

Quality of evidence form

Date: April, 2026

Intervention: EIDU

1. Studies

Please note that this is not intended as a systematic review; rather, it represents a targeted scoping of the evidence to highlight the most pertinent and methodologically robust studies for our purposes.

We reviewed the following studies providing evidence on EIDU's program.

Direct evidence

- Major, L., Daltry, R., Otieno, M., Otieno, K., Zhao, A., Sun, C., Hinks, J., & Friedberg, A. (2026). Digital personalised learning to improve literacy and numeracy outcomes: A randomised controlled trial in Kenyan pre-primary classrooms. *Research Papers in Education*. <https://doi.org/10.1080/02671522.2025.2605645>

Evidence on digital personalized learning

- Major, L., Francis, G. A., & Tsapali, M. (2021). The effectiveness of technology-supported personalised learning in low- and middle-income countries: A meta-analysis. *British Journal of Educational Technology*, 52(5), 1935--1964. <https://doi.org/10.1111/bjet.13116>
- Rodriguez-Segura, D. (2022). EdTech in developing countries: A review of the evidence. *World Bank Research Observer*, 37(2), 171--203. <https://doi.org/10.1093/wbro/lkab011>

Evidence around adaptive instruction/structured pedagogy programs

- Evans, D. K., & Popova, A. (2016). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. *World Bank Research Observer*, 31(2), 242-270. <https://doi.org/10.1093/wbro/lkw004>

- Snilstveit, Birte, Jennifer Stevenson, Daniel Phillips, Martina Vojtkova, Emma Gallagher, Tanja Schmidt, Hannah Jobse, Maisie Geelen, Maria Grazia Pastorello, and John Evers. 2015. Interventions for Improving Learning Outcomes and Access to Education in Low and Middle- Income Countries: A Systematic Review. 3ie Systematic Review 24. London: International Initiative for Impact Evaluation (3ie). <https://www.3ieimpact.org/evidence-hub/publications/systematic-reviews/interventions-improving-learning-outcomes-and-access>
- Global Education Evidence Advisory Panel. (2020, October). *Cost-effective approaches to improve global learning: What does recent evidence tell us are “smart buys” for improving learning in low- and middle-income countries?* World Bank Group. <https://documents1.worldbank.org/curated/en/719211603835247448/pdf/Cost-Effective-Approaches-to-Improve-Global-Learning-What-Does-Recent-Evidence-Tell-Us-Are-Smart-Buys-for-Improving-Learning-in-Low-and-Middle-Income-Countries.pdf>
- J-PAL. (2017). *Roll call: Getting children into school* (Policy Bulletin). Abdul Latif Jameel Poverty Action Lab. <https://www.povertyactionlab.org/sites/default/files/publication/roll-call-getting-children-into-school.pdf>

2. Appraisal

2.1) Study design

High. Evidence comes from studies with robust studies designs, meta-analyses, systematic reviews and RCTs.

[Major et al. \(2026\)](#) - The study is a stratified, two-arm, cluster-randomised controlled trial (cluster-RCT) with one treatment and one control group. It was conducted in Murang'a County, Kenya, across four sub-counties (Gatanga, Maragua, Kandara, and Mathioya). The final analytical sample comprised 1,955 pre-primary learners aged 4 to 6, drawn from 291 government schools. Data were collected at three timepoints: baseline (October 2022), midline (May 2023), and endline (October 2023), spanning approximately 13 months across four school terms. The intervention consisted of a classroom-integrated digital personalised learning (DPL) tool developed by EIDU, deployed on low-cost Android smartphones and aligned with the Kenyan national curriculum and the Tayari structured pedagogy programme (p. 2, 11, 12).

[Major et al. \(2021\)](#) - The researchers conducted a meta-analysis to synthesize findings from existing research. They performed a systematic search for studies published between

2007 and 2020, identifying 16 randomized controlled trials (RCTs). The study focused on school-aged learners (typically 6 to 15 years old) in low- and middle-income countries (LMICs). The final analysis included a total sample of 53,029 learners across five countries: China, India, Malawi, Russia, and Uruguay. The meta-analysis examined the effectiveness of technology that personalizes or adapts to a student's learning level, specifically targeting outcomes in mathematics and literacy (p. 1935, 1937, 1941, 1957).

[Rodriguez-Segura \(2022\)](#) - This is a comprehensive review of evidence regarding educational technology (EdTech) in developing countries. The author identifies and synthesizes 81 "core studies" that utilize credible causal identification frameworks, primarily from 36 low- and middle-income countries since 2002. The specific methods used in these underlying studies include randomized controlled trials (RCTs), which account for three out of four studies, along with four other types of quasi-experimental evaluations. The review organizes these studies into four thematic categories: Access to technology, technology-enabled behavioral interventions, improvements to instruction, and self-led learning (p. 172-173).

