications. *CALICO Journal*, 34(2), 243–258. https://doi.org/10.1558/cj.27623

Sari, B., Takacs, Z. K., & Bus, A. G. (2019). What are we downloading for our children? Best-selling children’s apps in four European countries. *Journal of Early Childhood Literacy, 19*(4), 515-532.

Shahjad and Mustafa, K. (2022). A Systematic Literature Review on Learning Apps Evaluation, *Journal of Information Technology Education,* 21, 663–700. https://doi.org/10.28945/5042.

Shoukry, L., Sturm, C., & Galal-Edeen, G. H. (2015). Pre-MEGa: A proposed framework for the design and evaluation of preschoolers' mobile educational games. In T. Sobh, & K. Elleithy, K. (Eds.), *Innovations and A Systematic Literature Review on Learning Apps Evaluation Advances in Computing, Informatics, Systems Sciences, Networking and Engineering,* 385–390. Springer. https://doi.org/10.1007/978-3-319-06773-5_52

Singal, A. G., Higgins, P. D., & Waljee, A. K. (2014). A primer on effectiveness and efficacy trials. *Clinical and Translational Gastroenterology*, *5*(1), e45.

Streiner, D. L. (2002). The 2 "Es" of research: Efficacy and effectiveness trials. *The Canadian Journal of Psychiatry*, *47*(6), 552-556.

Sweeney, P. & Moore, C. (2012). Mobile Apps for Learning Vocabulary: Categories, Evaluation and Design Criteria for Teachers and Developers, *International Journal of Computer-assisted Language Learning and Teaching*, 2(4), 1–16. https://doi.org/10.4018/ijcallt.2012100101.

Tahir, R., & Arif, F. (2014). Framework for evaluating the usability of mobile educational applications for children. *Proceedings of the Third International Conference on Informatics Engineering and Information Science*, Lodz, Poland, 156–170. https://www.researchgate.com/publication/265852684_Framework_for_Evaluating_the_Usability_of_Mobile_Educational_Applications_for_Children

Tarcea, G., Puchala, B., Berman, T., Scorzelli, G., Pascucci, V., Taufer, M., & Allison, J. (2022). The Materials Commons Data Repository. In *2022 IEEE 18th International Conference on e-Science (e-Science),* 405-406). IEEE.

Taylor, G. et al. (2022). Selecting educational apps for preschool children: How useful are website app rating systems? *British Journal of Educational Technology, 53*(5), 1262–1282. https://doi.org/10.1111/bjet.13199.

Tu, Y., Zou, D., & Zhang, R. (2020). A comprehensive framework for designing and evaluating vocabulary learning apps from multiple perspectives. *International Journal of Mobile Learning and Organization, 14(*3), 370– 397. https://doi.org/10.1504/IJMLO.2020.108199

UNESCO, 2023. *Global Education Monitoring Report: Technology in Education: A tool on whose terms?*, UNESCO. https://www.unesco.org/gem-report/en/technology

Urquhart, N., Lee, J., & Wood, E. (2023). Get That App!: Examining Parental Evaluations of Numeracy Apps. *Journal of Research in Childhood Education*, 1-15.

Vackova, P., Cermakova Lindroos, A. & Kucirkova, N. (2023). *Children's Digital Books: Development, Validation and Dissemination of Quality Criteria*. Stavanger: University of Stavanger. https://ebooks.uis.no/index.php/USPS/catalog/book/268

Vaiopoulou, J., Papadakis, S., Sifaki, E., Stamovlasis, D., & Kalogiannakis, M. (2021). Parents' perceptions of educational apps use for kindergarten children: Development and validation of a new instrument (PEAU-p) and exploration of parents' profiles. *Behavioral Sciences, 11*(6), 82.

Vaiopoulou, J., Papadakis, S., Sifaki, E., Kalogiannakis, M., & Stamovlasis, D. (2022). Classification and evaluation of educational apps for early childhood: Security matters. *Education and Information Technologies*, 1-32.

Vanbecelaere, S., Adam, T., Sieber, C., Clark-Wilson, A., Boody Adorno, K., & Haßler, B. (2023). *Towards Systemic EdTech Testbeds: A Global Perspective*. Global EdTech Testbeds Network. https://doi.org/10.53832/opendeved.0285

Vázquez-Cano, E., Quicios-García, M. P., Fombona, J., & Rodríguez-Arce, J. (2023). Latent factors on the design and adoption of gamified apps in primary education. *Education and Information Technologies*, 1-31.

