Rapid Evidence for AI in Education: An RCT Playbook

Sierra Leone

This playbook outlines the decision making behind the implementation of a rapid randomised controlled trial (RCT) in Sierra Leone (LearnLM Team, Google and Fab AI, 2026) measuring the impact of an in-classroom AI intervention on students' learning outcomes. It is not intended to be a comprehensive textbook on impact evaluation methodology, nor does it substitute the need for specialist expertise. This playbook is meant to be iterative, undergoing continuous refinement as we accumulate operational experience and evidence from the field. While we think there is value in sharing our experience early, it is critical to note that the external validity and generalisability of this framework to alternate national contexts or educational ecosystems have not yet been established. We remain committed to the dissemination of open science and will release successive iterations of this documentation as our thinking evolves and as we conclude subsequent RCTs in other global settings to construct a more comprehensive, cross-country evidence base. Please use the following form if you would like to share feedback on this playbook: goo.gle/LearnLM-SierraLeone-Playbook-Feedback

CONTENTS

Introduction

Causal evidence on the integration of artificial intelligence within educational ecosystems is limited (Fesler et al., 2026). Conventional impact evaluation methodologies in education need updating for the rapid life cycle of AI development, as they typically require multi-school-year timelines that are poorly suited to the rapid pace of AI evolution.

To address this evidence gap, we suggest a rapid-cycle field randomised controlled trial (RCT) framework, to provide rigorous causal (rather than correlational) evidence on a condensed timescale of school terms rather than school years. Our framework aims to generate timely, high-fidelity results that keep pace with the frontier of innovation. We believe that rapid-cycle field RCTs are needed to complement other existing evaluation frameworks (like tightly controlled, smaller-scale laboratory experiments and offline AI benchmarks) by allowing for rigorous measurement of how the intervention performs in the complex reality of physical classrooms on a meaningful scale. This playbook codifies our methodological approach, with the hope that other researchers and practitioners can draw on our experience, provide feedback and help improve the framework. Readers may note clinical terminology such as 'dosage' and 'treatment' used throughout this report; this reflects the educational sector's historical adoption of the medical model to standardise how efficacy is measured.

Central to our philosophy is a teacher-first, sociotechnical perspective. That is, AI is designed as a force multiplier for the educator, enhancing human connections and pedagogical direction rather than attempting to replace the central role of the teacher. Achieving this requires deep contextual co-design with local ecosystems to ensure that implementation is grounded in the realities of diverse global settings. Furthermore, our research prioritises scalability and efficiency; the intervention is structured to be cost-effective and sustainable in principle once the external research team concludes its engagement.

This methodological playbook is derived from the design, implementation, and analysis of a randomised controlled trial (RCT) of the Guided Learning feature in the Gemini app, within junior secondary schools in Sierra Leone. The implementation framework was co-designed by Fab AI and Google DeepMind, and executed in partnership with EducAid, Laterite, and Oxford MeasurEd (see Modular Partnership Structure and Institutional Insulation).

This playbook documents the methodological rationale, the trade-offs considered, and the decision rules that may be useful to adapt this rapid-cycle model to varied educational contexts.

Research Questions

The research agenda was structured around three foundational questions:

1. Does the intervention work?
2. How does the impact vary by how much the tool is used?
3. Does the intervention work equitably?

This list is not exhaustive of all the important questions we want to answer about the role of AI in education. Rather, it serves as a starting point for subsequent research cycles. We hope to investigate additional questions in our future studies, including the impact of AI on student metacognition, relational intelligence, and self-efficacy.

Methodological Note: In-Classroom Integration

For this study we adopted an in-classroom model of AI integration (rather than introducing AI as a homework tutor, an after-school resource or a tool for students to use in their own time). We view AI as a pedagogical complement rather than a substitute for human instruction, and want to understand its effectiveness as a tool integrated by teachers into their daily routine. This design preserves the central role of the educator in directing student interactions and ensures implementation is rooted in the realities of the daily classroom experience in local education systems.

The main question focuses on the **effect on learning outcomes** (see Intention-to-Treat (ITT) Estimand), specifically estimating whether the intervention causes a statistically significant improvement in student achievement:

What is the causal effect of a teacher-supported, in-class Gemini intervention on junior secondary students' learning in mathematics in Sierra Leone?

The intervention utilised Guided Learning — a tool in the Gemini app built to act as a teacher rather than an assistant. Instead of giving away the answer, Guided Learning invites the learner to actively engage in a conversation, breaking down the problem into smaller chunks and scaffolding the learner to do the work themselves, encouraging productive struggle and avoiding cognitive offloading.

We chose junior secondary level, as it optimised three factors: 1) educational returns, as the marginal returns to pedagogical interventions are typically maximised during earlier interventions; 2) requisite student literacy and cognitive maturity for meaningful technological engagement; and 3) targeting non-examination years to ensure implementation feasibility within existing school schedules (see Eligible Population and Sampling Framework).