[Evans and Popova, 2016](#) - Meta-review of systematic reviews

[Snijlsteit et al. \(2015\)](#) - Systematic review and meta-analysis

[GEEAP \(2020\)](#) - Systematic review

[J-PAL \(2017\)](#) - Evidence brief (akin to a systematic review but less technical)

Appraisal for study design

Level	Description
☒ High	Evidence comes from one or more systematic reviews, meta-analyses or experimental studies such as Randomised Controlled Trials (RCTs).
☐ Medium	Evidence comes from quasi-experimental designs. These are well designed but non-randomised studies.
☐ Low	Evidence comes from non-experimental studies (e.g., observational studies).
☐ Not considered	Evidence comes from case studies, anecdotal reports, expert opinion or narrative reviews.

2.2) Risk of bias

Overall we consider the risk of bias to be low across the reviewed studies. These follow robust methodologies and in the majority steps to minimise limitations.

Bias related to selection and allocation

- 1. Was true randomization used for assignment of participants to treatment groups?** Yes. Schools were randomly assigned to treatment or control groups using randomly generated integer lists in Microsoft Excel, following J-PAL methodological guidance. Learner selection within schools was also randomised, stratified by gender, using random integer lists (p. 12).
- 2. Was allocation to treatment groups concealed?** The paper does not explicitly report concealment of allocation. The study protocol was registered on an open-access repository prior to analysis, and the research team maintained independence from the DPL provider, but formal allocation concealment procedures are not described (p. 11, 12).
- 3. Were treatment groups similar at the baseline?** Yes. Chi-squared goodness-of-fit tests showed no statistically significant differences between treatment and control groups at baseline for total learner numbers ($p = .896$) or gender distribution ($p = .867$). Linear mixed model analysis of baseline IDELA scores also showed no significant difference in the final retained sample ($p = .638$) (p. 13, 14).

Bias related to administration of intervention/exposure

- 4. Were participants blind to treatment assignment?** No. Given the visible presence of DPL technology in treatment classrooms, participants and teachers were possibly aware of their assignment. The paper acknowledges this explicitly (p. 17-18).
- 5. Were those delivering the treatment blind to treatment assignment?** No. Teachers in treatment schools received training and implemented the DPL tool directly. Control group teachers were instructed not to change their practices, and ECDOs were allocated exclusively to either treatment or control schools to prevent contamination, but blinding of those delivering the treatment was not possible in this real-world implementation (p. 14-15).
- 6. Were treatment groups treated identically other than the intervention of interest?** Largely yes. Control schools continued with standard national curriculum teaching. ECDOs assigned to control schools received no intervention training. Measures were taken to prevent contamination by assigning ECDOs exclusively to one group and instructing control ECDOs not to alter practices (p. 14-15, 17-18).

Bias related to assessment, detection and measurement of the outcome

- 7. Were outcome assessors blind to treatment assignment?** Likely not fully. The paper acknowledges that enumerators may have been aware of treatment status

due to the visible presence of DPL devices in schools. However, enumerators were entirely independent of the treatment and had no conflict of interest. They were recruited through a rigorous independent process (p. 17-18).

- 8. Were outcomes measured in the same way for treatment groups?** Yes. The same assessment tool (IDELA) was administered by trained enumerators to both groups at all three timepoints, in the same languages and following the same protocols (p. 15-17).
- 9. Were outcomes measured in a reliable way?** Yes. The IDELA tool was developed by Save the Children and validated for children aged 3 to 6 across 11 countries, with established internal consistency, construct validity, and inter-rater and test-retest reliability. Enumerator inter-rater reliability was further strengthened through four days of training, six scoring exercises, and reached an intraclass correlation coefficient of 0.96 prior to data collection (p. 15-16).

Bias related to participant retention

- 10. Was follow up complete and if not, were differences between groups in terms of their follow up adequately described and analysed?** Follow-up was not fully complete. Overall attrition from baseline to endline was 32% (930 of 2,885 learners). The authors analysed attrition using logistic mixed model analysis and found that attrition did not statistically differ between treatment and control groups ($p = .786$), by gender ($p = .564$), or by overall baseline score ($p = .948$), indicating no attrition bias. The main driver of attrition was learners no longer enrolled, consistent across both school years. The final retained sample of 1,955 exceeded the adjusted minimum required, and observed effect sizes above 0.40 confirmed adequate statistical power (p. 13-15).

