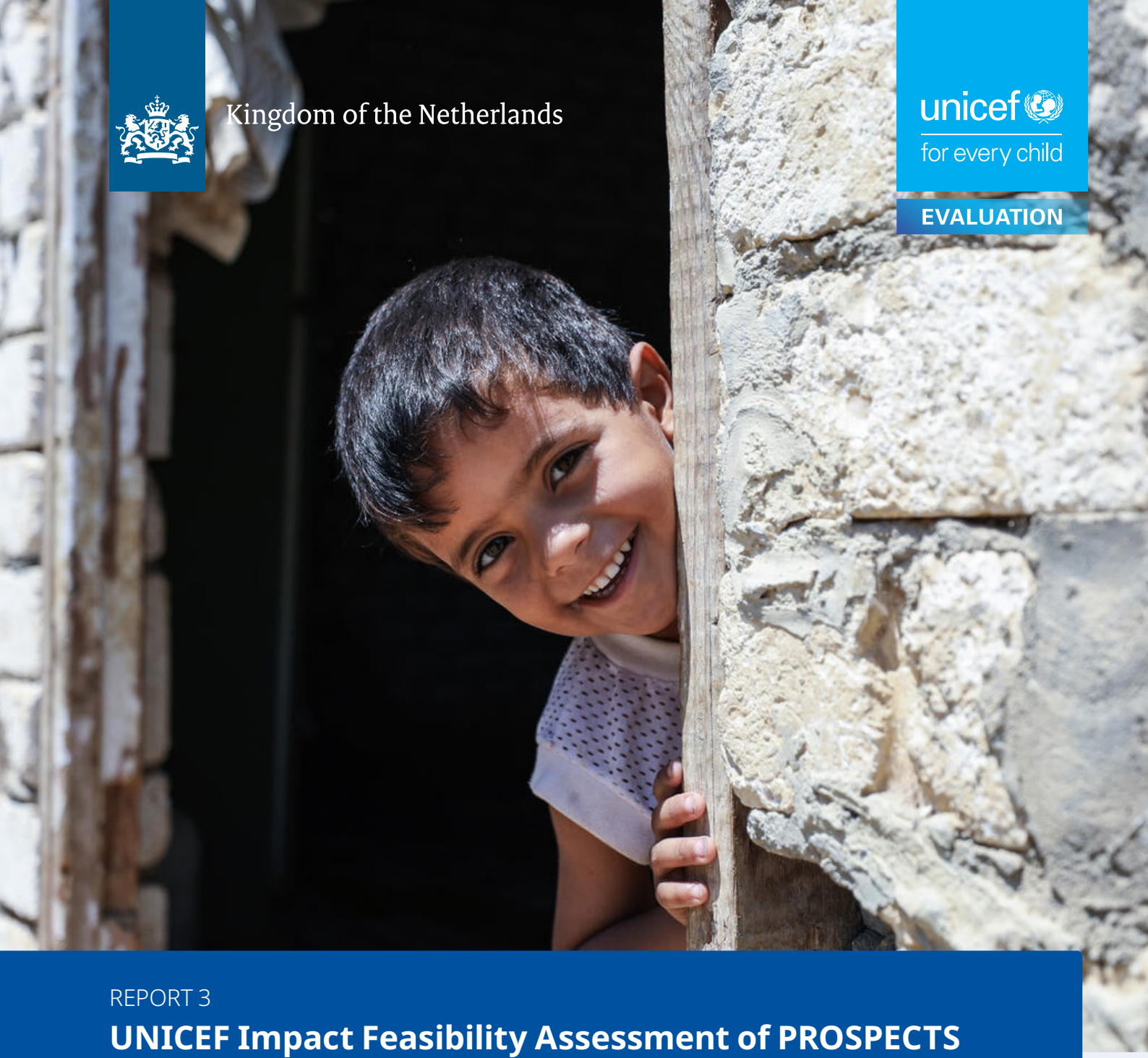


Kingdom of the Netherlands

unicef
for every child

EVALUATION

REPORT 3

UNICEF Impact Feasibility Assessment of PROSPECTS

Egypt Case Study

This report was prepared by Preksha Golchha, Jose Victor C. Giarola (Economic Policy Research Institute), with the guidance and supervision from Eduard Bonet Porqueras (Evaluation Office, UNICEF) and support from Innocent Kaba (Displacement and Migration Hub, UNICEF). The Impact Feasibility Assessment was commissioned by Tasha Gill and Rhonda Fleischer, the global lead and coordinator (Displacement and Migration Hub, UNICEF), and managed by Eduard Bonet Porqueras, with support from Innocent Kaba.

Acknowledgments: Earlier drafts of this report benefited from the valuable feedback provided by Amber Peterman, Andrew Kaiser-Tedesco and Dalia Bayoumi, who nonetheless bear no responsibility over any flaws that this published version may have. We would like to thank all the UNICEF PROSPECTS programme staff who have kindly shared ideas and especially access to documentation; Lauren Farwell and Khaled Khaled, Finally, we would like to express our sincere appreciation to the Netherlands Ministry of Foreign Affairs for funding, facilitating and guiding the PROSPECTS innovative partnership among United Nations agencies and multilateral financial institutions to find sustainable solutions to improve the well-being of displaced and host populations.

Suggested citation: United Nations Children's Fund, 'Impact Feasibility Assessment of PROSPECTS: Egypt Case Study', UNICEF: New York, 2025.

Cover photo: © UNICEF/UNI623680/Ahmed Mostafa
Design and layout: Elena Panetti

Please contact:
UNICEF
Evaluation Office
3 United Nations Plaza
New York, NY 10017, USA
Email: evalhelp@unicef.org
Website: www.unicef.org/evaluation/

Table of Contents

01

Motivation and Programme Description

As of 2025, Egypt hosts a substantial population of refugees and asylum-seekers—over 900,000 individuals from countries including Sudan, Syria, South Sudan, and others.[1] These forcibly displaced populations (FDPs) mostly reside in urban areas alongside host communities, with significant concentrations in Greater Cairo (Cairo, Giza, Qalubiya), Alexandria, Damietta, and Aswan.

The Government of Egypt allows refugee children to enrol in public schools, but the influx of new students has strained an education system already facing challenges. Public and community schools in vulnerable areas often grapple with overcrowded classrooms, limited resources, and inadequate infrastructure. Such conditions can undermine the quality of education and contribute to student dropouts, especially among disadvantaged groups. Socio-economic barriers also play a major role – for low-income Egyptian families and refugee households alike, costs of schooling (uniforms, supplies, transportation) and the opportunity cost of children's labour can lead to early dropout. The cumulative result is a risk of lower enrolment and retention in school for refugee children and poor host-community children, threatening to create a "lost generation" without proper education. This context underpins the need for targeted interventions to keep children in a protective learning environment and ensure they not only enrol in school but also stay and succeed.

The PROSPECTS Partnership was established as a multi-year, multi-country initiative to transform the way we respond to forced displacement crises. Financed by the Government of the Netherlands, PROSPECTS brings together five agencies – the International Labour Organization (ILO), the International Finance Corporation (IFC), the UN Refugee Agency (UNHCR), the UN Children's Fund (UNICEF), and the World Bank – to work in concert with national governments. It aims to bridge humanitarian and development efforts, enhancing the socio-economic inclusion of refugees and asylum-seekers while strengthening host communities' resilience. Education is one of PROSPECTS' key pillars, alongside protection and social protection and economic inclusion, given its central role in fostering self-reliance and social cohesion. In Egypt, PROSPECTS supports the national commitment to inclusive education (aligned with Egypt's Education Sector Plan and SDG4) by integrating refugees into public systems and improving education quality for all vulnerable children. There is strong interest from both development partners and the Government of Egypt in understanding "what works" to improve educational outcomes in displacement-affected communities, making rigorous evaluation a priority.

Under the PROSPECTS initiative in Egypt, UNICEF (in close collaboration with the Ministry of Education and Technical Education, UNHCR, and other partners) is implementing the "Learning in a Protective Environment for Increased Retention" programme. The goal of this integrated programme is to increase the enrolment and retention of children in schools by creating safer, more inclusive learning environments and alleviating economic barriers to schooling. Rather than a single intervention, it combines three complementary components that together address the multi-faceted causes of dropout:

1 Available here

Teacher Training for Inclusive and Protective Education: The programme provides extensive in-service training to teachers (and school staff) in target schools on child-centred, inclusive pedagogy and classroom management. Teachers are trained in techniques to engage learners who are at risk of falling behind or dropping out – for instance, psychosocial support, identifying and supporting children with learning difficulties, and gender-responsive teaching methods. By strengthening teacher capacity, the programme aims to improve classroom instruction quality and make school a more supportive place for all learners.

Cash Grants to Vulnerable Families: To tackle financial barriers, the programme provides cash support to the most vulnerable households with school-aged children (both refugees and Egyptians) in the target communities. These education-linked cash grants are intended to offset direct costs of schooling (such as fees, uniforms, and learning materials) and reduce the pressure for children to engage in work or other coping mechanisms. This cash "safety net" component is designed to incentivize enrolment and regular attendance – leveraging global evidence that cash transfers can boost school participation. It also provides a modest economic relief to refugee families and host community households facing hardship, contributing to their overall well-being.