The second inquiry addresses **compliance and dosage effects** – does more exposure to the tool produce greater learning gains. We utilised a Treatment-on-the-Treated (TOT) analysis, estimating the effect of each additional hour of Gemini use on student outcomes, as well as the effect of meeting the intended dosage threshold (see Treatment-on-the-Treated (TOT) and Instrumental Variables).

The final inquiry examines the **heterogeneity of treatment effects** (see Estimation of Heterogeneous Treatment Effects). Recognising that EdTech impact may be non-uniform, we decomposed the effect across pre-specified student characteristics—including grade level, baseline mathematics proficiency, baseline English literacy, gender, and age—to identify for whom the intervention is most effective.

Modular Partnership Structure and Institutional Insulation

To safeguard the internal validity of the RCT, we engineered a modular partnership structure. By delegating core implementation, data collection, and analytical functions to distinct organisations, the study aimed to neutralise potential conflicts of interest and to insulate outcome measurement against bias. Fab AI and Google DeepMind jointly spearheaded the overall research and impact evaluation design.

Implementation Partner: EducAid

EducAid led the local execution of the intervention, leveraging over 30 years of operational experience in Sierra Leone. Their mandate encompassed school engagement and delivery of teacher training workshops (see Teacher Training and Pedagogical Protocol), and the recruitment of field monitors to ensure implementation fidelity (see Monitoring Implementation Fidelity). This localised expertise was critical for navigating the realities of Sierra Leone and securing active institutional buy-in.

Data Collection Partner: Laterite

Laterite managed all facets of primary data collection and fieldwork logistics, including the administration of standardised paper-based testing, and the digitisation of results (see Data Collection Technical Specifications). Methodological independence was maintained by separating the data collection team from the implementation staff, thereby mitigating enumerator bias and ensuring standardised assessment conditions.

Assessment Development Partner: Oxford MeasurEd

Oxford MeasurEd was responsible for the psychometric design and analysis of the learning assessments (see Learning Outcomes Measurement). They developed curriculum-aligned instruments and utilised Item Response Theory (IRT) to generate scaled ability estimates. Their insulation from the intervention delivery and data collection teams ensured that the measurement of learning gains was conducted without knowledge of treatment assignment, providing an additional layer of rigor.

Methodology

We implemented a preregistered (see The Pre-Analysis Plan) cluster randomised RCT. As discussed previously, a fundamental objective of this experimental design is to eliminate selection bias and thus isolate the causal treatment effect of Gemini Guided Learning on student achievement. The purpose of random assignment is to make treatment and control groups statistically equivalent at baseline across both observable and unobservable characteristics, ensuring that any divergent learning trajectories can be attributed to the intervention rather than to other variables.

The study utilised a two-arm experimental framework with the classroom as the primary unit of randomisation. In the treatment arm, students engaged with Guided Learning for approximately 50% of their weekly mathematics instructional time. Conversely, the control arm continued with business-as-usual instruction. To maintain internal validity, all teachers (treatment and control) participated in identical training protocols prior to randomisation, as detailed in the Teacher Training and Pedagogical Protocol section. Furthermore, in many instances, the same teachers taught both treatment and control classrooms.

> **Methodological Note: Two-Arm Experimental Design**
>
> A central design decision in any RCT is the selection of the counterfactual: what condition should the control group experience? In an ideal setting with unlimited resources, multiple control arms could isolate distinct causal contributions—for example, paired practice without AI to identify the benefits of the collaborative structure alone, or interaction with a general-purpose AI system to isolate the effect of the learning science insights incorporated into Guided

Learning. However, each additional arm requires a disproportionately larger sample to maintain statistical power. Given logistical and budgetary constraints, a two-arm RCT was selected to estimate the aggregate effect of the full intervention package compared to the counterfactual of standard instruction, answering the most policy-relevant question—does introducing this package improve learning beyond current practice?

Methodological Note: Selection of Mathematics

Mathematics was selected as the focal subject due to persistent learning deficits in this subject in the region and its suitability for structured, step-by-step scaffolding. Additionally, mathematics achievement can be measured with high precision using standardised assessments, which optimises the statistical power of the evaluation and reduces measurement error in the primary outcome.

Methodological Note: 50% Instructional Dosage

The 50% allocation of weekly mathematics instruction to Guided Learning constitutes a pragmatic trade-off. It balanced the need for sufficient treatment exposure to detect causal effects within a condensed eight-week cycle against the wish to maintain AI-free teacher-led pedagogical direction. This dosage serves as an exploratory starting point; subsequent research cycles will be required to understand the optimal dosage.

Experimental Evaluation Framework

Clustered Randomisation and Cluster Unit Definition

In this study **classroom** serves as the primary unit of randomisation. This configuration was selected as a pragmatic balance—it allowed us to maintain sufficient statistical power to detect policy-relevant effects while acknowledging the logistical and budgetary constraints.