Statistical Conclusion Validity

- 11. Were participants analysed in the groups to which they were randomized?** Yes. The study followed an intention-to-treat approach, analysing learners in their assigned groups. The study protocol was pre-registered prior to analysis (p. 11-12).
- 12. Was appropriate statistical analysis used?** Yes. Linear and logistic mixed models were used to account for the clustered nature of the RCT, adding school as a random effect. Gain scores were calculated by subtracting baseline from midline and endline scores, then standardised using pooled standard deviations. Cohen's d was calculated using Jamovi 2.6.44. All models were assessed for required assumptions including variance, independence, and normal distribution (p. 17-18).
- 13. Was the trial design appropriate and any deviations from the standard RCT design (individual randomization, parallel groups) accounted for in the conduct and analysis of the trial?** Yes. The cluster-RCT design was appropriate given that the unit of assignment was the school rather than the individual learner, reducing contamination risk. The use of sub-counties as strata for randomisation, rather than

fully randomising across the county, was necessitated by existing ECDO zoning arrangements and is clearly explained and justified. This deviation is transparently reported and was accounted for in the analytical approach (p. 11-12).

[Major et al. \(2021\)](#)

Risk of bias assessment questions:

1. **Is the review question clearly and explicitly stated?** Yes. The study explicitly asks: "What is the effectiveness of technology-supported personalised learning in improving learning outcomes (mathematics and literacy) for school-aged children in LMICS?" (p. 1937)
2. **Were the inclusion criteria appropriate for the review question?** Yes. Eligibility criteria required studies to be published between 2007 and 2020, involve learners aged 5 to 18 in LMICs, feature technology-supported personalized learning, use a comparison control group in an RCT, and measure academic performance in mathematics or literacy (p. 1941).
3. **Was the search strategy appropriate?** Yes. The search involved automated queries of major databases (Google Scholar, Scopus, Web of Science, ERIC) and the Searchable Publication Database (SPuD), which indexes over 3 million records (p. 1941).
4. **Were the sources and resources used to search for studies adequate?** Yes. Researchers also accepted "grey literature," conducted forward and backward citation snowballing, and contacted study authors for further recommendations (p. 1941).
5. **Were the criteria for appraising studies appropriate?** Yes. Studies were assessed using a framework aligned with the Building Evidence in Education (2015) guidance, covering conceptual framing, contextual detail, research design, validity/reliability, ethics, and interpretation (p. 1942).
6. **Was critical appraisal conducted by two or more reviewers independently?** Not explicitly for all studies. The text states the assessment was "led by MT" and that a second rater (LM) randomly appraised six included studies to test validity (p. 1942).
7. **Were there methods to minimize errors in data extraction?** Yes. Studies were coded by one author (MT) with other authors independently reviewing the extracted data (p. 1943).
8. **Were the methods used to combine studies appropriate?** Yes. The authors used the Dersimonian and Laird method to determine summary effects and calculated composite effect size estimates for studies with multiple treatment arms or outcomes (p. 1944, 1958).
9. **Was the likelihood of publication bias assessed?** Yes. The researchers used a funnel plot and the "trim-and-fill" method to identify and correct for potential publication bias (p. 1949).
10. **Were recommendations for policy and/or practice supported by the reported data?** Yes. Recommendations regarding adaptive features and intervention intensity are directly linked to the meta-regression and sensitivity analysis results (p. 1936, 1955, 1956).

- 11. Were the specific directives for new research appropriate?** Yes. The authors suggest future research into cost-effectiveness, scaling, optimum duration, and the role of teachers, addressing gaps identified during the review (p. 1956).

[Rodriguez-Segura \(2022\)](#)

Risk of bias assessment questions:

- 1. Is the review question clearly and explicitly stated?** Yes. The review explicitly states five research questions: (a) which thematic areas and outcomes show the most promise, (b) which interventions have little effectiveness, (c) under what contextual circumstances they work best, (d) what are the current knowledge gaps, and (e) how cost structures influence scalability (p. 173).
- 2. Were the inclusion criteria appropriate for the review question?** Yes. Inclusion criteria required papers to use experimental or quasi-experimental evaluation methods, involve a technological tool for an educational policy issue, and be set in a developing country (p. 175).
- 3. Was the search strategy appropriate?** Yes. The search included scientific repositories, working paper series from research and international organizations, and forward and backward tracing from key papers (p. 172).
- 4. Were the sources and resources used to search for studies adequate?** Yes. The review utilized scientific repositories and working paper series from renowned international organizations (p. 172).
- 5. Were the criteria for appraising studies appropriate?** Yes. Studies were included based on having "credible causal identification frameworks" (p. 171-172). Studies were analysed using a similar framework as used by Escueta et al. (2020). See Appendix section titled 'Classification of Studies'.
- 6. Was critical appraisal conducted by two or more reviewers independently?** No. The document does not explicitly state whether the appraisal was conducted by two or more reviewers independently; it is written from the perspective of a single author (p. 172).
- 7. Were there methods to minimize errors in data extraction?** The document does not explicitly state the specific methods used to minimize errors during data extraction (p. 175).
- 8. Were the methods used to combine studies appropriate?** The author chose a qualitative description of trends instead of a formal meta-analysis to avoid masking crucial heterogeneity across diverse interventions and contexts (Appendix).
- 9. Was the likelihood of publication bias assessed?** The document does not explicitly state that a formal assessment of publication bias was conducted (p. 172).
- 10. Were recommendations for policy and/or practice supported by the reported data?** Yes. For example, the finding that mere provision of hardware is unlikely to improve learning is supported by multiple studies cited in the review (p. 178).
- 11. Were the specific directives for new research appropriate?** Yes. The review identifies gaps and frontiers in scientific knowledge as one of its core research questions (p. 173).

