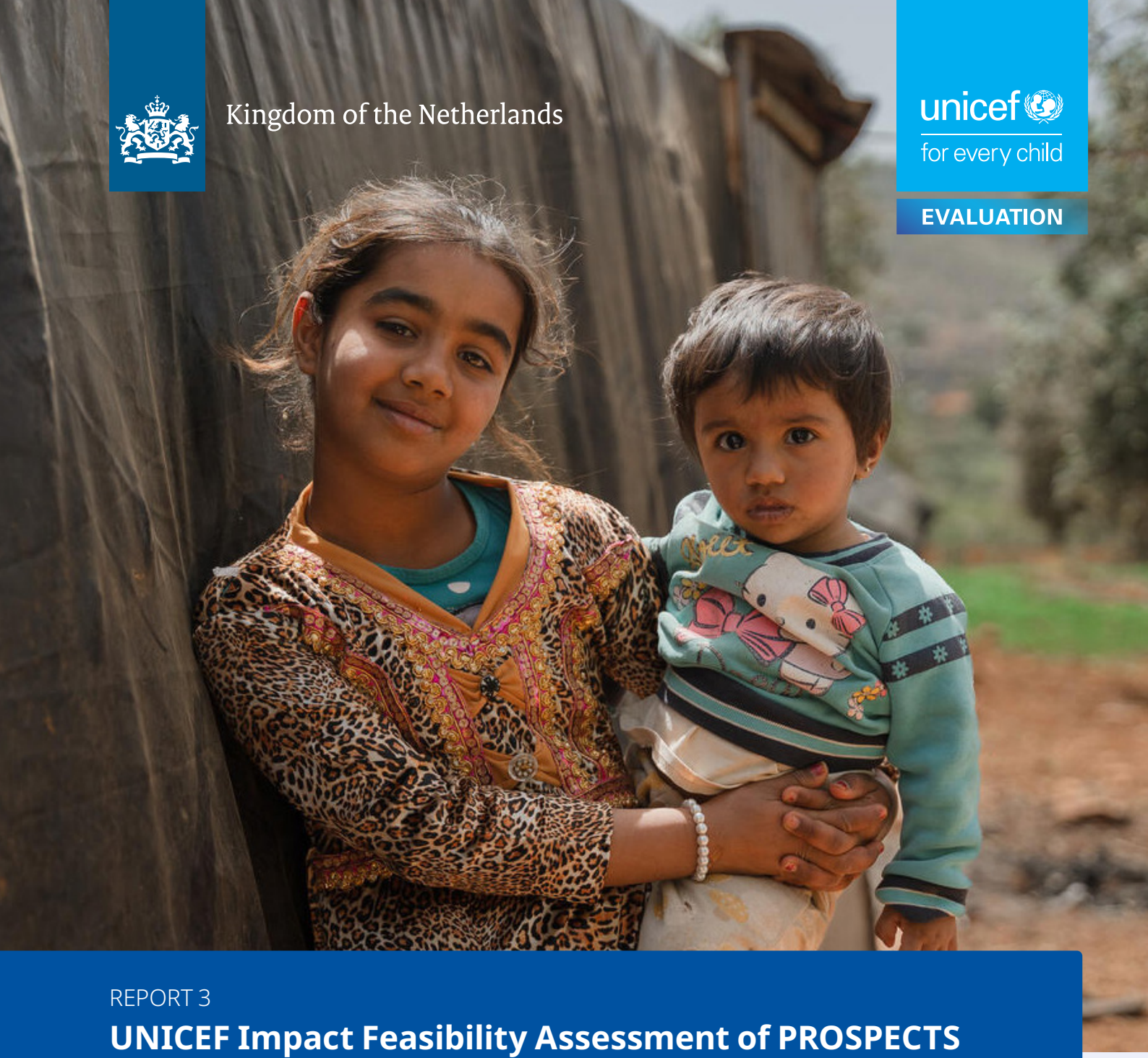
Kingdom of the Netherlands

unicef
for every child

EVALUATION



REPORT 3

UNICEF Impact Feasibility Assessment of PROSPECTS

Jordan 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 especially by Gabrielle Tremblay, 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: Jordan Case Study', UNICEF: New York, 2025.