Methodological Note: Selection of Randomisation Level

School-level randomisation, while theoretically superior for eliminating within-school spillovers (e.g. through treatment and control students discussions), was deemed unsuitable, due to significant budgetary and infrastructure requirements (e.g., electricity and connectivity). **Individual-level randomisation** was methodologically incompatible with the teacher-led, in-class delivery model, which requires unified pedagogical instruction to all students within the same class. **Classroom-level randomisation** provided the optimal balance between statistical sensitivity and implementation feasibility, with contamination risks mitigated through strict device management protocols and the presence of dedicated field monitors (see Monitoring Implementation Fidelity).

To optimise statistical precision and control for between-school differences, we implemented a stratified randomisation protocol blocked at the school-by-grade level. This blocking strategy ensures that randomisation occurs within each school-by-grade combination, maximising the comparability of treatment and control arms. Within each school-by-grade block, classrooms were assigned

to the treatment or control condition with an equal probability. This experimental allocation resulted in a final sample of $N = 48$ classrooms (24 treatment, 24 control) distributed across 12 government-supported schools.

Eligible Population and Sampling Framework

The primary cohort comprised students in Junior Secondary levels JSS1 (Grade 7) and JSS2 (Grade 8) aged 13 years or older, situated within government-supported institutions in the Port Loko District of Sierra Leone. These schools were chosen because they possessed a record of prior engagement with our implementation partner, EducAid. This existing partnership facilitated the rapid institutional engagement necessary to meet the condensed recruitment timeline of the study. The JSS3 (Grade 9) cohort was intentionally excluded from the research design to avoid disrupting high-stakes national examination preparation and to maintain institutional alignment with the schools' academic priorities.

School eligibility was determined by a set of inclusion criteria designed to verify prerequisite infrastructure and institutional readiness. Specifically, participating schools were required to demonstrate stable electricity access and satisfy a minimum threshold of two classrooms per grade level to facilitate within-school randomisation.

While the resulting sample provides high internal validity, it is important to acknowledge that it is not perfectly representative of the national educational landscape. A comprehensive national representative design would require random selection from an exhaustive registry of all schools; however, such an approach is often incompatible with technology-intensive interventions due to pervasive institutional constraints in infrastructure. Therefore, the findings of this evaluation apply most directly to educational settings that share these baseline characteristics, and generalisations regarding external validity should be approached with appropriate technical caution.

> **Methodological Note: Representativeness vs. Implementation Feasibility**
>
> Criteria-based sampling is a standard methodological trade-off in rapid-cycle research and pilot-stage evaluations. This approach prioritises implementation feasibility by testing the intervention under conditions where it can be realistically sustained. While this design inherently limits broad external validity, it ensures that the causal evidence base is established within an environment capable of supporting the prerequisite infrastructure.

Within-School Cluster Selection

Following the identification of the eligible schools, we looked to determine specific classroom participation. To maintain the scientific validity of the causal estimates, we ensured that all classroom selections are finalised and fixed prior to the commencement of random assignment. This protocol precludes the post-hoc substitution of non-selected classrooms into the study arms, as such deviations would compromise the random assignment process and could introduce selection bias. In the event of cluster-level attrition following randomisation, the methodological framework dictates the acceptance of reduced sample size over the introduction of non-randomised replacements.

Sampling Considerations

Sampling considerations were driven by the need to achieve sufficient statistical power to detect policy-relevant changes in learning outcomes in the most cost effective manner. The primary parameters for the power analysis included the Minimum Detectable Effect (MDE), the Intra-cluster Correlation Coefficient (ICC), and the explanatory power of baseline covariates (R^2).

Minimum Detectable Effect

This parameter serves as a threshold for both inferential precision and policy significance. The selected value should aim to ensure the study is sufficiently powered to detect learning gains that would justify the fiscal and operational costs of national-scale implementation.

Intra-Cluster Correlation

The ICC quantifies the proportion of total outcome variance attributable to between-cluster differences. Higher ICC values signify greater within-classroom similarity of students, which necessitates a larger number of clusters to maintain statistical sensitivity.

Methodological Note: Estimating and Applying the ICC

The ICC is subject to significant variation based on geographical context, academic domain, and the level of clustering. In Sub-Saharan Africa, the reported school-level mathematics achievement ICCs range from 0.08 (Seychelles) to 0.55 (South Africa) (Kelcey et al., 2016). This means that the gap in educational quality between different schools varies wildly depending on where you are; in the Seychelles, most schools produce relatively equal maths results, but in South Africa, the specific school a child attends correlates with a substantial portion of their success. For rigorous trial design, researchers should distinguish between unconditional and conditional ICCs—the latter of which accounts for baseline covariates and block fixed effects—as power calculations should ideally reflect the conditional variance structure to avoid underpowering the analysis. Precise parameterisation is important to sample size planning; underestimation of the ICC risks structural underpowering, while overestimation results in inefficient resource allocation and unnecessarily large sampling frameworks.

Covariates and Explained Variance

Baseline covariates, such as students' starting knowledge measured by their pre-trial mathematics and literacy scores, can be controlled for to make the statistical model more sensitive. When designing the study, R_1^2 coefficient reflects our assumption on how much of the background noise among students in the same classroom would be filtered by such past scores. Similarly, R_2^2 estimates how much of the performance gap between different classrooms would be controlled for by classroom characteristics. By relying on safely moderate estimates rather than overly optimistic ones, the trial can remain powerful enough to detect the Minimum Detectable Effect (MDE).