[Evans and Popova \(2016\)](#)

Evans and Popova's (2016) study provides a valuable meta-review of existing systematic reviews on educational interventions in developing countries. It offers insightful analysis and practical recommendations, even though the study does not follow all standard protocols of a systematic review. We consider that the study has a low risk of bias.

- **Is the review question clearly and explicitly stated?** Yes. The authors aim to understand why systematic reviews on educational interventions in developing countries yield divergent conclusions and to identify consistent findings across these reviews.
- **Were the inclusion criteria appropriate for the review question?** Yes. The study examines six systematic reviews that assess interventions aimed at improving learning outcomes in developing countries. The selection focuses on reviews that analyze experimental and quasi-experimental studies reporting on student learning outcomes.
- **Was the search strategy appropriate?** Yes. While the authors analyze existing systematic reviews, they do not conduct a new systematic search for primary studies. Their approach is more of a meta-review, assessing the methodologies and findings of existing reviews rather than conducting a fresh systematic review themselves.
- **Were the sources and resources used to search for studies adequate?** Partially. Since the study does not involve a new search for primary studies but rather analyzes existing systematic reviews, this criterion is not directly applicable. They describe that they “selected six systematic reviews that meet four key criteria: they cover multiple developing countries, focus on student learning outcomes, review experimental or quasi-experimental evidence, and assess a wide range of education interventions”. It should be caveated that this approach may have missed other relevant reviews and limits reproducibility.
- **Were the criteria for appraising studies appropriate?** Yes. The authors critically assess the methodologies of the included systematic reviews, examining factors such as the selection of studies, categorization of interventions, and the consistency of findings across reviews.
- **Was critical appraisal conducted by two or more reviewers independently?** Not specified. The paper does not detail whether multiple reviewers independently appraised the included systematic reviews.
- **Were there methods to minimize errors in data extraction?** Not specified. The study does not elaborate on the procedures used for data extraction from the included reviews, so it's unclear whether steps were taken to minimize errors.
- **Were the methods used to combine studies appropriate?** Yes. The authors employ a narrative synthesis to compare and contrast the findings of the included systematic reviews. They analyze differences in study inclusion, categorization of interventions, and conclusions drawn, which is appropriate for a meta-review.
- **Was the likelihood of publication bias assessed?** Not directly. The study discusses the potential for publication bias in the context of the included systematic reviews but does not conduct a formal assessment, such as a funnel plot analysis, to evaluate publication bias across the primary studies.
- **Were recommendations for policy and/or practice supported by the reported data?** Yes. The authors identify two classes of interventions—pedagogical

approaches tailored to student learning levels and repeated teacher training linked to specific tasks—as consistently effective across multiple reviews. These recommendations are grounded in the data analyzed.

- **Were the specific directives for new research appropriate?** Yes. The authors suggest that future systematic reviews should combine narrative synthesis with meta-analysis, conduct more exhaustive searches, and maintain low aggregation of intervention categories to enhance the reliability and applicability of findings.

[Snilstveit et al. \(2015\)](#)

We consider this systematic review and meta-analysis to be of a high standard. We hence determine a low risk of publication bias for this study. Snilstveit et al. (2015) conducted a rigorous and comprehensive systematic review, adhering to established methodological standards. The review's findings and recommendations are well-supported by the data, making it a valuable resource for policymakers, practitioners, and researchers interested in education interventions in low- and middle-income countries.