WASH Improvements in Schools: The programme invests in improving Water, Sanitation and Hygiene (WASH) facilities in the selected schools. This includes constructing or refurbishing gender-segregated latrines, ensuring clean drinking water supply, and promoting hygiene practices among students. Safe and adequate WASH facilities are fundamental to a "protective" learning environment – they particularly help keep girls in school (reducing absences related to hygiene issues) and improve all students' health, dignity, and concentration. Based on initial assessments, this arm of the intervention will be piloted in five schools, delivered in close consultation with FDPs and HCs. Results from this pilot will inform future interventions and guide government initiatives towards sustained improvements in school-based WASH through additional funding sources.

Geographically, the programme is focused on community and public schools in Greater Cairo (Cairo, Giza, Qalubiya governorates), Alexandria, Damietta, and Aswan. These areas were selected due to their high concentration of refugee families and underserved host communities, as well as varying regional challenges. Both formal public schools and community-based education centres are included to ensure broad coverage of vulnerable children. By targeting schools that serve both refugees and Egyptians, the programme explicitly fosters inclusive education – refugees attend the same schools as host-community peers, accessing the same improved facilities and instruction, which helps integrate them into the national education system. Meanwhile, host-community children benefit equally from the programme inputs, mitigating any resentment and building social cohesion through shared positive experiences in school.

Box 1

Learning in a Protective Environment for Increased Retention Programme Selection

The Learning in a Protective Environment for Increased Retention programme was selected for Stage 3 of the Impact Feasibility Assessment via a standardized process consisting of: 1) an analysis of **country context** and 2) an **intervention context** by mapping of PROSPECTS interventions in the 8 PROSPECTS countries (with 33 interventions included in total). Key considerations in the country context were: 1) Political interest and will from government and partners to understand what works and to what extent ministries would support the system changes necessary to scale up a successful intervention; 2) the operational facility, including potential risks to a successful evaluation; 3) the prioritization based on knowledge gaps – as assessed on the IFA Stage 1 Rapid Review; and 4) the national data and evaluation capacity, including the existence of strong research institutions in the country and high-quality sources of secondary data. Key considerations in the intervention context were: 1) the scale and scalability of programming, which considers whether interventions are large enough to support rigorous impact evaluations; 2) previous or planned impact evaluations; 3) the potential for future expansion; 4) the knowledge gains, which prioritizes interventions capable of addressing knowledge gaps identified during Phase One of the IFA (Rapid Review);[2] and 5) the type of programming, which assessed interventions based on ToC Integration and partners integration.

- **Country Context:** Egypt has scored moderately high in terms of country context, with high prioritization based on knowledge gaps, moderately high political will and high national evaluation capacity.
- **Intervention Context:** The Learning in a Protective Environment for Increased Retention programme was ranked as the top priority in Egypt (among five interventions included), as it met almost all assessment criteria, including those relating to scale and scalability, plans for future expansion, no existing impact evaluation and knowledge gains and type of programming based on partners and ToC integration.

The Learning in a Protective Environment for Increased Retention programme is one of secondary priority interventions shortlisted for Stage 3 (alongside three priority interventions interventions in Ethiopia, Jordan and Uganda) for which impact evaluation plan is being developed.

Importantly, this integrated approach addresses multiple impediments to education simultaneously. The teacher training component improves the quality of instruction and attentiveness to students' needs; the cash grants address demand-side barriers by reducing households' economic burdens; and the WASH upgrades improve the safety and health conditions of schools. The expectation is that together these arms of the intervention create a virtuous cycle: children are more motivated and able to attend school regularly, and once there, they find a conducive environment that encourages them to continue attending and learning. This holistic design reflects a core principle of PROSPECTS – that protecting and educating children in displacement contexts requires coordinated action on several fronts (education, child protection, and social protection).

The Learning in a Protective Environment for Increased Retention programme was selected as a promising secondary intervention to develop an impact evaluation plan based on the systematic progress in Stage 2 of the Impact Feasibility Assessment, which includes the assessment of both country- and intervention-level factors (see Box 1).

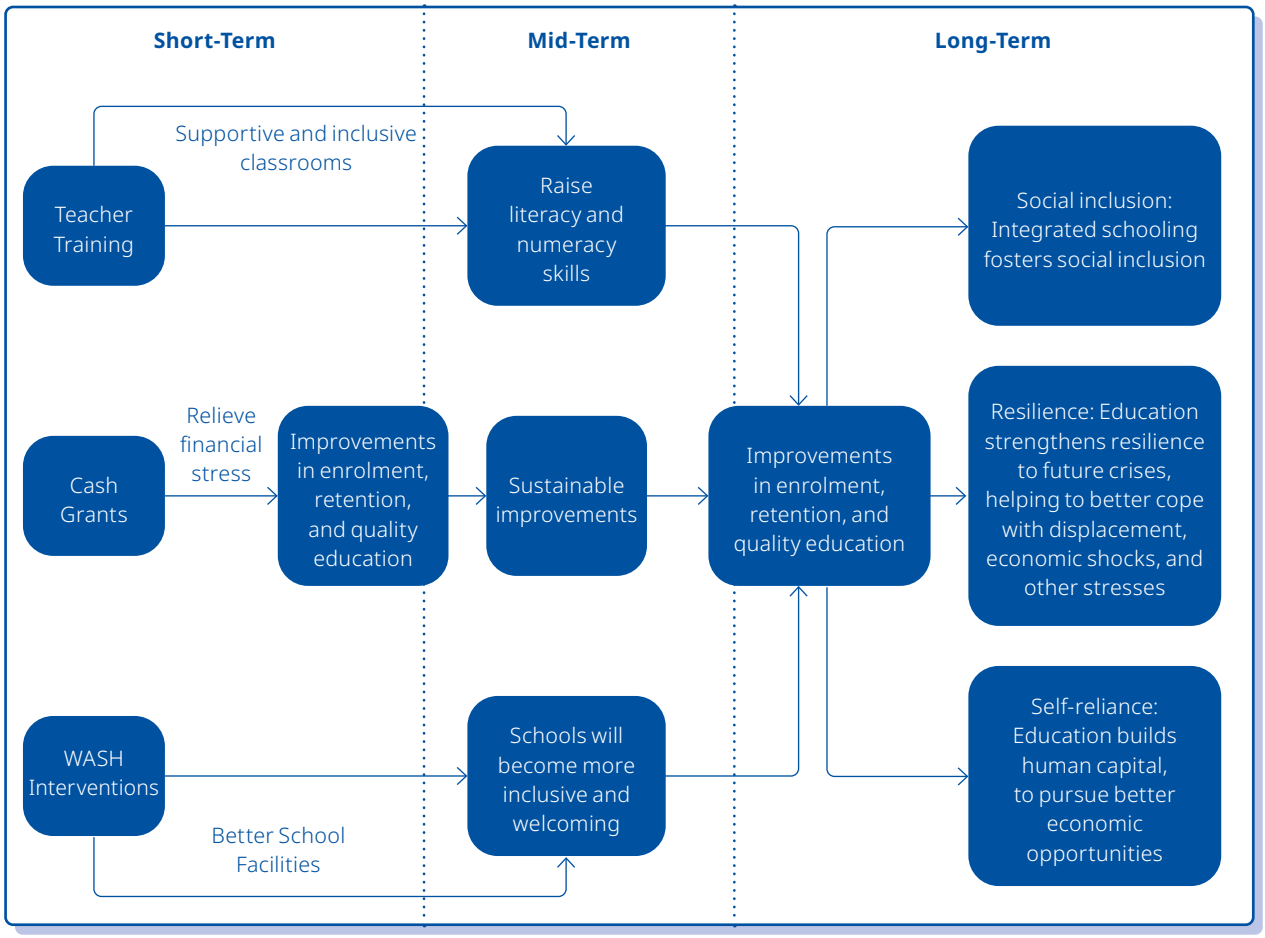
2 It is important to note that evidence gaps identified in the rapid review were limited to PROSPECTS countries and forcibly displaced populations. A detailed explanation of this decision can be found in the Section 2.1 "Selection Criteria" of the Stage 1 Impact Feasibility Assessment Report.

02

Impact Pathways and Key Outcomes of the Intervention

The theory of change for the Learning in a Protective Environment programme is anchored in the broader PROSPECTS Partnership results framework, which ultimately seeks to enhance self-reliance, social inclusion, and resilience among refugees and host communities. Education is a critical pillar toward these high-level impacts. Figure 1 (below) illustrates the programme's impact pathways, linking the short-term outcomes achieved through the intervention to intermediate outcomes in education, and finally to the intermediate impacts (PROSPECTS' overarching goals). In narrative form, the impact pathway is described below.

Figure 1. Impact Pathways



The programme aims for immediate improvements in student enrolment, attendance, and retention by addressing key barriers through teacher training, cash grants, and WASH interventions. Teacher training promotes engaging, supportive, and inclusive classrooms, motivating students—particularly refugees and girls—to regularly attend and remain in school. Cash grants relieve financial stress, enabling families to afford essential school expenses, reducing dropout rates, and encouraging new enrolments among vulnerable students. Improved WASH facilities decrease absenteeism due to illness and hygiene concerns, particularly benefiting adolescent girls by enabling consistent school attendance. Collectively, these interventions should rapidly translate into measurable gains in attendance, lower dropout rates, and improved grade progression.

Over the mid-term, the programme will shift from ensuring enrolment to strengthening educational quality and inclusion. Sustained teacher training will raise literacy and numeracy skills, improving overall learning outcomes. Schools will become more inclusive and welcoming, integrating refugee and marginalized students, narrowing achievement gaps, and encouraging continuity, especially among girls. Strengthened school management practices—such as attendance monitoring and student participation—will enhance the resilience of educational systems, ensuring sustainable improvements beyond immediate interventions. This increased access to quality education supports PROSPECTS' objectives of equitable and inclusive education for all students.