Methodological Note: Sample Size Optimisation Under Constraints

In educational ecosystems characterised by infrastructure and logistical constraints, achieving theoretically ideal cluster volumes is frequently infeasible. When facing such limitations, researchers can optimise the existing sample by strengthening the baseline covariate set to maximise R^2 or accepting a higher MDE threshold.

Learning Outcomes Measurement

The study utilised a baseline-endline experimental design. The baseline assessment (administered before the intervention began) serves as the primary covariate within the statistical analysis framework, hypothesised to absorb between-student variance and enhance the statistical power of the ITT estimate (see Intention-to-Treat (ITT) Estimand). It also helps establish pre-treatment comparability between groups. The endline assessment (administered immediately after the intervention concluded) was designed to differ from the baseline to prevent recall bias and ceiling effects, whilst still measuring the same underlying mathematical constructs.

Furthermore, a baseline English reading comprehension instrument was administered to evaluate literacy as a potential moderator of treatment efficacy. Given that the primary language of interaction with Gemini was English, this control was designed to investigate the heterogeneity of effects across varying levels of linguistic proficiency.

Methodological Note: Baseline Administration

Baseline assessments should occur prior to randomisation. This avoids any risk of the assessment process being influenced by treatment assignment, but more importantly, it allows for the use of the baseline data to verify that the randomisation produced balanced groups. In smaller sample RCTs, random assignment can still produce imbalances in key characteristics between the treatment and control groups. Baseline data enables researchers to detect such imbalances and, if necessary, adjust the analytical models to account for them. It is good practice to specify these adjustments as part of study preregistration (see The Pre-Analysis Plan).

Assessment Instrument Development

Assessment instruments were developed by Oxford MeasurEd to ensure alignment with the Sierra Leonean Junior Secondary mathematics curriculum. The item pool contained content targeting specific curricular domains covered during the intervention cycle as well as foundational competencies and broader mathematical constructs. This dual-focus design ensures the assessment is sensitive to intervention effects while maintaining the capacity to evaluate general subject mastery.

Before large-scale administration, cognitive interviews were conducted with a small sample of local students to verify that the assessment items were contextually appropriate, clearly articulated, and demonstrated an appropriate difficulty distribution for the target population.

Methodological Note: Curricular Alignment

Assessments strictly aligned with the intervention content maximise the likelihood of detecting positive effects but risk confounding learning with "teaching-to-the-test" artifacts. Conversely, excessively broad instruments may fail to detect nuanced pedagogical gains. Our framework employs a hybrid sampling of items to distinguish between narrow content acquisition and the transfer of foundational mathematical skills, balancing sensitivity with broader construct validity.

Item Response Theory Scaling

Student achievement was estimated using a 3-parameter logistic Item Response Theory (IRT) model. Unlike raw percentage scores, IRT models treat achievement as a latent ability, accounting for varying item difficulty, item discrimination, and the pseudo-guessing parameter. This ensures that a student's score reflects their underlying mastery rather than the idiosyncratic difficulty of a specific test form.

To enable comparability, the baseline and endline instruments were linked via concurrent calibration. This psychometric process anchors both test forms to a common interval scale, allowing for the direct quantification of learning gains over the trial period. Final ability estimates were transformed to a scale with a mean of 500 and a standard deviation of 100 to facilitate interpretation across educational stakeholders.

Methodological Note: The Necessity of IRT

Utilising unlinked raw scores for impact evaluation is methodologically unsound, as any measured change would be confounded by differential test difficulty. IRT concurrent calibration resolves this by placing all assessments on a unified scale, isolating genuine learning progress. This analysis was performed by Oxford MeasurEd blinded to treatment assignment, ensuring that psychometric scaling remained objective and free from evaluative bias.

Surveys and Qualitative Protocols

To optimise the precision of the analytical model, we deployed a suite of baseline sociodemographic surveys designed for variance absorption and the facilitation of heterogeneity analysis.

To complement the quantitative findings, we implemented a qualitative protocol consisting of semi-structured Focus Group Discussions (FGDs) with the mathematics teachers near the end of the intervention period (see Qualitative Analytic Framework and Thematic Validation for the analytical methodology). The protocol for these discussions was designed to capture teachers' first-hand experiences of integrating Gemini into their lessons, including their use of the tool for lesson planning, their role in guiding student interactions during class, and their overall impressions and recommendations. The qualitative data serves to contextualise the quantitative results and identify potential moderators of efficacy that are not captured via field monitor logs or standardised assessments.

Logistical execution involved four virtual FGDs, each comprising a cohort of four to five teachers, with session durations ranging from 60 to 90 minutes. Data processing utilised Gemini 2.5 Pro for initial interview transcription, followed by a rigorous phase of manual verification by the research team. All educator responses were de-identified.