- **Is the review question clearly and explicitly stated?** Yes, see section 1.4. The review aims to assess the effectiveness of education interventions on improving learning outcomes and access to education in low- and middle-income countries. This objective is clearly articulated in the report.
- **Were the inclusion criteria appropriate for the review question?** Yes, see section 2.1. The authors included studies that evaluated the impact of education interventions on learning outcomes and access, focusing on low- and middle-income countries. They considered both randomized and quasi-experimental studies, ensuring relevance to the review question.
- **Was the search strategy appropriate?** Yes, see section 2.1. A comprehensive search strategy was employed, covering multiple databases and sources to identify relevant studies. The authors also searched grey literature and contacted experts to minimize publication bias.
- **Were the sources and resources used to search for studies adequate?** Yes, see section 2.1. The review utilized a wide range of sources, including academic databases, grey literature repositories, and organizational websites. This thorough approach ensured a comprehensive collection of relevant studies.
- **Were the criteria for appraising studies appropriate?** Yes – see section 2.3. The authors applied standardized criteria to assess the quality and risk of bias of included studies, ensuring that the appraisal was systematic and appropriate for the review's objectives.
- **Was critical appraisal conducted by two or more reviewers independently?** Yes, the review process involved multiple reviewers independently assessing studies to minimize bias and errors in appraisal. More specifically, they had two reviewers with a third one that was consulted to address disagreements (see page 23).
- **Were there methods to minimize errors in data extraction?** Yes, see section 2.3. Data extraction was performed using standardized forms, and multiple reviewers cross-checked the extracted data to reduce errors and ensure consistency.
- **Were the methods used to combine studies appropriate?** Yes, see section 2.5 and 2.6. The authors employed meta-analytic techniques where feasible and provided

narrative syntheses when meta-analysis was not possible, appropriately combining study results.

- **Was the likelihood of publication bias assessed?** Yes, see section 2.4. The review assessed publication bias through funnel plots and other statistical methods, acknowledging its potential impact on the findings.
- **Were recommendations for policy and/or practice supported by the reported data?** Yes, see section 10. The authors provided policy and practice recommendations grounded in the evidence synthesized from the included studies.
- **Were the specific directives for new research appropriate?** Yes, see section 10.7.

[GEEAP \(2020\)](#) - low risk of bias is considered for this study. The review question is clearly stated, the inclusion criteria of studies is appropriate for the objectives of the research, the criteria used for assessing the studies is considered appropriate and the recommendations for policy, practice and future research are supported by the data. Example:

“Approach to Classification. To classify interventions, the panel reviewed a range of recent, rigorous evaluative research on education and learning, primarily in low- and middle-income countries, that, where possible, included some measure of cost-effectiveness. The interventions discussed in this note were chosen because they have been rigorously tested using methodologies that can distinguish the causal effect of an intervention. They are also backed up by a body of other evidence, including evidence that the problem that the intervention was designed to address is widespread in low- and middle-income countries.” ([GEEAP, 2020](#), p.8)

The report does not explicitly describe the search strategy, methods to minimize data extraction errors, whether appraisal by multiple reviewers, or assessment of publication bias was conducted. However, it was produced by a panel of experienced researchers and education policy experts forming the Global Education Evidence Advisory Panel (GEEAP) — a joint initiative of the World Bank, FCDO, and Building Evidence in Education (BE2). Given the expertise of the panel and the collaborative nature of the project producing this review, it is reasonable to infer that appropriate quality assurance measures were applied.

[J-PAL \(2017\)](#) - moderate to low risk of bias is considered for this study. This is a research-based policy summary or evidence brief, synthesizing key findings from high-quality randomized evaluations, many of which may be part of J-PAL's database. It is not a systematic review strictly speaking however, J-PAL is a leading organization in impact evaluations. TLYCS considers they are a trustworthy source with expert evaluators. Although the report is brief, and cannot be assessed in the same way as a systematic review, the objective of the review is clearly stated, as well as the key inclusion criteria for the studies. Namely:

- Objective: “the objective is to draw general lessons about what types of strategies are most effective at improving student participation in school” ([J-PAL, 2017](#), p.3)

- Selection criteria: “Papers were included in this review if they were high-quality randomized evaluations in a developing country, had school participation as an outcome, and either evaluated an intervention designed to increase school participation or provided a plausible theory of change that could address participation. Quasi-randomized studies are also referenced (in footnotes) when they helped to interpret the results from randomized evaluations.” ([J-PAL 2017](#), p.3)

58 RCTs were reviewed. The way the information was appraised and synthesized (methods to combine reporting on studies) is considered appropriate for the objectives of the study.

Some caveats include that the search strategy is not specified, neither whether the appraisal was conducted by two or more reviewers independently, nor the likelihood of publication bias was assessed. However, the study was conducted and reviewed by a team of experts in impact evaluation and education:

- Bulletin Authors: Rachel Glennerster, Meghan Mahoney, Meagan Neal, Luke Strathmann, Caitlin Tulloch
- Reviewers: Karthik Muralidharan, Phil Oreopoulos

Weighing caveats and what is known about the methodology of the review with the expertise of authors we consider the risk of publication bias to be low to moderate for this study.