In the longer term, sustained improvements in enrolment, retention, and quality education will contribute significantly to self-reliance, social inclusion, and resilience among refugee and host communities. Education builds human capital, enabling youth—both refugee and host—to pursue better economic opportunities and reduce dependency on aid. Integrated schooling fosters social inclusion, breaking down barriers between refugee and host populations, promoting community cohesion. At individual, family, and systemic levels, improved education strengthens resilience to future crises, helping children and communities better cope with displacement, economic shocks, and other stresses. Though these impacts extend beyond the immediate evaluation scope, early indicators such as improved educational transitions, reduced gaps between vulnerable groups, and strengthened community acceptance will signal progress toward PROSPECTS' long-term vision.

03

Evaluation Questions

The evaluation will focus on assessing the impacts of the **Learning in a Protective Environment programme**, examining whether it achieves its intended outcomes and contributes meaningfully to PROSPECTS' broader goals of self-reliance, social inclusion, and resilience. Guided by the programme's impact pathway, the evaluation will answer the following key questions:

- **How has the programme affected student enrolment, attendance, and retention in targeted schools?**
  - Have dropout rates decreased more in intervention schools compared to matched comparison schools?
  - Are students in programme schools more likely to progress smoothly through grade levels?
- **To what extent has the programme improved educational quality, particularly teaching practices and student learning outcomes?**
  - Are trained teachers demonstrating more inclusive, participatory, and child-centred teaching practices compared to teachers in similar non-programme schools?
  - To what extent do students in intervention schools improve their learning outcomes in key subjects such as literacy and numeracy?
- **What mechanisms within the school environment explain observed impacts on enrolment, attendance, and learning?**
  - Which programme components—teacher training, cash grants, or WASH improvements—are most strongly associated with improved student outcomes?
  - How do changes in teachers' attitudes, classroom practices, or community engagement contribute to creating safer, more inclusive learning environments?
- **How have improvements in WASH facilities specifically influenced student attendance, retention, and educational participation?**
  - Are adolescent girls and refugee students especially benefiting from improved WASH facilities, enabling more consistent attendance and reducing dropout?
- **How do programme impacts differ between refugee students and host-community students?**
  - Are refugee children experiencing comparable or greater gains in enrolment, retention, and learning outcomes compared to host-community children?
  - What specific factors (such as language barriers, displacement experiences, household support) might explain observed differences in outcomes?

- **To what extent does the programme differently impact girls and boys in intervention schools?**
  - Are girls benefiting equally or more substantially in terms of attendance, retention, and academic performance?
  - How do gender-sensitive interventions (e.g., improved WASH facilities or inclusive teaching practices) contribute to narrowing educational gaps?
- **What early evidence is there that the programme contributes to broader social inclusion and resilience outcomes?**
  - Are students, families, and educators reporting improved integration and social cohesion between refugee and host communities?
  - Do participants perceive that improved educational participation is enhancing their resilience and capacity to cope with economic or social challenges?

These evaluation questions will be addressed through a combination of quantitative analysis and qualitative exploration, emphasizing the identification of causal impacts, understanding the mechanisms driving these impacts, and exploring subgroup differences. All questions align closely with the PROSPECTS Theory of Change, ensuring the evaluation's findings provide actionable insights to policymakers and practitioners on how effectively the programme contributes to broader development goals in education and beyond.

04 Evaluation Design

Key Considerations for Clarification Before Finalizing Evaluation Design

This section is based on information currently available from PROSPECTS programme documents (e.g. the Egypt multi-year country programme plan). Communication with the country office during the preparation of this report was limited, so the proposed evaluation design relies primarily on desk review. Consequently, important information gaps remain – particularly regarding the target beneficiaries, implementation plan, and rollout timeline of the "Learning in a Protective Environment for Increased Retention" programme. Addressing these gaps through discussions with the country team is essential. The following non-exhaustive list of questions highlights key areas that need clarification before finalizing the impact evaluation design:

- **Implementation timing and phasing:** What is the planned schedule for rolling out the programme components across the target areas? Will all components – teacher training, cash transfers, and WASH facility upgrades – be launched concurrently in each school, or will they be introduced sequentially (e.g. training first, then cash assistance, followed by WASH improvements)? Furthermore, what is the duration of each intervention component? For instance, will teacher training and cash support be delivered over multiple academic years, or concentrated within a single school year as a pilot?
- **Geographic coverage and school selection:** How many schools are targeted in each of the focus regions (Greater Cairo, Alexandria, Damietta, and Aswan), and what criteria were used to select these schools? Are the chosen schools among the most vulnerable (e.g. those with high refugee enrolment or poor infrastructure), or were they selected for convenience/pilot purposes? Clarifying the distribution of intervention schools by governorate – and whether implementation will be simultaneous across regions or staggered – is important for ensuring the evaluation captures context-specific effects and that comparison groups can be appropriately matched to each area.
- **Eligibility and targeting criteria for households:** On what basis are beneficiary households selected for the cash grants? What defines "vulnerable" in this context – is it based on poverty thresholds, refugee status, school attendance records, or other vulnerability assessments? It is crucial to know whether both refugee and host-community families are eligible for cash assistance under the programme (and in what proportions), and whether any conditionality is attached (for example, requiring children's school attendance). These criteria will affect not only programme operations but also the evaluation sampling strategy (to ensure that the surveyed population includes the intended beneficiaries).
- **Integration of intervention components:** To what extent will the three components (teacher training, cash transfers, and WASH improvements) overlap in the same schools and communities? For example, will the five pilot schools receiving WASH facility upgrades also have teachers trained and cash distributed to their students' families, thereby receiving the full package? Or are some components implemented in distinct sites (e.g. WASH only in a subset of schools)? Understanding how the components converge is critical, as it determines whether the evaluation should consider the programme as a single integrated treatment or if there are variations in exposure (which might require subgroup analysis or adjusted impact

estimates for schools that did not receive all components).

- **Target groups and school levels:** Which student populations and grade levels is the programme focusing on? Are all grades within the selected schools included, or is the intervention targeting specific levels (for instance, upper primary and lower secondary grades where dropout risk is highest)? Additionally, the programme spans both formal public schools and community-based education centres – will the intervention approach (and evaluation) be uniform across these different school types, or are there adaptations for community schools? Clarification on the target groups (including the ratio of refugee to Egyptian students in these schools and any particular focus on girls or other vulnerable subgroups) will ensure the evaluation design appropriately stratifies samples and addresses equity considerations.

Addressing the above questions in collaboration with the country office will help tailor the evaluation design to Egypt's context, ensuring it is practical and able to capture the programme's true effects.

The potential evaluation design options outlined below, are seemingly feasible options with the information available at this moment. They may need to be refined or even revised once the information gaps and operational details are available, favouring whenever feasible and retaining good quality designs, those design options with lighter requirements in data collection. The designs recommended should use counterfactuals, as this element is critical in the assessment of causality in the changes observed (or not observed). Nevertheless, the following designs are not always the best ideal options in this sense, but just feasible options with the information at hand.

Overall Design

The evaluation will employ a mixed-methods, quasi-experimental design to rigorously assess the impact of the programme. A mixed-methods approach is well-suited for this multi-component intervention as it allows us to quantify the "what works" in terms of outcomes, while also exploring the "how and why" through qualitative inquiry. Quantitative methods (described below) will be used to establish causality – isolating the programme's effect on key outcomes like retention – by comparing intervention beneficiaries to a credible counterfactual (comparison group). Qualitative methods will be integrated to capture the experiences of teachers, students, and families, shedding light on implementation processes, contextual factors, and unintended effects that numbers alone cannot reveal. This dual approach ensures we not only measure impact but also understand the mechanisms behind observed changes and the perspectives of those involved.

Quantitative Evaluation Designs

Ideally, impact evaluations of programmes like this are done with an experimental design for maximum internal validity. At this stage, EPRI does not have full information on the implementation plan, and, thus, we are unable to infer if the intervention has already started to be implemented. This has consequences for the impact evaluation design. We assume two scenarios and propose 3 different designs depending on them. First, we assume that interventions have not yet started, and that randomization of schools is still possible. Under this scenario, EPRI proposes a Randomized Control Trial (RCT) as a feasible option for impact evaluation. Secondly, we assume that the pilots are already taking place and, thus, under this scenario the implementation in Egypt is already underway in all targeted areas, and true randomization was not built into the programme. Recognizing the practical constraints, our secondary design is a quasi-experimental Propensity Score Matching with Difference-in-Differences (PSM-DiD) approach. This combines matching (to select a comparable control group) with a difference-in-differences analysis (to net out baseline differences and trends), providing a robust estimate of impact. Additionally, we have a fallback design (Retrospective Matching) in case baseline data for DiD is unavailable or other issues arise; this would rely on endline comparisons using statistical matching and recall data.

The Retrospective Matching is presented as a last resort review option. It is considered a design of last resort when prospective methods are not possible, as its causal inference strength is significantly lower. The absence of a true baseline and randomization means we have to make stronger assumptions. Recall bias is a concern: respondents may not accurately remember past conditions, which could blur true changes. Unobserved differences can also remain – we might not capture all how program communities differed from non-program ones initially. As a result, impact estimates from this design are interpreted more cautiously, indicating potential effects rather than definitive proof.