Implementation Protocol and Experimental Integrity

The validity of an RCT is fundamentally dependent on the precision of the operational execution. To safeguard experimental integrity, we engineered a suite of implementation protocols designed to mitigate contamination risks and ensure that the intervention delivery remained consistent with the preregistered research design. This framework helps ensure that any observed divergent learning trajectories are attributable to the causal impact of the treatment rather than operational variances.

Consent Protocols

To provide a primary ethical safeguard for the minor-aged cohort, we implemented a rigorous active parental consent protocol. This means the acquisition of signed authorisation from parents or legal guardians as a prerequisite for student participation, across both the treatment and control groups. This also serves as a methodological safeguard: implementing the consent protocol uniformly ensured that both groups were subject to the same participation filter, preserving the comparability established by randomisation. In contrast, in some evaluation designs, consent is collected only from treatment participants, with outcomes compared against routinely available or administrative data from the wider student population. This filters one arm by parental willingness to participate while leaving the other unfiltered, risking selection bias that compromises the equivalence of the trial arms.

Monitoring Implementation Fidelity

Implementation fidelity—defined as the degree to which the intervention is delivered in accordance with its experimental design—constitutes a critical determinant of the trial's internal validity. Sub-optimal tool utilisation in treatment clusters or unauthorised access to the treatment within control clusters (contamination) would result in a bias of the estimated treatment effect. To mitigate these risks, we deployed a multi-layered monitoring system designed to safeguard experimental integrity and minimise measurement error.

Field Monitor Deployment and Operational Support

One field monitor was stationed at each of the 12 participating institutions for the duration of the intervention. Monitors executed core operational protocols, including: (i) the preparation and distribution of hardware prior to treatment sessions; (ii) systematic recording of attendance for both experimental arms; (iii) reactive technical support; and (iv) comprehensive logging of operational variances—such as crossover between classrooms, device failures, and schedule deviations—via a standardised daily tracker.

To preserve realistic classroom dynamics and minimise observer effects, monitors were instructed to remain external to the classroom environment during active instruction. Entry was restricted to teacher-initiated requests for technical troubleshooting.

> **Methodological Note: Ecologically Valid Monitoring**
>
> The experimental design prioritised ecological validity by maintaining conditions that approximate standard pedagogical practice. Constant internal monitoring by field observers risks introducing Hawthorne effects, wherein teachers and students alter their behaviors due to the presence of an observer, thereby biasing the results. By limiting monitor involvement to reactive technical support, we reduced the risk that observed behaviours reflected the presence of an observer rather than the intervention itself.

Monitor Training and Instrumentation

Field monitors participated in a training program covering experimental design, non-contamination protocols, and ethical consent management. This included hands-on technical training in Gemini troubleshooting and the utilisation of digital tracking instruments to ensure standardised incident reporting across all schools.

Fidelity Tracking and Quality Assurance

Structured monitoring forms captured granular data for each treatment session, including instructor presence, student attendance, and instances of cross-arm contamination. Weekly aggregated summaries allowed the research team and EducAid to identify classrooms falling below the intended dosage thresholds and follow up with field monitors to understand the underlying barriers. These data were further enriched through two working sessions between field monitors to resolve technical barriers. To ensure continuous quality assurance, follow-up virtual reviews were conducted between field monitors and Fab AI. This iterative feedback loop allowed for the escalation of unresolved technical issues and the maintenance of high-fidelity implementation across different school environments.

> **Methodological Note: Mitigating Contamination**
>
> In cluster RCTs where the same instructor may facilitate both experimental arms, the risk of contamination is heightened. Our primary safeguards included: (i) centralised physical management of hardware by monitors (tablets were only distributed during treatment lessons); (ii) restricted access to Gemini credentials for control participants; and (iii) immediate reporting of protocol violations. Any deviations in school timetables or classroom assignments were logged in real-time to maintain the design-based identification of treatment effects.

Teacher Training and Pedagogical Protocol

We utilised a cascade model of professional development to optimise for institutional scalability. An expert trainer facilitated an intensive single-day technical session for a core cohort of EducAid trainers, who subsequently delivered a standardised single-day curriculum to all mathematics teachers in the study sample. This curriculum integrated three core technical components: (i) hardware and interface familiarisation (tablet operation and Gemini navigation); (ii) generative AI fundamentals, addressing both its capabilities and associated technical risks; and (iii) applied pedagogical integration for Guided Learning.

The pedagogical protocol proposed a structured four-part lesson framework: (1) a teacher-led introduction to establish learning objectives and check prior knowledge; (2) a main activity where students worked in pairs with Gemini; (3) a consolidation phase where the class discussed and reviewed what they had learned; and (4) a plenary to summarise key takeaways. Teachers were shown how to prepare each lesson in advance by defining clear learning objectives, writing starter prompts for students, and drafting question stems for the chalkboard to scaffold interactions with Gemini. The training also covered how to embed educational and regional context into prompts (e.g., specifying the student's grade level and the setting of Sierra Leone) so that Gemini could ground its responses appropriately.