Appraisal for risk of bias

Level	Description
☒ Low	The <i>majority</i> of RCTs, systematic review(s) or meta-analysis considered show a <i>low</i> risk of bias.
☐ Medium	Across <i>some</i> of the considered studies the risk of bias may be <i>moderate</i> , but not high.
☐ High	A <i>high</i> risk of bias is identified across studies.

2.3) Imprecision

Moderate. Based on the reviewed literature, the potential for imprecision varies significantly across study types and intervention categories, leading to an overall assessment of moderate imprecision.

In high-quality individual evaluations, such as Major et al. (2026), imprecision is low for primary outcomes due to robust a priori powering and a sample size (n=1,955) that exceeded the validity threshold, yielding highly significant effects ($p < .001$) (p. 11-12). However, imprecision increases to a moderate level in that same study when observing

subgroup analyses and temporal shifts, where smaller samples introduce greater uncertainty (p. 18-20). Similarly, in meta-analyses like Major et al. (2021), a massive total sample size (n=53,029) provides high statistical power, yet imprecision remains a concern because the data is geographically concentrated, with over half the studies originating from a single country (p. 1935, 1955).

In broader reviews, imprecision is more pronounced. Rodriguez-Segura (2022) identifies "imprecise null effects" in categories like hardware access, where studies often fail to reach statistical significance (p. 178). Other major reviews, such as Snilstveit et al. (2015) and GEEAP (2020), exhibit moderate imprecision due to high heterogeneity in results across structured pedagogy and a lack of reported confidence intervals for specific metrics like Learning-Adjusted Years of School (LAYS). Conversely, Evans and Popova (2016) find low imprecision in core instructional interventions, noting that different reviews report consistent effect sizes (0.22–0.25 SD) for the same underlying studies.

[Major et al. \(2026\)](#) - The study was powered a priori, assuming a minimum detectable effect size of 0.20 SD, 80% power, an intraclass correlation of 0.2, and a significance level of 0.05, requiring a minimum sample of 220 schools. The final analytical sample of 1,955 learners across 291 schools exceeded the adjusted minimum threshold, and the authors note that observed effect sizes above 0.40 confirm the sample is sufficient to maintain study validity. The primary outcome (overall IDELA score) yielded a large and highly significant effect size of 0.534 SD ($F(1, 286) = 49.285, p < .001$), and domain-level effects were similarly precise and significant. Imprecision is therefore of low concern for the primary outcome. However, for the subgroup analyses (achievement level, gender, class size), sample sizes within subgroups are smaller and effect sizes are more variable, introducing somewhat greater uncertainty for those estimates. The midline-to-endline period showed near-zero and non-significant effects (0.054 SD overall, $p = .396$), which may reflect a genuine plateau rather than imprecision, but the shorter data collection window in that period introduces some ambiguity. Overall, imprecision is low for the primary outcomes but moderate for subgroup and temporal analyses (p. 11-12, 18-20).

[Major et al. \(2021\)](#) - In this study, imprecision was addressed by the meta-analytic approach, which aimed to improve the precision of estimates by pooling data. While the total sample size was large (53,029 learners), the authors noted that the limited number of studies (16) from only five countries (with nine of those coming from China), requires care when interpreting findings. The overall effect size was 0.18 with a p-value of 0.001, which the authors described as a "statistically significant if moderate-positive" effect. Overall, imprecision is of some concern due to the concentration of data from a single research lab and a limited number of countries, although the large total number of participants provides a significant degree of statistical power for the primary findings (p. 1935, 1948, 1955).

[Rodriguez-Segura \(2022\)](#) - Imprecision is noted in several studies where effects were either null or statistically insignificant. For instance, the median effect of studies in the "Access to technology" category is described as an "imprecise null effect" (p. 178). Additionally, a study on e-readers in Nigeria found that students receiving curriculum materials had "large, imprecise effects" (Appendix). The review generally deals with imprecision by opting for a qualitative synthesis rather than aggregate meta-parameters that might hide such variation (Appendix). Overall, imprecision is a significant concern across many studies, particularly those involving hardware provision without pedagogical support, leading to results that often fail to reach statistical significance (p. 178).

[Evans and Popova \(2016\)](#)

Evans and Popova's meta-review reveals minor variation in reported effect sizes for interventions that tailor instruction to student learning levels across systematic reviews, suggesting low statistical imprecision in the core evidence. However, broader inconsistencies in categorization, study inclusion, and reporting practices across reviews introduce moderate interpretive uncertainty, which should be considered when translating evidence into policy or practice.