The evaluation is designed to be adaptive – the evaluation team will implement the strongest feasible design given field conditions, while having fallback options to maintain rigor if the ideal conditions cannot be met. In all scenarios, we plan to integrate baseline and endline data collection (except in a purely retrospective case) to observe changes over time. Below, we describe each design option, including methodology, requirements, feasibility, and potential limitations, and conclude with the rationale for the chosen approach.

Primary Design: Randomized Control Trial (RCT)

If conditions allow for true experimental design, an the primary evaluation approach is a clustered Randomized Controlled Trial (RCT) at the school level. Under this design, a subset of schools that have not yet received the programme would be randomly selected to implement the full intervention package, while others serve as a control group. This scenario assumes that programme rollout can be controlled such that some eligible schools temporarily delay receiving the intervention for the sake of the evaluation. Random assignment of the intervention would provide the strongest evidence of impact by ensuring that, on average, the treatment and control groups are equivalent in both observed and unobserved characteristics.

Methodology and Implementation: The RCT would be organized around the remaining schools in the target regions that have not started the programme. From this pool, for example, ten schools could be chosen for the study, with five schools randomly assigned to receive the full package of interventions (teacher training, cash grants to their students' households, and WASH facility improvements), and the other five schools assigned to serve as controls during the evaluation period. The implementation steps would be as follows:

Baseline (pre-intervention): Conduct a baseline survey and collect key metrics in all selected schools before the programme begins. This establishes pre-intervention levels of outcomes (e.g. enrolment, attendance, dropout rates, and learning indicators) and verifies that the treatment and control groups are statistically similar at the start, given randomization.

Intervention rollout: Immediately after baseline, deliver the integrated programme in the five treatment schools. This means providing the teacher training to all relevant educators in those schools, disbursing the education-focused cash grants to the identified vulnerable households connected to those schools, and implementing the planned WASH infrastructure upgrades on the school premises. The control schools continue with "business-as-usual" – they do not receive any of the new programme components

during this phase. Implementation would typically cover one full academic year (or another defined period), during which the programme is expected to affect student attendance, retention, and learning conditions in the treated schools.

Endline (post-intervention): After the intervention period (for instance, at the end of the school year), conduct an endline survey across both the treatment and control schools. At this point, all schools are measured on the same indicators as at baseline (such as retention rates, average attendance, and test scores in core subjects). The impact of the programme can then be estimated by comparing outcomes between the two groups – essentially measuring how much the treatment schools improved relative to the control schools over the period.

Post-evaluation phase: After the endline data collection, the control schools would receive the intervention as well (this could be immediate or in a subsequent phase), ensuring that all selected schools eventually benefit. This "delayed treatment" approach upholds ethical considerations by not permanently denying the programme to any school, and it aligns with the programme's intent to expand coverage. It also helps maintain government and community support for the evaluation, as stakeholders know the control group's exclusion is only temporary and for learning purposes.

By following these steps, the RCT treats the phased introduction of the programme as a deliberate experiment. The assumption here is that there is buy-in from the Ministry of Education and implementing partners to use random selection for the rollout in those remaining schools. It requires that no school or community is prioritised for political or other reasons outside of the random assignment – in other words, the selection must be fair and unpredictable. If certain schools have been pre-selected due to urgent needs or other criteria, introducing randomness may involve stratification (for example, grouping schools by region or school type, then randomly picking some from each group) to ensure balance and acceptability. We assume also that the programme can be delivered independently to the chosen treatment schools without "leaking" into control schools. This means, for instance, that teachers from control schools are not inadvertently trained or transferred into treatment schools, and families in control communities do not receive the new cash grants. In practice, geographic separation or administrative arrangements may be used to minimize spillover – for example, treatment and control schools could be in different districts to reduce the chance that non-participating schools benefit indirectly from WASH facilities or other components nearby. With these conditions met, the RCT design offers high internal validity: any differences in outcomes between the two groups at endline can confidently be attributed to the programme, since randomization ensures no systematic bias in who received the intervention.

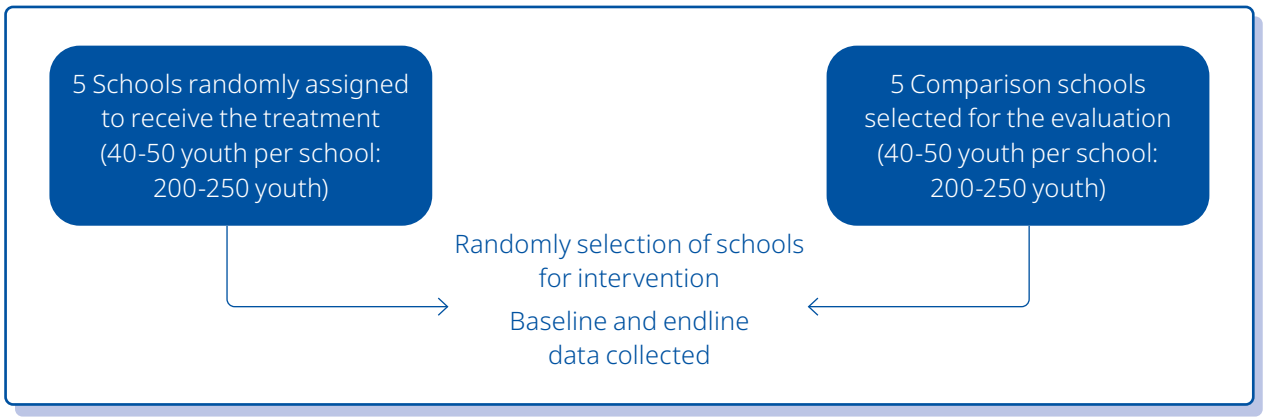
Executing this RCT will require close coordination with local education authorities. The evaluation team must ensure that the random assignment is properly implemented and adhered to: once the five treatment schools are selected, the plan must commit to rolling out the programme in those schools and not in the control schools until the evaluation's planned endline. Any deviation (for example, if authorities decide mid-way to extend the cash grants or WASH upgrades to all schools due to external pressure) would compromise the design. Therefore, maintaining the integrity of the experiment is a key operational challenge. Regular monitoring will be needed to confirm that control schools remain uninfluenced by the programme during the study period. Logistically, the RCT is simpler than a stepped-wedge design – all treatment schools start at the same time – but it still demands robust monitoring and record-keeping. We assume that the implementing partners can deliver the teacher training and WASH construction on schedule in the treatment schools, and that there is sufficient capacity to distribute cash grants efficiently to the selected families. Any significant delays or implementation failures in the treatment group could dilute the measured impact, so those risks will be tracked as part of the process evaluation.

Sample Size and Power

A notable limitation of this RCT design is the small number of clusters (schools) available for randomization. With only 5 treatment schools and 5 control schools, the trial may face limited statistical power to detect modest effects. In cluster-randomized trials, the number of clusters (schools, in this case) is a critical driver of power – and common guidelines often recommend a few dozen clusters per arm for robust detection of smaller impacts. Here, due to the pilot nature of the programme (only five schools initially receiving WASH and the full package), we are constrained to 10 clusters in total.

To mitigate this limitation, the evaluation will increase the student sample size within each school as much as feasible. We plan to survey or collect outcome data from about 40–50 students per school (for example, all pupils in a particular grade cohort or a random sample of students across relevant grades). This yields roughly 400–500 students in total across the 10 schools, which helps improve precision in estimating outcomes like average attendance or test scores. By incorporating baseline measurements and using techniques such as analysis of covariance (controlling for baseline values), we can further improve the ability to detect differences by reducing unexplained variance. Despite these measures, the minimum detectable effect size is relatively larger than in a bigger trial – we estimate that the design would be powered to detect only fairly substantial impacts (for instance, on the order of a 15–20 percentage point increase in attendance or a similarly large improvement in test scores) with conventional statistical confidence. Smaller effects might not register as statistically significant given the cluster count and the expected intra-class correlation of education outcomes within schools.

Figure 2. Randomized Control Trial (RCT)



Secondary Design: Propensity Score Matching (PSM) with Difference in Differences (DiD)

This is the chosen secondary design for the evaluation. It is a quasi-experimental approach that aims to simulate the counterfactual by selecting non-participating schools/communities that resemble the participating ones, and then comparing changes in outcomes over time between the two groups. The design has two key steps:

Propensity Score Matching (PSM) to Select a Comparison Group: The evaluation team will use propensity score matching to pair each "treatment" unit (schools or perhaps school catchment areas that received the programme) with one or more "control" units (schools that did not receive the programme) that have similar observed characteristics. The propensity score in this context is the probability of a school being selected for the programme, predicted from baseline characteristics. Likely matching variables include: school size (enrolment numbers), school type (primary/secondary, public/community), location characteristics (rural/urban, governorate or district), historical dropout rates, percentage of refugee students (if known), and community socio-economic indicators (poverty rate, etc.). The evaluation team will draw on Ministry of Education data and possibly UNHCR/UNICEF data to obtain these baseline covariates. The idea is to account for the factors that influenced where the programme was implemented – for example, if the Ministry/UNICEF deliberately chose schools with higher dropout problems, that needs to be accounted for. By matching on these factors, we create a comparison group of schools that looks statistically similar to the intervention schools at the outset, except for not having the programme.

The matching will likely be done at the school level (since the interventions are delivered at school/community level). We anticipate having a pool of potential comparison schools either in the same governorates (non-targeted schools in Cairo, Giza, Qalubiya, Alexandria, Damietta, Aswan) or in similar governorates that were not included (ensuring they have similar demographics). It's possible the programme targeted all schools meeting certain criteria; if so, we might match to schools just below the cutoff of selection. Each intervention school could be matched to, say, one or two control schools with the closest propensity scores. The evaluation team will check the balance after matching – i.e., ensure that the matched control schools have baseline characteristics (enrolment, pre-program dropout, etc.) not significantly different from the intervention group. A well-matched control group provides the foundation for a credible difference-in-differences analysis.