Students worked in pairs, with "driver" and "navigator" roles that swapped each lesson. The driver typed whilst the navigator took notes and helped think through questions. This collaborative structure was grounded in evidence on the benefits of pair work for learning and was also a practical response to device constraints (a 2:1 student-to-device ratio).

Methodological Note: Training Intensity and Scalability

A single day of training is relatively light compared to more intensive professional development models. This was a deliberate choice, as we wanted to test the intervention under conditions that could realistically be replicated at scale, rather than under artificially intensive support and observation conditions that would be prohibitively expensive to maintain at scale.

While the current results demonstrate significant causal impact, it is hypothesised that more sustained pedagogical support could yield larger gains. Investigating the upper bounds of efficacy through varied instructional support models remains an open question for future exploration.

Data Collection Technical Specifications

To isolate the causal effect of Guided Learning from general digital literacy gains, all assessments were administered under standardised conditions via paper-based instruments. This methodological choice aimed to eliminate the risk of differential device effects, ensuring that the treatment group's eight-week exposure to tablets did not inflate achievement estimates through superior familiarity with digital interfaces. Paper-based administration further optimised the operational robustness of the study within the Sierra Leonean context, as it removed dependencies on local electricity grids and internet connectivity during assessment windows. Data collection was managed by Laterite.

Methodological Note: Digital Versus Paper Assessment

In contexts characterised by equivalent baseline digital literacy across experimental arms, digital administration protocols may offer superior operational efficiency and data integrity. However, to safeguard internal validity within interventions involving technology exposure, paper-based instrumentation may help eliminate confounding effects arising from differential device familiarity.

Analytical Framework

The analytical framework serves to translate the experimental design into robust statistical estimates capable of addressing the primary research questions. This section describes the methodological rationale for the preregistration of the analysis plan and summarises the core quantitative and qualitative approaches employed to ensure the precision of the study findings.

The Pre-Analysis Plan

A Pre-Analysis Plan (PAP) constitutes a formal technical codification of the statistical protocols and decision rules established prior to the accessibility of outcome data. The PAP for the current evaluation was preregistered with the AEA RCT Registry (AEARCTR-0016651) and finalised preceding the acquisition of endline data. This protocol specifies the statistical models, outcome definitions, subgroup analyses, and decision rules to be utilised in the impact analysis. The implementation of a PAP serves to mitigate the risks associated with specification searching and "p-hacking". This commitment ensures that the reported causal estimates reflect pre-defined hypotheses rather than arbitrary data mining. Our preregistered PAP defined the primary outcome as the IRT-scaled mathematics score and established the Intention-to-Treat (ITT) as the primary estimand. It further codified the treatment-on-the-treated (TOT) instrumental variables approach, the clustering of standard errors at the classroom level, and the inclusion of school-by-grade block fixed effects.

Methodological Note: Writing a Good PAP

A robust PAP should provide a codification of statistical protocols, including proactively incorporating formal contingency frameworks to address potential operational challenges, such as non-random attrition, instrumental variable weakness, or missing baseline data. By documenting which analyses were planned in advance, the PAP establishes a clear boundary between confirmatory and exploratory findings, helping readers to interpret results with the appropriate degree of evidential weight.

Quantitative Analytical Protocols

Intention-to-Treat Estimand

Intention-to-Treat (ITT) framework was used to quantify the causal impact of being assigned to the Gemini Guided Learning treatment condition on mathematics achievement. By evaluating participants based on their initial random allocation—regardless of subsequent adherence or actual tool utilisation—the ITT preserves the randomisation and therefore maintains the causal interpretation.

Even if some treatment students were absent from Gemini lessons, the ITT remains an unbiased estimate of the effect of being assigned to the program as it was actually delivered, inclusive of real-world non-compliance.

We executed three hierarchically structured regression specifications to ensure inferential robustness. *Specification 1* is a design-based model including only the treatment indicator, estimated for all students present at endline. *Specification 2* introduces baseline mathematics and English literacy scores as precision-enhancing covariates. It is estimated for a balanced panel of students who have both the baseline and endline scores. *Specification 3* incorporates granular demographic controls: gender and age. All models utilise standard errors clustered at the classroom level and school-by-grade block fixed effects to account for the stratified experimental design. This multi-model approach enables the assessment of sensitivity to covariate conditioning and student attrition between baseline and endline assessments. Additionally, a Difference-in-Differences (DiD) estimation was conducted on the balanced panel to further validate the robustness of the treatment effect.

Treatment-on-the-Treated (TOT) and Instrumental Variables

To isolate the causal efficacy of actual tool exposure—defined as cumulative hours of Guided Learning engagement—we implemented a **Treatment-on-the-Treated (TOT)** analysis. Recognising that dosage accumulation is determined by student characteristics, including their motivation and school attendance, a naive comparison between high- and low-exposure students would yield biased results. Consequently, we employed an instrumental variables approach, using random classroom assignment as an instrument for individual student dosage.