- For the same underlying studies (e.g., Banerjee et al., 2007; Duflo et al., 2011), reviews such as McEwan (2015) and Conn (2014) report similar effect sizes (around 0.22–0.25 standard deviations) for instructional interventions adapted to student level. This small variation suggests that the core evidence on the effectiveness of these interventions is relatively robust, and does not indicate substantial imprecision in effect size estimates.
-

[Snilstveit et al., \(2015\)](#) - Systematic review and meta-analysis. They assessed "heterogeneity of effect sizes graphically, and tested for heterogeneity formally by calculating the Q-statistic, the I^2 , and Tau^2 to provide an overall estimate of the amount of variability in the distribution of the true effect sizes" [Snilstveit et al., \(2015\)](#). Analyses were undertaken by type of intervention. Focusing on remedial teaching, they note that the number of studies considered for this type of intervention was small and that there was substantial heterogeneity across studies. For structured pedagogy they considered 21 studies and found high heterogeneity in results for learning outcomes in maths and language, but low heterogeneity for composite scores and cognitive skills. Hence we consider a moderate level of heterogeneity overall.

[GEEAP \(2020\)](#) - Figure 2 in Appendix D shows the Learning-Adjusted Years of School (LAYS) gained, by intervention category. The graph shows variation in the number of studies considered for each intervention category. They however do not report confidence intervals for the calculated number of LAYS. Hence we cannot reliably assess imprecision. Given this uncertainty we consider there is likely moderate imprecision.

[J-PAL \(2017\)](#) - In Figure 1 they present the results with confidence intervals for each estimate (page 8 and 9). While there is variation across types of intervention in the magnitude of the confidence intervals, imprecision is considered to be moderate.

Appraisal for imprecision

Level	Description
☐ Low	The studies have <i>high</i> statistical power to detect significant differences. Confidence intervals are sufficiently narrow meaning that statistical imprecision has negligible impacts on decision making. This typically means that the average study has large sample sizes.
☒ Medium	The studies only have <i>moderate</i> statistical power to detect significant differences. Confidence intervals may introduce some uncertainty into decision making. Some of the considered studies have small sample sizes.
☐ High	The studies have <i>low</i> statistical power to detect significant differences with confidence intervals that introduce uncertainty into decision making. This typically means that the studies have small sample sizes.

2.4) Inconsistency

Moderate. Based on the reviewed literature, the potential for inconsistency is moderate. While direct evidence from Major et al. (2026) demonstrates high internal consistency across learning domains and subgroups, it reveals temporal inconsistency as effects plateaued after the midline. Broader meta-analyses highlight significant heterogeneity (Major et al., 2021; Rodriguez-Segura, 2022), often driven by whether technology complements or replaces teachers. However, for interventions on adaptive learning, consistency is higher. Evidence from Evans and Popova (2016) and Snilstveit et al. (2015) consistently supports the effectiveness of tailored instruction, suggesting that variability is largely explained by implementation quality and software design.

[Major et al. \(2026\)](#) - Within the study, results are largely consistent: treatment effects were statistically significant across all three primary outcome domains (overall IDELA, numeracy,

literacy) and across 11 of 14 individual assessment items. Effects were consistent across gender and class size subgroups, with no significant interaction effects detected. The main internal inconsistency is the marked difference between the baseline-to-midline period (effect sizes of 0.496, 0.445, and 0.400 SD for overall, numeracy and literacy respectively, all $p < .001$) and the midline-to-endline period (effect sizes of 0.054, 0.011, and 0.087 SD, all non-significant). The authors discuss possible explanations including a content ceiling effect and potential novelty effects, but acknowledge that the reasons remain uncertain. This temporal inconsistency is the primary within-study source of variability and limits conclusions about the sustained effectiveness of the intervention beyond the first half of implementation (p. 18-24).

[Major et al. \(2021\)](#) - The authors assessed this using the I^2 statistic and found significant heterogeneity ($I^2 = 83.59\%$, $p = 0.001$ for the main analysis). This variation was explored through meta-regression, which showed that "HIGH" personalization features (adaptive/adjusting) led to significantly higher effect sizes (0.35) than "MEDIUM" features (0.13). Overall, there is high inconsistency across the reviewed studies, which is largely explained by differences in the levels of personalization and the software used, rather than the learning domain (math vs. literacy) (p. 1948, 1950).

[Rodriguez-Segura \(2022\)](#) - Inconsistency is evident across different types of EdTech interventions. For example, while some handheld device interventions found positive effects (Pitchford 2015; Mensch and Haberland 2018), others found null effects (Habyarimana and Sabarwal 2018) (p. 177, 178). Similarly, computer-assisted learning (CAL) showed positive effects when used as a complement but negative effects when used as a substitute for regular instruction (Appendix).