Difference-in-Differences (DiD) Analysis: With a matched control group in hand, we then compare the change in outcomes from baseline (pre-intervention) to endline (post-intervention) between the treatment and control groups. The difference-in-differences estimator will effectively subtract out any common trends affecting both groups, as well as any baseline level differences (since we focus on changes). Concretely, if retention improved by X% in the intervention schools and by Y% in the comparison schools over the same period, the DiD impact estimate would be (X – Y)%, attributing that difference to the programme. This approach controls for time-varying factors like national education trends, seasonal effects, or policies that affected all schools. It also handles any initial differences between schools that are fixed over time (e.g., if intervention schools were somewhat larger on average, that constant difference won't bias the change measure).

The validity of DiD rests on the assumption of parallel trends – that in the absence of the intervention, the outcome trajectory for the treatment group would have followed the same path as that of the comparison group. We can investigate this by leveraging historical data: for example, using EMIS records we can check if enrolment or dropout rates in the years prior to the programme were moving similarly in the selected intervention vs. control schools. If the trends diverged before the programme, we may adjust the model or refine matching to improve parallelism. Selection bias remains a potential concern if unobservable differences exist between intervention and comparison schools. To mitigate this, robustness checks—such as placebo tests and sensitivity analyses with alternative matching approaches—will be conducted.

The PSM-DiD design is feasible as it requires no programme disruption and utilizes existing data. Key requirements include: (a) sufficient comparison schools not targeted by the programme, (b)

reliable baseline data for intervention and comparison schools, and (c) ensuring no other major concurrent interventions affect comparison schools.

Sample Size and Power

Given the relatively small number of participating schools (five treated schools), careful attention must be given to sample size planning to maintain statistical robustness. Recognizing potential limitations arising from the quasi-experimental design and propensity score matching (PSM), the evaluation strategy incorporates deliberate oversampling in the comparison group to ensure sufficient statistical power and methodological rigor.

Within each of the five treatment schools and the matched comparison schools, the evaluation team will sample approximately 40 to 50 students. This translates into an anticipated treatment group of approximately 200 to 250 students (5 schools × 40–50 students each), and a comparison group comprising roughly 200 to 250 students (5 comparison schools × 40–50 students each). This overall approach provides a total sample size ranging from approximately 400 to 500 students, sufficient for detecting moderate-to-large program effects.

Key assumptions underlying these estimates include:

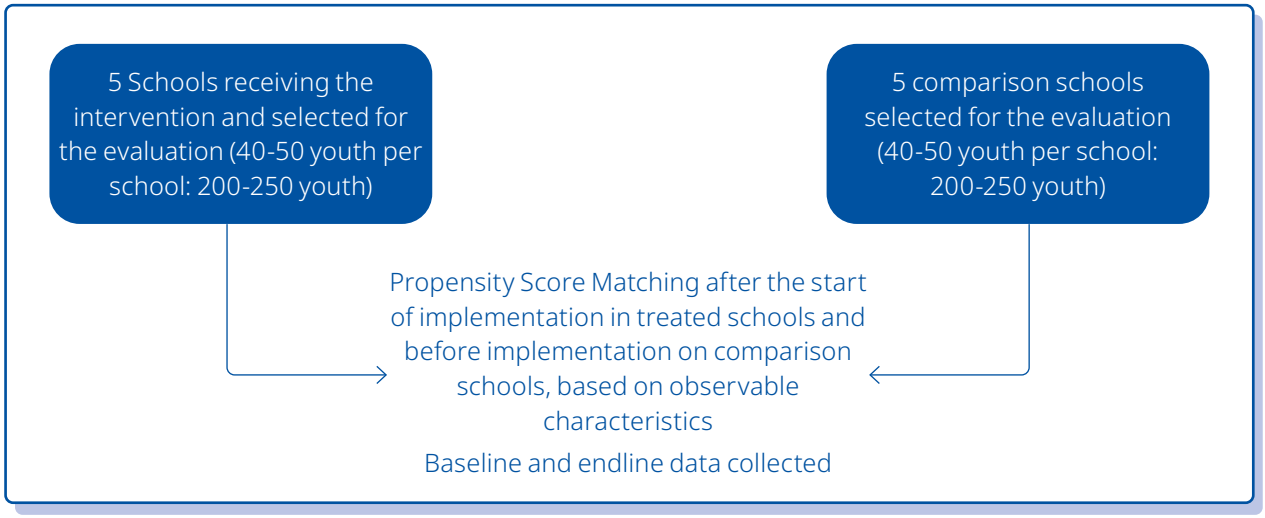
- **Significance level (α):** 0.05
- **Power (1–β):** Approximately 0.80, though actual achieved power may be lower given the small number of clusters.
- **Intra-class correlation (ICC):** Estimated between 0.10 and 0.20, in line with previous education-focused studies in comparable settings.
- **Cluster size (average students per school):** Approximately 40–50.

Under these conditions, it is anticipated that the Minimum Detectable Effect Size (MDES) may be somewhat larger—likely between 0.35 to 0.50 standard deviations for continuous outcomes such as test scores or retention rates—given the relatively limited number of clusters. While smaller effects might be challenging to identify reliably, moderate-to-large program impacts should still be detectable.

The evaluation will also aim to examine differential impacts among specific subgroups such as Syrian refugee students and girls. Given the smaller sample size, subgroup analyses will be primarily exploratory, and the evaluation team may adopt targeted oversampling of critical subgroups within schools, if feasible, to improve the robustness of these analyses.

Sensitivity analyses will be carried out to assess how variations in ICC, attrition rates, and subgroup sizes could influence statistical power and the precision of effect estimates. These analyses will inform interpretation and strengthen confidence in the evaluation's findings, given the inherent limitations of working with a small number of intervention schools.

Figure 3. Propensity Score Matching with Difference in Differences



Alternative Design 1: Retrospective Matching

If the Difference-in-Differences (DiD) approach cannot be fully implemented due to limited or poor-quality baseline data—particularly missing or incomplete baseline data for comparison schools—the evaluation team will adopt a retrospective matched evaluation design using endline-only data. Under this scenario, Propensity Score Matching (PSM) would be employed at the evaluation's endline stage to identify comparison schools similar to the five intervention schools, based on observable endline characteristics. In the absence of comprehensive baseline measures, this approach heavily relies on statistical adjustments and careful matching at endline to estimate the programme's effects.

Given the limited scale of the evaluation—covering only five treatment schools—the evaluation team will match each intervention school to two or three closely comparable non-participating schools, resulting in approximately 10 to 15 matched comparison schools. Matching variables will include observable characteristics at endline, such as current enrolment numbers, dropout rates, proportion of refugee students, community poverty levels, and geographical and demographic indicators. This matching process aims to reconstruct a plausible counterfactual scenario retrospectively, compensating partially for the absence of direct pre-intervention data.

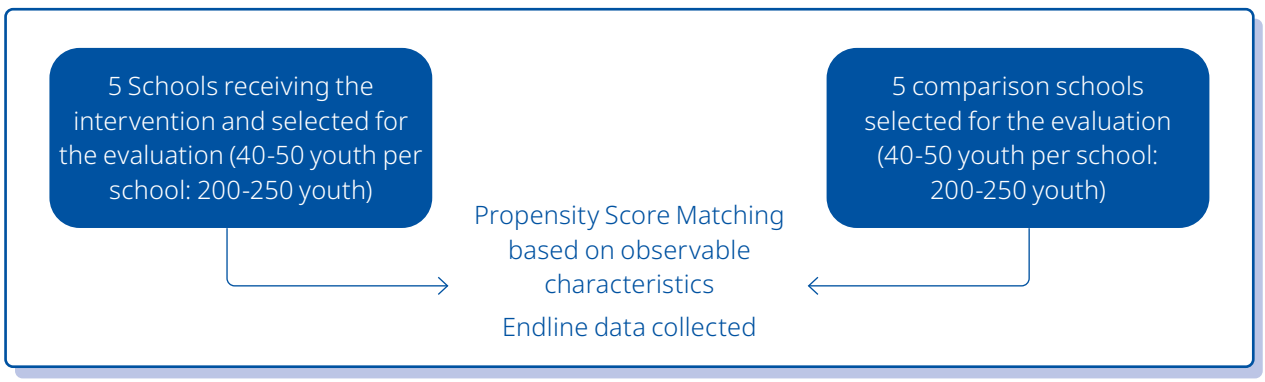
In a purely retrospective, cross-sectional analysis conducted solely at endline, the ability to make causal claims is inherently weaker compared to a prospective DiD approach. Observed differences between groups may partially reflect pre-existing, unobservable selection factors rather than direct programme impacts. To mitigate this limitation, multivariate regression analysis on the matched samples will be performed to statistically control for observable differences between groups. Additionally, retrospective recall questions—asking respondents such as teachers, principals, or students to report prior conditions—may be integrated into data collection tools to approximate baseline conditions, including enrolment patterns, attendance, dropout rates, and pre-intervention teaching practices. Wherever feasible, administrative records (e.g., Ministry of Education data or historical school registers) will also be utilized to reconstruct approximate baseline conditions.