Random assignment to a treatment classroom strongly predicts how many hours of Gemini a student receives, but is unrelated to individual student characteristics, thus neutralising confounding student-level traits. The resulting instrumental variable coefficient quantifies the causal effect of an additional hour of the full intervention for "compliers" whose exposure was determined by their classroom's treatment status. We also estimated the impact of achieving the intended dosage threshold of 12 hours using the same instrumental variable strategy, providing a rigorous test of the intervention's intended instructional density.

> **Methodological Note: Dual Reporting of ITT and TOT**
>
> The concurrent reporting of ITT and TOT estimands is important for a comprehensive evaluation of pedagogical interventions. The ITT provides the policy-relevant estimate of program impact under real-world conditions, encompassing factors such as operational friction and non-compliance. A robust ITT demonstrates that the intervention package is effective as a deliverable educational service, providing a scientific foundation for longitudinal scalability and institutional adoption. Where TOT estimates are significant but ITT estimates are not, this may suggest that effects are concentrated among students who engaged with the tool, though the reasons for limited engagement would require further investigation.

Estimation of Heterogeneous Treatment Effects

To identify potential non-uniformity in the intervention's impact, we executed a series of pre-specified heterogeneity analyses. These analyses decomposed the treatment effect across primary student-level dimensions, including grade level (JSS1 vs. JSS2), baseline English reading comprehension, baseline mathematics proficiency, gender, and age. This allows us to explore whether the intervention's efficacy varies across diverse student profiles. However, it should be noted that the experimental

design was primarily powered for the aggregate ITT effect; consequently, estimates of heterogeneous treatment effects may possess lower statistical power and should be interpreted as exploratory rather than definitive evidence of subgroup-specific causal impacts.

Clustered Variance and Stratification

The experimental design utilised the classroom as the primary unit of randomisation. Students within the same classroom share common environmental and pedagogical influences, violating the assumption of independent observations. Failure to adjust for this would understate the true uncertainty around the treatment effect, leading to artificially narrow confidence intervals and inflated significance levels. Consequently, all regression specifications employ standard errors clustered at the classroom level.

Furthermore, to align the analysis with the randomisation design, we incorporated school-by-grade block fixed effects. This stratified approach absorbs systematic differences across schools and grades, ensuring that the treatment effect is identified from within-block comparisons. This enhances the precision and internal validity of the causal estimates.

> **Methodological Note: The Criticality of Clustered Inference**
>
> A fundamental requirement for rigorous clustered RCT evaluation is that the analysis accounts for the clustered assignment. Analysing student-level data without appropriate clustering frequently leads to spurious significance findings due to artificially narrowed confidence intervals. It is imperative to maintain design-based identification by clustering standard errors at the level of randomisation and including fixed effects for all stratification variables used during the experimental allocation.

Assessment of Attrition

Attrition refers to the loss of participants between the start and the end of a study. In our context, it means students who were assessed at baseline but were not available for the endline assessment, whether because they left the school, were absent on the day of testing, or any other reason. Attrition matters because if the students who drop out are systematically different from those who remain, the final sample may no longer be balanced between treatment and control, undermining the comparability that randomisation was designed to achieve.

It is important to distinguish between student-level mobility and cluster-level attrition (e.g. the loss of an entire classroom). While student-level attrition is anticipated in field studies and, if roughly equal across arms, is manageable, cluster-level attrition can structurally underpower the trial by removing a primary unit of randomisation. In our case, cluster-level integrity was perfectly maintained with zero classroom attrition.

Attrition was tracked by measuring how many students in each arm completed the endline assessment, reporting follow up rates by treatment arm and by school and grade block. The realised follow-up rates demonstrated high stability: 90.7% for the control group and 93.3% for the treatment group. To verify the orthogonality of participant loss to treatment assignment, we executed a formal attrition analysis regressing endline observation on the treatment indicator and baseline covariates (prior test scores, gender, and age). This helps determine whether attrition is random or whether it is correlated with factors that could bias the results. The results showed that attrition was 3 percentage points higher in the control group than in the treatment group. However, this difference

was not significantly related to students' baseline mathematics levels (captured by the interaction term between the treatment condition and the baseline mathematics score, measuring whether the students who dropped out had systematically higher- or lower-performance at baseline by condition). This suggests that the students lost to attrition in each arm were not systematically different at baseline. This reduces concerns that attrition biased the ITT estimates, and we therefore did not apply weighting or trimming adjustments.

Methodological Note: Decision Rules for Handling Attrition

Our PAP specified a hierarchical framework for managing attrition based on its realised magnitude and nature. Under conditions where follow-up rates are balanced and non-differential, a standard complete-case analysis can proceed. Some students will be missing, but if the loss is roughly equal in both arms and unrelated to treatment, the remaining sample preserves the comparability created by randomisation.

In instances of differential attrition (our PAP set the threshold of 10 percentage point difference between arms), the internal validity of the results is potentially compromised. Such imbalances can potentially be addressed by the application of appropriate statistical adjustment or bounding methods (e.g. Inverse Probability Weighting, Lee bounds). It is good practice to specify these adjustments as part of study preregistration (see The Pre-Analysis Plan). Adherence to the preregistered decision rules ensures that the impact evaluation remains transparent and resilient to the systematic biases inherent in longitudinal field research.