Vincent, T. (n.d.a). Educational app evaluation checklist. Squarespace. https://static.squarespace.com/static/50eca855e4b0939ae8bb12d9/50ecb58ee4b0b16f176a9e7d/50ecb593e4b0b16f176aa976/1330884481041/Vincent_App_Checklist.pdf

Verwimp, C., Snellings, P., Wiers, R. W., & Tijms, J. (2023). A randomised proof-of-concept trial on the effectiveness of a game-based training of phoneme-grapheme correspondences in pre-readers. *Journal of Computer Assisted Learning* (in press).

Wadhwa, M., Zheng, J., & Cook, T. D. (2023). How Consistent Are Meanings of "Evidence-Based"? A Comparative Review of 12 Clearinghouses that Rate the Effectiveness of Educational Programs. *Review of Educational Research*, 00346543231152262.

Wang, Y. Y., Wang, Y. S., Lin, H. H., & Tsai, T. H. (2019). Developing and validating a model for assessing paid mobile learning app success. *Interactive Learning Environments*, 27(4), 458–477. https://doi.org/10.1080/10494820.2018.1484773

Wang, A. I., & Tahir, R. (2020). The effect of using Kahoot! for learning–A literature review. *Computers & Education, 149*, 103818.

XPRIZE. (2019). *Global learning XPRIZE executive summary*. https://www.xprize.org/prizes/global-learning/ articles/glexp-executive-summary (Accessed January 8, 2023).

# APPENDIX 1

**List of included frameworks (academic)**

1. Baloh, M., Zupanc, K., Kosir, D., Bosnić, Z., & Scepanović, S. (2015)
2. Barr, R., & Kirkorian, H. (2023)
3. Booton, S. A., Kolancali, P., & Murphy, V. A. (2023)
4. Campbell, L. O., Gunter, G., & Braga, J. (2015)
5. Chatzopoulos, A., Karaflis, A., Kalogiannakis, M., Tzerachoglou, A., Cheirchanteri, G., Sfyroera, E., & Sklavounou, E. O. (2023)
6. Chen, X. (2016)
7. Cherner, T., Dix, J., & Lee, C. (2014)
8. Cherner, T., Fegely, A., Lee, C. Y., & Santaniello, L. (2016)
9. Goodwin, K., & Kucirkova, N. (2012)
10. Herodotou, C. (2021/2023)
11. Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015)
12. Huntington, B., Goulding, J., & Pitchford, N. J. (2023)
13. Hussain, A., Mkpojiogu, E. O. C., & Hassan, F. (2018)
14. Ibrahim, N.K. et al. (2019)
15. Israelson, M. H. (2015)
16. Kalogiannakis, M., & Papadakis, S. (2017)
17. Kay, R. (2018a)
18. Kay, R. (2018b)
19. Kay, R., Lesage, A., & Tepylo, D. (2019)
20. Khan, A. I., Al-Khanjari, Z., & Sarrab, M. (2017)
21. Kolak, J., Norgate, S. H., Monaghan, P., & Taylor, G. (2021)
22. Konca, A. S., Izci, B., & Simsar, A. (2023)
23. Kucirkova, N. (2018)
24. Lee, C. Y., & Cherner, T. S. (2015)
25. Lisenbee, P. S. (2018)
26. Lubniewski, K. L., McArthur, C. L., & Harriott, W. A. (2017)
27. Lytras, M.D. et al. (2019)
28. Mallawaarachchi, S.R. et al. (2023)
29. Martín-Monje, E., Arús, J., Rodríguez-Arancón, P., & Calle-Martínez, C. (2014)
30. Meyer, M., Zosh, J. M., McLaren, C., Robb, M., McCaffery, H., Golinkoff, R. M., Hirsh-Pasek, K., & Radesky, J. (2021)
31. Outhwaite, L. A., Early, E., Herodotou, C., & Van Herwegen, J. (2023)
32. Papadakis, S., Kalogiannakis, M., & Zaranis, N. (2017)
33. Papadakis, S., Vaiopoulou, J., Kalogiannakis, M., & Stamovlasis, D. (2020)
34. Rosell-Aguilar, F. (2017)
35. Sari, B., Takacs, Z. K., & Bus, A. G. (2019)
36. Shoukry, L., Sturm, C., & Galal-Edeen, G. H. (2015)
37. Sweeney, P. and Moore, C. (2012)
38. Tahir, R., & Arif, F. (2014)
39. Taylor, G. et al. (2022)
40. Tu, Y., Zou, D., & Zhang, R. (2020)