Given the inherent constraints and potential measurement errors associated with recall data, results from this evaluation will be presented cautiously. For instance, findings might be summarized as follows: "At endline, treated schools exhibited X% higher student retention compared to matched comparison schools, suggesting a potential positive programme effect, although selection bias and recall limitations cannot be entirely ruled out."

Sample Size and Power Considerations

The evaluation will sample approximately 40 to 50 students per school yielding approximately 200 to 250 students in the five treated and control schools. Consequently, the total anticipated student sample size ranges from 400 to 500 students, significantly enhancing measurement precision and analytical robustness.

Figure 4. Retrospective Matching



Sensitivity analyses will further assess how variations in assumptions—such as ICC, subgroup distributions, and recall error—may influence the evaluation results. Although internal validity remains lower than prospective designs, this adjusted retrospective strategy provides credible, indicative evidence of programme impact within clearly articulated methodological limitations.

Summary of Research Designs

Table 1 below summarizes the key features of the three core design options, highlighting their requirements, strengths, and best-use scenarios:

Table 1. Evaluation Design Options

Design Option	Randomized Control Trial (RCT	PSM+DiD	Retrospective Matching
Randomization	Yes, cluster-level (school randomization)	No (uses observed comparison)	No (program already implemented)
Baseline Data	Yes	Yes – strongly preferred (for matching & DiD)	Not collected prior (reconstructed via recall)
Timing of Comparison Setup	Prospective – comparison groups defined before implementation	Quasi-prospective – comparison group identified after baseline but before they receive intervention (or from non-participants)	Retrospective – comparison group identified after implementation, at endline
Matching Method	Random assignment ensures balance	Statistical matching on propensity score (uses baseline covariates)	Post-hoc matching on recalled baseline and time-invariant traits
Primary Analysis	Simple OLS regressions, once baseline characteristics are balanced	Matched Difference-in-Differences (baseline vs endline changes in matched samples)	Endline group differences, using reconstructed baseline for context
Causal Inference Strength	**High** – strongest internal validity due to randomization; unbiased estimate of impact	**Moderate** – controls for observable differences; some risk of bias from unobservables remains	**Low–Moderate** – indicative results only; higher uncertainty and potential recall bias
Best Use Case	Randomization is feasible and ethical, and baseline data can be collected.	Program rollout was non-random, but baseline data exist (or can be collected) and some units haven't received the intervention yet.	Program already at scale or no baseline was done. Serves as a last-resort method to evaluate impacts after the fact

05

Key outcome indicators

To answer the evaluation questions and measure the programme's success, the evaluation team will track a set of key performance indicators aligned with the short-term and intermediate outcomes defined in the Theory of Change. These indicators will be disaggregated by important categories (gender, nationality/refugee status, school type, etc.) to examine equity and inclusion. Table 2 below summarizes the core indicators, along with their level (short-term vs intermediate outcome) and data sources.

Table 2. Key Indicators by Outcome Level and Data Sources

Outcome Level	Key Indicators (Illustrative)	Data Source & Frequency
Short-Term Outcomes	# of teachers completing professional development trainings[3]	School EMIS records (term-wise); Headcount surveys; Project records; Annual school survey.
	Net Enrolment Rate (by gender, nationality)[4,5]	
	Attendance Rate (% days attended)	
	Dropout Rate (annual, by grade)[6]	
	# of Out-of-school children (re) enrolled	
	Gender-specific attendance (girls' attendance %)	

3 Results framework indicator 1.1a) Number of teachers/ facilitators/ TVET trainers completing professional development trainings

4 Results framework indicator 1.2a) Number of new FDPs/HCs enrolled in pre-primary/primary/secondary education (formal and non-formal)

5 Results framework indicator PROSPECTS contribution to GCR indicator 2.2.1 – Proportion of refugee children enrolled in the national education system (primary and secondary)

6 Results framework indicator 1.2b) Number of FDPs/HCs retained in Primary/Secondary Education Program, Including Acceleration Education Programs and Early Childhood Education

Intermediate Outcomes	Test scores in reading and math (average)	Learning assessment at baseline & endline; Ministry exam data (annual); Classroom observation (baseline & endline); Facility checklist; Student/ teacher surveys; EMIS; Endline survey.
	% students passing year-end exams	
	Teacher practice score (observation rating)	
	Student-teacher ratio	
	Toilets per 100 students (by gender)	
	Student satisfaction with safety (survey)	
	Retention gap (refugee vs host)	
	Transition rate to secondary education	
Intermediate Impacts (for context)	Perceived inclusion (qualitative reports)	Focus groups, interviews (endline); Endline household survey (if done).
	Refugee-host difference in outcomes (quantitative proxy)	
	Parental confidence in children's future (survey/qual)	

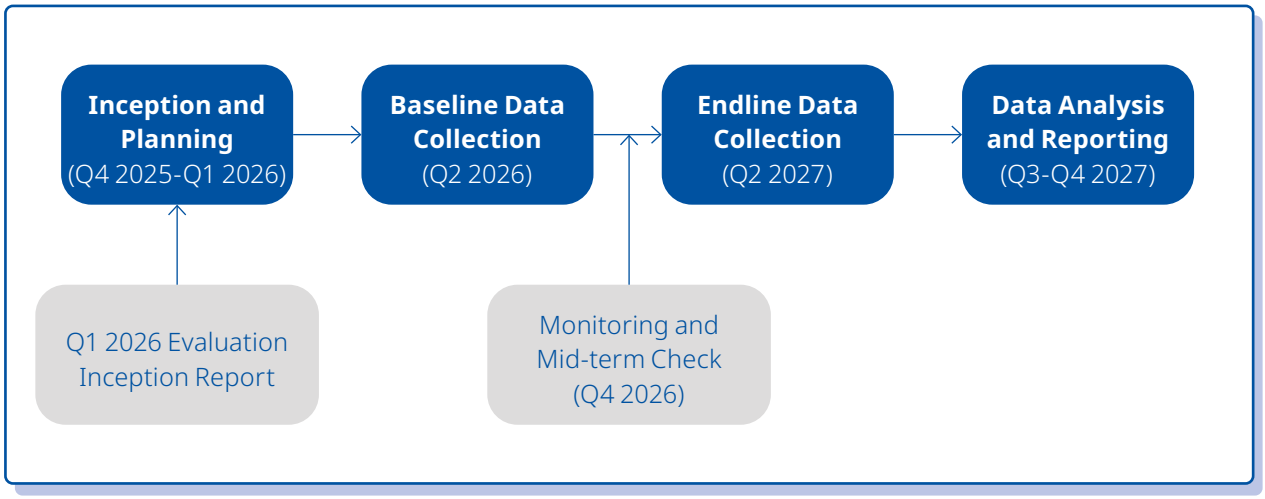
By tracking short-term indicators like attendance and dropout, we can detect early effects of the programme; by tracking intermediate indicators like learning outcomes and inclusion measures, we can assess whether the programme is on track to achieve its higher objectives. These indicators collectively provide a comprehensive picture of the programme's performance. The evaluation team will use them not only to evaluate impact at the end but also for ongoing monitoring (some data like attendance can be reviewed periodically to inform mid-course corrections if needed).

06

Timeline

The evaluation will be carried out over approximately 24-30 months, synchronized with the academic calendar and programme implementation schedule. Figure 5 below presents a tentative timeline with key phases and a more detailed version is presented below with milestones. The timeline assumes the programme is ongoing and will continue through the end of 2026, allowing us to observe at least one full school year (preferably two) of implementation effects.

Figure 5. Evaluation Timeline



Inception and Planning (Q4 2025 – Q1 2026): Establish the evaluation team and detailed plan. This includes recruiting any external research partners or consultants, engaging with government stakeholders, and refining the design and instruments. By the end of Q1 2026, the evaluation team will have completed the evaluation inception report, obtained necessary approvals (including ethical clearance), and finalized sampling strategies and matching criteria. Also, during this phase, the evaluation team will compile existing data (e.g., retrieve EMIS records from 2024 and 2025 for baseline info) and conduct a thorough review of programme documents to inform instruments

Baseline Data Collection (Q2 2026): A baseline survey and assessment will be conducted before the end of the 2025/2026 school year (or early 2026/2027, depending on logistics) to capture pre- or early-intervention conditions. Since the programme already started, baseline outcomes will rely significantly on pre-program data from the 2025 school year. Baseline collection will involve gathering school records from selected schools, administering learning assessments to a student sample (preferably at the end of the school year), and conducting surveys (household socio-economic data, past schooling, teacher knowledge assessments). Qualitative work, including key informant interviews (KIIs), will occur in Q2 2026. By the end of Q2 2026, we aim to fully document

pre-intervention conditions. If the intervention started fully in 2025, baseline data will partly be retrospective. For analysis, the 2025 school year serves as the "baseline year," and 2026 as the first "treatment year".

Monitoring and Mid-term Check (Q4 2026): While no formal midline survey is planned (due to our baseline-endline design), the evaluation team will conduct a mid-term review at the end of the 2026 school year (Q4 2026). We'll collect 2026 administrative data (enrollment/dropout) from intervention and comparison schools as interim data (first-year differences). Additionally, we'll conduct qualitative midline assessments, including focus group discussions (FGDs) with selected teachers and parents, to gather feedback on programme progress, implementation issues, and early outcomes. If resources allow, a brief teacher survey on knowledge gained from training may be included. This mid-term check will be formative, providing implementers with insights for potential adjustments and helping refine endline assessment instruments.