The remaining systematic reviews mention some variations across the studies they assessed ([GEEAP \(2020\)](#), [J-PAL \(2017\)](#)). However, it should be noted that the meta-analyses find that interventions that tailor teacher instruction to children's learning level have positive impacts on learning outcomes ([Snilstveit et al., \(2015\)](#); [Evans and Popova \(2016\)](#)).

Appraisal for inconsistency

Level	Description
☐ Low	The estimated effects of the intervention are broadly consistent across the studies (e.g., RCTs). There may be some minor variation.
☒ Medium	The estimated effects are broadly consistent across the studies, although there may be some moderate variation.

☐ High	There is inconsistency in the estimated effects across studies.
-------------------------------	---

2.5) Indirectness

Low. Evidence comes from highly relevant LMICs contexts, including direct evidence of EIDU's operations in Kenya.

[Major et al. \(2026\)](#) - The study is highly directly relevant to the intervention being assessed. It evaluates EIDU's specific DPL tool, as deployed in real government pre-primary schools in Kenya, integrated with the national curriculum and the Tayari structured pedagogy programme, on low-cost Android smartphones. The population (pre-primary learners aged 4 to 6 in government schools in a lower-middle-income country), the setting (291 schools across four sub-counties in Kenya), and the outcomes (IDELA-assessed numeracy and literacy) are directly relevant to EIDU's programme as described. The study was conducted over a full school year in a real-world, non-lab setting, further enhancing its applicability.

[Major et al. \(2021\)](#) - This study targeted LMICs specifically to minimize indirectness from high-income contexts. Evidence was retrieved from Malawi (representing Sub-Saharan Africa) and India (representing South Asia). While the findings are relevant to the broad regions, they may not perfectly generalize to every specific LMIC. Overall indirectness is moderate; while the study focuses on LMICs, the high concentration of studies in China may limit the direct applicability of findings to specific countries like Kenya (where EIDU operates) that were not included in the primary data (p. 1954, 1955, 1958).

[Rodriguez-Segura \(2022\)](#) - Indirectness is addressed by the author's decision to exclude evidence from developed countries, arguing that those results may translate differently in developing contexts due to infrastructure and access issues (p. 172). Regarding TLYCS interests, the review explicitly includes evidence from Sub-Saharan Africa and South Asia (p. 174). Specifically, it mentions that EdTech firms are concentrated in Kenya (where EIDU operates), South Africa, and Nigeria (p. 173). Several studies cited in the review were conducted in Kenya, such as Habyarimana and Jack (2018) and Duflo et al. (2011) (Appendix).

The other multi-country analyses that include African settings ([Evans and Popova, 2016](#); [Snilstveit et al., 2015](#)).

Appraisal for indirectness

Level	Description
-------	-------------

☒ Low	The RCTs or quasi-experimental studies study the intervention directly as it is implemented by the prospective RNP or in highly relevant contexts.
☐ Medium	The studies are only moderately relevant to the context in which the intervention is implemented.
☐ High	The studies have very low relevance to the context in which the intervention is implemented.

3. Summary

Criteria	Appraisal		
Study design	☒ High	☐ Medium	☐ Low
Risk of bias	☐ High	☐ Medium	☒ Low
Imprecision	☐ High	☒ Medium	☐ Low
Inconsistency	☐ High	☒ Medium	☐ Low
Indirectness	☐ High	☐ Medium	☒ Low

4. Quality of evidence final score

Please award the prospective RNP a final *quality of evidence* score:

Level	Description
☒ High	There is a wide range of well conducted experimental and quasi-experimental evidence showing the effectiveness of the intervention (i.e., the studies show a low risk of bias and the estimates are precisely estimated). This includes evidence in the geography where the organization operates. There is at least one study assessing the impact of the intervention as implemented by the RNP (direct evidence). There is broad consistency on the effect of the intervention across considered studies.
☐ Medium	There is a medium range of experimental, quasi-experimental and non-experimental evidence showing the effectiveness of the intervention. Some studies have a moderate risk of bias, but none show a high risk. Some studies show moderate imprecision. There is primarily indirect evidence for the intervention. There is moderate variation of the results across the

	studies.
☐ Low	There is a low range of studies showing the effectiveness of the intervention. There is a high risk of bias and imprecision across some studies. There is only indirect evidence on the effectiveness of the intervention (none coming from studies where intervention is implemented by prospective RNP). There is high variation of the results across studies.

Please provide any comments to back up the decision if considered necessary. Only one quality score should be awarded to the RNP even if it implements more than one intervention. Hence, to inform this final score, first award an intervention specific score. Take the average across the interventions' final scores.