41. Urquhart, N., Lee, J., & Wood, E. (2023)
42. Vackova, P., Cermakova, A. & Kucirkova, N. (2023)
43. Vaiopoulou, J., Papadakis, S., Sifaki, E., Kalogiannakis, M., & Stamovlasis, D. (2022)
44. Vaiopoulou, J., Papadakis, S., Sifaki, E., Stamovlasis, D., & Kalogiannakis, M. (2021)
45. Vázquez-Cano, E., Quicios-García, M. P., Fombona, J., & Rodríguez-Arce, J. (2023)
46. Vincent, T. (n.d.a)
47. Wang, Y. Y., Wang, Y. S., Lin, H. H., & Tsai, T. H. (2019)

**List of commercial or national/policy frameworks**

1. AERO Standards of Evidence, Australian Education Research Organisation (AERO), 2021
2. EdSurge Product Index & Decision Guide, EdSurge, 2012
3. WiKIT Evidence-ready Service
4. LearnPlatform Impact Ready Audit
5. EdTech Developer's Guide, Office of Educational Technology, 2015
6. EdTech Evidence Toolkit, Office of Educational Technology, 2023
7. EdTech Standards of Evidence, What Worked Education, 2022
8. EdTech Tulna Standards, EdTech Tulna, 2021
9. EdTech Impact Quality Framework
10. Education Alliance Finland
11. LeanLab Pilot Ready Audit
12. EEF Padlock rating in Toolkit Guide, EEF, 2023
13. Barber, M. and Rizvi, S Efficacy Framework for Pearson (2013)
14. ESSA Tiers of Evidence, US Gov, 2015
15. Evaluation Taxonomy, Learning Assembly, 2017
16. ISTE Seal of alignment framework, ISTE, 2023
17. NESTA Standards of Evidence, Queensland Standards of Evidence, Queensland Department of Education, 2023
18. Research-Based Design certification, Digital Promise, 2019

## Appendix 2

The scoring rubric used in this report is based on the EVER framework (Kucirkova, Brod & Gaab, 2023). This Appendix contains some examples used in the scoring of the individual frameworks.

| Qualitative studies criteria | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Credibility** | Data were collected only through one qual method | Data were collected by at least two different methods and findings compared | Data and findings compared across at least two different methods, and two different analysts | Data/findings compared by at least two different analysts, with at least two different methods, and with some participants | Data/findings compared by at least two different analysts, with at least two different methods, and a representative sub-sample of participants |
| **Member validation** | Only one researcher interpreted the data | A group of researchers from the same research team interpreted a selection of the data | A group of researchers interpreted the data together, with clear description of how consensus was reached | Interpretation of findings was verified with independent researchers | Interpretation of findings was verified with independent researchers and participants |

| | | | | | |
|---|---|---|---|---|---|
| **Reflexivity** | Minimal researcher reflexivity throughout the project | Some awareness of the importance of researcher reflexivity noted/ documented | Mention of researcher reflexivity, but no systematic effort | Report of frequent and thorough researcher reflexivity but no close documentatio n of the process | Evidence of deep researcher reflexivity throughout the study, with supporting evidence |
| **Theoretical saturation** | There is little correspondence between the data and the hypothesis/<br>theory for the study | One or two aspects of the theory/ hypothesis are supported by the data | Some components of the theory/<br>hypothesis are supported by the data | Most components of the theory/<br>hypothesis are supported by the data | New data are no longer triggering theory<br>/hypothesis revision |

| Quantitative studies criteria | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Internal validity** | Power calculation shows insufficient statistical power OR big attrition rates OR no control group | Sample size on the margin of good statistical power; OR considerable attrition rates OR no standardized treatment conditions | Sample size sufficient for good statistical power AND/OR well-documented attrition AND/OR standardized treatment conditions | Sample size sufficient for high statistical power AND/OR low attrition AND/OR standardized treatment conditions | Sample size sufficient for high statistical power, detailed description of the intervention context AND low attrition rate or attrition well-documented AND standardized treatment conditions with a control group |
| **External validity** | Attempt made but poor randomization of participants | Some random-ization of participants | Appropriate participant randomization | Use of random or stratified sampling | Use random or stratified sampling |
| **Reliability** | N/A | N/A | N/A | Study was Replicated in other contexts | Study was Replicated in other contexts |
| **Objectivity** | N/A | N/A | N/A | N/A | The relationships between dependent and independent variables verified |