Endline Data Collection (Q2 2027): Major endline activities will occur around the end of the 2026/2027 academic year (April–June 2027, depending on exam schedules). By then, the programme will have been active for at least one full year (likely two for early starters in 2025), allowing outcomes to emerge. Endline data collection will mirror the baseline but measure post-intervention conditions:

- Administer endline learning assessments to the same cohort (or equivalent grades) to measure literacy and numeracy gains.
- Conduct household surveys (final education status, schooling expenditures, economic well-being), teacher surveys (practices, attitudes, knowledge tests), and principal surveys (school stats, management changes).
- Perform classroom observations and WASH facility inspections in a subset of schools.
- Retrieve final administrative data (enrollment, attendance, exam results for 2027) from schools or EMIS.
- Carry out qualitative fieldwork: FGDs with students (girls, boys, refugees), parents (mothers' and fathers' groups), and KIIs with teachers, principals, and programme staff to explore perceptions of change, inclusion, and community effects.

Fieldwork will last 6–8 weeks across regions (Cairo to Aswan) with simultaneous teams. Data collection will be completed by the end of Q2 2027.

Data Analysis and Reporting (Q3–Q4 2027): After endline data collection, the evaluation team will clean and analyze the data in Q3 2027. Quantitative analysis will compute difference-in-differences estimates, run statistical tests, and conduct subgroup analyses. Qualitative data (KIIs and FGDs transcripts) will be coded and analyzed for key themes. Initial draft findings will be available by early Q4 2027, followed by a preliminary results validation workshop with key stakeholders (e.g., Ministry, UNICEF, UNHCR, donors). A draft Evaluation Report will be prepared, incorporating stakeholder feedback and peer review (if conducted). The final Evaluation Design Report will be updated with actual findings and finalized by the end of 2027.

07

Data Collection Methods

The evaluation team will utilize a mixed-methods data collection strategy, combining quantitative surveys/assessments with qualitative fieldwork. This ensures we gather both the hard data needed for statistical analysis and the rich contextual information to interpret those numbers. All data collection instruments will be carefully designed to align with the key indicators and evaluation questions and will be pilot-tested in similar communities before full rollout. The methods are summarized below by type:

Quantitative Data Collection

School Administrative Data: The evaluation team will collect existing administrative records from the Ministry of Education's EMIS and school registers, including enrolment (by grade/gender), attendance (where available), dropout lists, and exam results. Data will cover two years pre-programme and each year during the programme (baseline: retrospective to pre-2025; midline: 2026; endline: 2027).

School Survey: In sampled schools (intervention and comparison), the evaluation team will administer a structured survey to head teachers to gather data on school characteristics (teacher numbers, facilities, programmes, costs) and challenges with attendance. At baseline, this aids matching and context; at endline, it captures changes (e.g., teacher training).

Household Survey: The evaluation team will survey parents/caregivers of a subset of students (20–30 households per school, around 600–900 total) using two-stage sampling (schools and households). Surveys will cover demographics, poverty indicators, education history, costs, attitudes, child labour, and for cash grant recipients, usage and satisfaction.

Student Assessment: To measure learning outcomes, the evaluation team will test cohorts from primary and lower secondary at baseline and endline. Tests (30–45 mins) will assess literacy and numeracy using standardized instruments aligned with the curriculum. A short student survey will capture background, school experience, and aspirations.

Teacher Survey and Assessment: Teachers (3–5 per school) will complete surveys at baseline and endline covering demographics, qualifications, training, and attitudes towards inclusive education. A knowledge test or vignettes on pedagogy will assess training impacts. Surveys will be administered during school visits (around 30 minutes).

Classroom Observations: Structured observations will assess teaching practices and classroom environment using a standardized tool (adapted from CLASS/UNICEF checklists). Observations (1–2 per school, full lesson duration) will be conducted blindly (where possible) at baseline and endline to supplement teacher survey findings.

WASH Facility Checklist: Enumerators will complete a WASH checklist during school visits, recording toilet facilities, cleanliness, water access, and maintenance needs. At endline, this will verify WASH component outputs and allow analysis of correlations with outcomes like girls' attendance.

Qualitative Data Collection

The qualitative component is integral for understanding the nuances of implementation and the experiences of beneficiaries. The evaluation team will employ several qualitative methods.

Key Informant Interviews (KIIs)

The evaluation team will conduct semi-structured interviews with approximately 20–30 stakeholders, mainly at endline:

- School Administrators: Principals/head teachers from intervention and comparison schools to gather perspectives on programme impacts and challenges.
- Education Officials: Ministry of Education and Ministry of Social Solidarity representatives, and UNICEF and UNHCR officers, to discuss policy alignment, system support, and sustainability.
- Programme Implementers: UNICEF staff to reflect on implementation fidelity, challenges, and perceived effectiveness.
- Community Leaders/Members: NGO partners, refugee leaders, and religious figures to explore shifts in community attitudes toward education.

Interviews (45–60 minutes) will use tailored guides to cover relevance, effectiveness, inclusion, and sustainability.

Focus Group Discussions (FGDs)

- Parents/Caregivers: Groups of 6–8 mothers and fathers (separately), including refugee-only groups where possible, to discuss education attitudes, programme impacts, and barriers.

Document Review

Finally, the evaluation will review relevant documents and monitoring data. This includes the programme's documents, training curricula and manuals used, progress reports, and any existing monitoring data (e.g., lists of recipients, school monitoring checklists, previous evaluation reports of related projects).

This desk review helps verify what was planned vs. delivered (e.g., how many teachers were actually trained, what criteria were used for cash selection), and situates our findings within existing evidence. It also aids in triangulation – if monitoring reports claim a certain outcome, the evaluation team will confirm that with our independent data collection.

08 Analysis Methods

The evaluation will employ rigorous analysis methods tailored to each design option, with a clear plan for quantitative and qualitative data analysis. The aim is to produce credible estimates of impact and insightful explanations, while addressing the assumptions and potential biases inherent in the chosen design. We outline the analysis approach for the primary design (PSM-DiD) and alternatives (SW-CRT, retrospective matching), followed by how the evaluation team will analyse qualitative data and integrate findings.

Quantitative Analysis

Propensity Score Matching with Difference-in-Differences

The evaluation team will use a Difference-in-Differences (DiD) approach implemented through regression analysis for each key outcome (e.g., attendance, dropout, test scores). The preferred DiD regression model is:

$$Y_{it} = \alpha + \beta_1 (Post_t) + \beta_2 (Treatment_i) + \beta_3 (Treatment_i \times Post_t) + \gamma X_{i0} + \varepsilon_{it}$$

Where:

- Y_{it} is the outcome for household or individual i at time t (e.g., attendance, dropout, test scores),
- $Treatment_i$ indicates assignment to treatment,
- $Post_t$ is a time dummy for endline,
- $Treatment_i \times Post_t$ captures the DiD treatment effect,
- X_{i0} are pre-intervention covariates, including baseline values of the outcome where available.

Given that treatment is at the school level and students within the school will benefit from the programme, the evaluation team will use cluster-robust standard errors clustered at the school level to account for intra-cluster correlation.

The validity of the DiD approach hinges on the parallel trends assumption – that in absence of the programme, the trend in outcomes would have been the same in both groups. The evaluation team will test this assumption by examining pre-program trends using historical data. For instance, we might compare 2019-2024 trends in enrolment or dropout between intervention and matched control schools; if they are parallel, it increases confidence. If the parallel trend assumption looks violated, the evaluation team will either refine the matching (to better align trends) or use analysis adjustments like including group-specific time trends (though that weakens identification).

A critical part of our analysis will be examining whether the programme's impact differs across subgroups such as gender, refugee status, and region. The evaluation team will implement this by extending the DiD regression with interaction terms involving subgroup indicators. For example, to see if impacts differ for refugees vs. host-community students, we include a triple interaction: Treatment × Post × Refugee. Similarly, for gender, we'd interact Treatment × Post × Female. These models will be run if sample sizes permit; we recognize that adding interactions reduces statistical power. If the sample of certain subgroups (e.g., refugees in secondary school) is small, we might instead do subgroup-specific DiD analyses (running the model on the subset of refugees and then on hosts separately) as a simpler approach.

Retrospective Matching: In case no baseline data is collected, a retrospective matched design will reconstruct pre-intervention conditions using recall-based indicators and secondary sources (e.g., administrative data). Treated and untreated schools will be matched post hoc on pre-treatment characteristics, and post-intervention outcomes will be compared using cross-sectional regressions or adjusted comparisons. Cross-sectional regression models with extensive covariate adjustment will be used.

Qualitative Analysis

The qualitative data collected from KIIs and FGDs will be analysed through a thematic content analysis approach. The evaluation team will begin by developing a coding framework that reflects both the evaluation questions and emergent themes from an initial review of transcripts. Likely code categories include barriers to education (sub-coded by economic, socio-cultural, and quality factors), perceived programme benefits (broken down by component such as teacher training, cash transfers, and WASH improvements), implementation challenges, changes in teaching practices, student attitudes, gender dynamics, refugee-host relations, and suggestions or unintended effects.

Once refined, the coding will be systematically applied to all transcripts using qualitative analysis software such as NVivo. Analysis will focus on identifying patterns within each theme. The team will examine whether perspectives differ across groups, such as between teachers and parents, or between regions like Cairo and Aswan. Special attention will be paid to uncovering the mechanisms of change; qualitative narratives will help explain how observed quantitative improvements, such as increased attendance, actually occurred in practice.

Direct quotes will be included in the final reporting to give authentic voice to the findings, carefully anonymized and selected to illustrate key points. A transparent audit trail documenting the coding and theme development process will be maintained to enhance the credibility and rigor of the qualitative analysis.