If attrition is significantly predicted by baseline characteristics (such as prior test scores, gender, or age) but not by the treatment assignment itself, methods such as Inverse Probability Weighting (IPW) can be used to re-weight the endline sample. This adjustment accounts for potential non-random missingness (the fact that the students who dropped out may differ from those who stayed), by calibrating the endline sample to match the baseline demographic and ability distribution, thereby preserving the generalisability of the findings within the specified sampling frame. This assumes that missingness is conditionally random given the observed baseline information.

The most critical scenario occurs when attrition is both differential across treatment arms and related to baseline outcome levels, such as prior mathematics achievement. In this case, the observed endline sample may no longer be comparable across experimental arms, raising concerns that estimated treatment effects partly reflect selective retention rather than the intervention itself. When attrition differs by treatment status but is not systematically related to baseline outcome levels or other key covariates, the risk of attrition bias is reduced, although the issue should still be reported transparently.

Qualitative Analytic Framework and Thematic Validation

Qualitative data derived from educator FGDs were subjected to thematic analysis (Braun and Clarke, 2006). We used a codebook approach using a strict manual coding protocol. AI integration was explicitly restricted to the generation of initial de-identified transcripts via Gemini 2.5 Pro. All subsequent interpretive phases—including coding, hierarchical grouping, and thematic synthesis—were executed exclusively by human researchers to preserve the integrity of the inductive approach.

The analysis proceeded following a formalised seven-stage protocol: (i) **Familiarisation**, in which all coders read every transcript and recorded free-form observations; (ii) **Open Coding** on a subset of transcripts, followed by a team meeting to consolidate a candidate code list; (iii) **Codebook Development**, in which codes were grouped, defined, and organised hierarchically by two analysis leads; (iv) **Pilot Coding** to test the codebook's consistency and revise definitions before full application; (v) **Systematic Coding** of all transcripts; (vi) **Reconciliation** in which disagreements were resolved by each coder pair; and (vii) **Thematic Refinement**, in which the analysis leads finalised the thematic structure, and selected representative quotes. This staged process was designed to balance the interpretive flexibility that qualitative data requires with the transparency and consistency needed to make the analysis reproducible.

Discussion

This playbook suggests that maintaining high levels of scientific rigour on a condensed timescale when running an RCT for an AI-powered classroom intervention is feasible, but necessitates a series of methodological trade-offs.

We discussed the methodological rationale and decision rules used by our teams to design and implement the Sierra Leone RCT (LearnLM Team, Google and Fab AI, 2026), which we hope will serve as a useful resource for others to use and build upon. Alongside using robust statistical methods, it was also important to structure the evaluation process in a way that safeguards its scientific integrity. This was achieved through two controls: institutional insulation and preregistration. By working with a modular partnership structure, with separation of roles between implementation, data collection, and analysis, the study aimed to avoid potential evaluator bias and ensure that outcome measurements remained as objective as possible.

Specific methodological trade-offs, including the selection of a two-arm experimental design and a criteria-based framework for sampling the participating schools, were calibrated to balance statistical power with logistical constraints. While a multi-arm design could have disaggregated the contributions of the intervention's individual components, the two-arm model maximised the precision of the aggregate treatment effect estimate within the available sample size. Similarly, targeting schools with prerequisite infrastructure and a preexisting relationship ensured enhanced implementation fidelity, establishing a "minimum viable" baseline for future replication in more diverse school populations.

Ultimately, while the results of this first RCT in Sierra Leone (LearnLM Team, Google and Fab AI, 2026) establish a robust causal baseline for learning gains for an in-classroom teacher-led AI intervention, they underscore the necessity for subsequent research cycles. The questions we focused on–whether the intervention works, whether it works equitably, and how impact varies with use–are foundational, but they are not the whole picture. The field will need evidence on mechanisms, on cost-effectiveness, on longer-term effects, on impacts on teachers' practice, on other outcomes such as motivation and self-efficacy, and on potential harms including over-reliance and changes in academic

integrity. Some of these questions are answerable through further RCTs; others will need qualitative, mixed-methods, or longitudinal designs. This document is an early input to a wider collective effort, not a definitive account of how AI in education should be evaluated.

References

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa.

Lily Fesler, J Martinez Claeys, Chris Agnew, and Susanna Loeb. The evidence base on ai in k-12: A 2026 review, 2026.

Ben Kelcey, Zuchao Shen, and Jessaca Spybrook. Intraclass correlation coefficients for designing cluster-randomized trials in sub-Saharan Africa education. *Evaluation Review*, 2016. doi: 10.1177/0193841X16660246.

LearnLM Team, Google and Fab AI. Teaching with Gemini: Measuring the impact of Guided Learning on student mathematics progress in Sierra Leone, May 2026. URL http://goo.gle/LearnLM-SierraLeone-May26.