Triangulation of Quantitative and Qualitative Findings

Triangulation will be a central pillar of the analysis, combining quantitative and qualitative findings to enhance the validity and richness of the evaluation. After conducting separate analyses, the team will systematically compare results across data types, mapping areas of convergence and divergence for each evaluation question. For example, if the DiD analysis shows increased enrolment, the evaluation team will check whether qualitative interviews with school staff and parents corroborate this trend and attribute it to programme activities.

Where qualitative insights align with quantitative trends, they will strengthen causal interpretations; where discrepancies emerge, such as external factors influencing outcomes, the evaluation team will adjust conclusions accordingly. Triangulation will also help identify outcomes not directly measured quantitatively, with qualitative findings cross-validated against available quantitative proxies where possible. Triangulation procedures and outcomes will be fully documented in the methodology section to ensure transparency and rigor.

09

Estimated Resources for Data Collection

This section will be further refined in collaboration with UNICEF's Egypt Country Office, as it depends on the feasibility of each proposed quantitative evaluation design and the proposing data collection methods. At this stage, we lack sufficient information on the number of schools and children benefiting from the Learning in a Protective Environment for Increased Retention programme.

10

Feasibility and Limitations

Designing a rigorous evaluation in a real-world program context inevitably comes with feasibility considerations and limitations. We address here the practical feasibility of our proposed evaluation design in Egypt's context and discuss the key limitations, along with how we plan to mitigate them.

Feasibility Considerations

A major feasibility factor is that the programme has already started across all target areas. This means we cannot influence the implementation schedule for evaluation purposes, but it also means the intervention is in place to be observed. The evaluation will be embedded in an ongoing program, requiring close coordination with implementers to avoid disrupting activities while collecting data. The timeline we proposed aligns with school schedules, which is feasible as long as necessary permissions (from Ministry and perhaps each school or district) are obtained.

All target governorates (Greater Cairo, Alexandria, Damietta, Aswan) are accessible by road or air. We do not foresee security issues impeding fieldwork. However, given the dispersion, simultaneous teams are needed. With regards to data, the Ministry of Education's EMIS data is crucial; the evaluation team will need formal access. Feasibility is improved by the fact that UNICEF is a trusted partner to MoE. We must ensure data sharing agreements are in place, and data privacy is respected.

One potential challenge is finding a suitable comparison group of schools/communities. Since the intervention targeted certain governorates and presumably the most vulnerable communities within them, one might worry that all similar communities are already covered. However, in practice, even within the same governorates, not every school got the interventions, and there are similar schools that didn't. The feasibility of matching depends on the richness of data to characterize them. We plan to use both EMIS and possibly census or survey data to match on community characteristics. If we find that it's hard to get a perfect match (limitation), we might opt for a comparison group that is "close enough" and then rely on DiD to adjust, acknowledging some bias.

Potential Limitations and Mitigation

A key limitation of the evaluation is the non-randomized design, which introduces potential selection bias. While we use propensity score matching and a Difference-in-Differences (DiD) framework to control for observable differences, unmeasured confounders may still affect outcomes. For instance, differences in local leadership or baseline trends could influence results regardless of the programme.

Mitigation: We mitigate this through careful matching, pre-trend analysis, and cautious interpretation—reporting associations rather than definitive causal claims. Qualitative data further strengthens our findings by helping confirm causal pathways or uncovering alternative explanations behind observed changes.

Spillover and contamination are also possible, especially in cases where comparison schools are near intervention sites. These effects could dilute or distort impact estimates: shared district-level practices might improve both groups (understating impact), while student transfers from control to treatment schools could inflate programme effects.

Mitigation: We minimize this risk by selecting geographically distinct comparison units and including survey questions to detect exposure in control areas.

Programme implementation likely varied across sites, affecting both effectiveness and our ability to attribute specific outcomes to specific components. Since we did not design the evaluation to isolate each element (e.g., cash vs. WASH), we assess the intervention as a package.

Mitigation: Qualitative feedback and exploratory analysis (e.g., comparing outcomes by WASH completion status) will help interpret which components may have driven observed effects, but precise attribution remains limited.

Data quality may also constrain analysis. Attendance records may be incomplete, test participation may be biased by absenteeism, and retrospective recall can be unreliable.

Mitigation: To address this, we cross-check data sources (e.g., household surveys vs. school records), implement bounded recall techniques, and conduct make-up testing where feasible. Nonetheless, gaps or inconsistencies may persist, particularly in more remote or under-resourced settings.

In terms of external validity, findings are grounded in specific governorates and implementation contexts. These results may not generalize to other regions or countries without adaptation, especially given differing infrastructure, cultural norms, or policy environments.

Mitigation: The evaluation team will clearly situate findings within their local context while identifying elements that may be transferable, such as the value of cash support in vulnerable communities.

Finally, the short evaluation timeframe (1–2 years) limits our ability to assess long-term outcomes like resilience or self-reliance. Some benefits, like improved learning, may take time to manifest, while others may fade if support ends.

Mitigation: Our conclusions focus on early impacts and intermediate outcomes, with the recommendation that future follow-up be considered to assess sustainability and longer-term change.

11 Ethical Considerations

This evaluation will uphold the highest ethical standards, in line with UNICEF's Procedure for Ethical Standards in Research, Evaluation, Data Collection and Analysis and relevant national ethics guideline. Given that the subject matter involves children, refugees, and vulnerable families, the evaluation team will take special care at every step to protect participants' rights, safety, and dignity. Key ethical considerations include:

Informed Consent and Assent

Participation in data collection will be fully voluntary. The evaluation team will obtain informed consent from all adult participants (teachers, principals, parents, etc.) and parental consent for all minors involved in student assessments or interviews. Consent forms will outline the purpose of the study, what participation involves, any risks/benefits, and emphasize that declining will not affect their schooling or access to services. The evaluation team will present consent forms in Arabic (with a verbal explanation as needed for low-literacy parents) and give copies to participants. For refugee families, the evaluation team will clarify that data is for evaluation only and will not affect any aid or status determination (to alleviate any fear of saying "no"). Participants can withdraw at any time or skip any question they are uncomfortable with.

Confidentiality and Data Privacy

The evaluation team will ensure strict confidentiality of all personal data. Student and teacher surveys will use unique ID codes; no names will be reported in our datasets or publications. Any identifying information (like student names for panel tracking) will be stored securely and separately from outcome data, with access limited to the core evaluation team. Digital data on tablets will be encrypted and transferred to secure servers. The evaluation team will comply with data protection regulations (in line with UNICEF's data privacy guidelines) when handling personal data. In reports, results will be presented in aggregate form – e.g. school-level or group-level statistics – so that no individual or specific school is identifiable for sensitive outcomes. Audio recordings from qualitative interviews will be erased after transcription, and transcripts will be anonymized (replace real names with pseudonyms or codes).

Protection from Harm

The evaluation instruments will be designed to minimize any risk of harm or distress. Questions will be phrased sensitively, avoiding any triggering language especially for vulnerable youth (some may have trauma backgrounds). Enumerators and facilitators will be trained on child protection and gender-sensitive approaches. If during the course of data collection a respondent shows signs of distress or reveals a serious issue (e.g., abuse, self-harm thoughts), the team will have protocols to respond – such as pausing the interview, and making referrals to professional support services (UNICEF and partners have protection mechanisms in these communities). The evaluation will coordinate with the program's safeguarding officers to handle any such cases. Discussions in groups will be managed to ensure respectful listening and no stigmatization; for instance, girls will be in female-only FGDs to allow free expression, and any discussion of sensitive topics (e.g., gender-based violence, if it arises in context of life skills or safety) will be handled with confidentiality.

Cultural Sensitivity

The team will be trained in the cultural norms of the communities. For example, when scheduling interviews with mothers in conservative areas, using female staff and appropriate dress code; obtaining permissions through community leaders where customary; and being mindful of gender dynamics (like not asking young girls questions that might be inappropriate in local culture, and being careful when discussing topics like early marriage or child labour). We'll engage local informants or community facilitators to advise on any sensitive approaches. Questions about income or family issues will be phrased respectfully and only asked if necessary for the evaluation (and often such data can be gleaned from existing sources to avoid probing).

Child Protection Protocols

Enumerators and researchers will sign a **Child Safeguarding agreement**. If, during the course of data collection, a child discloses something indicating they are at risk (e.g. abuse, exploitation), the enumerator will not probe further (to avoid causing further trauma) but will report it to the evaluation field supervisor. We have a referral system in place: such cases will be referred to the appropriate child protection authorities or services (likely through UNICEF's protection team in Jordan, which can intervene or inform relevant national mechanisms) in line with national protocols. The evaluation team will inform participants of this limitation to confidentiality: i.e., if a child is in danger, we are obliged to seek help. This way, the evaluation will not turn a blind eye to serious issues encountered, fulfilling an ethical duty of care.

Feedback and Accountability to Participants

The evaluation is designed not just to extract data, but to ultimately benefit the communities by improving programs. The evaluation will practice reciprocity by feeding back results to the schools and participants in an accessible format (e.g., school meetings or brief summaries in local language) after the study, so they can see what was learned and how it will be used. This respects participants' contribution and ensures accountability.

The evaluation will protect participants' rights and welfare throughout the study by upholding these ethical standards. Ethical diligence is particularly paramount given that many participants are minors and refugees who have experienced vulnerabilities. The evaluation's conduct will reflect UNICEF's core principles of humanity, respect, and integrity.

May 2025