Cover photo: © UNICEF/UNI578937/Saleh Elaiwa
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

Jordan remains a major host of Syrian refugees with over 731,000 registered refugees, leading to protracted strain on public services like education. Even before COVID-19, approximately 31.4% of Syrian refugee children (ages 6–15) were out of school.

The pandemic exacerbated learning gaps – a 2022 national diagnostic assessment found many students lacked basic literacy and numeracy skills for their grade. In addition to severe learning loss, the education system has faced weak school leadership and insufficient accountability for student retention. These challenges underline the need for new approaches to strengthen foundational learning and keep vulnerable children in school, especially girls and refugees who face higher dropout risks.

In response, the PROSPECTS Partnership (funded by the Netherlands and involving UNICEF, World Bank, ILO, UNHCR, and IFC) was established to integrate humanitarian and development efforts for displaced populations. PROSPECTS aims to enhance socio-economic inclusion of refugees and host communities by expanding access to quality education, employment, and protection, while strengthening resilience. In Jordan's education sector, PROSPECTS 2.0 supports "Developing Retention Mechanisms" – a multi-component intervention to address barriers to education for vulnerable Jordanian and Syrian children by improving teaching quality, reducing dropouts, and bolstering foundational skills.

A core part of the Retention Mechanisms is the **Schools Teaching for Learning Recovery (ST4LR) programme**, which consists of interlinked components targeting schools. This evaluation will focus on three key ST4LR components:[1]

Teacher of the Future (TOF): A teacher training course to build educators' capacity to deliver inclusive, gender-sensitive, and effective instruction using digital tools and innovative pedagogy. In Year 1 of PROSPECTS 2.0, TOF was rolled out in 50 vulnerable public schools (training ~4 teachers per school in Arabic and Math subjects). An additional 100 vulnerable schools are planned for TOF training in Year 2. By strengthening teachers' skills (e.g. universal learning design, engaging at-risk learners), TOF aims to improve classroom teaching and learning quality.

Reading Recovery Programme (RRP) and Maths Accelerated Programme (MAP): Targeted remedial education initiatives to bridge foundational learning gaps in literacy and numeracy for students at risk of dropping out. These programs provide supplemental outside of school hours, but inside schools, to help struggling learners reach grade-level competence. RRP and MAP will be implemented in Year 2 in the initial 50 TOF schools, focusing on key grades (e.g. early primary for reading, later primary for math) so that students can read and solve math problems at the expected level. This is expected to improve academic performance and confidence, reducing the dropout risk as students progress to higher grades. In subsequent years, RRP/MAP are slated to expand to the same 100 additional schools receiving TOF, eventually reaching all 150 target schools.

1 Other elements of the Retention Mechanisms, such as research on out-of-school children and school leadership support, complement ST4LR but are not the focus of this evaluation.

Box 1

Schools Teaching for Learning Recovery (ST4LR) Programme Selection

The Schools Teaching for Learning Recovery (ST4LR) 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);[1] and 5) the type of programming, which assessed interventions based on ToC Integration and partners integration.

- **Country Context:** Jordan has scored high in terms of country context, with high political will, operational facility, and national evaluation capacity and moderately prioritization based on knowledge gaps.
- **Intervention Context:** The Schools Teaching for Learning Recovery (ST4LR) programme was ranked as the top priority in Jordan (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.

The ST4LR programme is one of three interventions shortlisted as a priority for Stage 3 (alongside interventions in Ethiopia and Uganda) and one additional intervention as secondary priority (Egypt) for which impact evaluation plan is being developed.

The Schools Teaching for Learning Recovery (ST4LR) programme was selected as a promising 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).

By focusing on teacher quality and remedial learning, ST4LR directly addresses the barriers identified in Jordan's education context. The initiative builds on approaches already piloted with success – for example, the reading recovery model has shown positive outcomes in Syrian refugee camp schools. TOF leverages new digital teaching methods and gender-responsive pedagogy to engage learners who might otherwise fall behind. Overall, ST4LR is designed to improve the schooling experience in vulnerable schools – making education more accessible, equitable, and effective – so that children stay enrolled and succeed academically. This is particularly critical in Jordan's context of refugee inclusion fatigue and COVID-19 learning loss, where innovative retention mechanisms are needed to prevent a "lost generation" of learners.

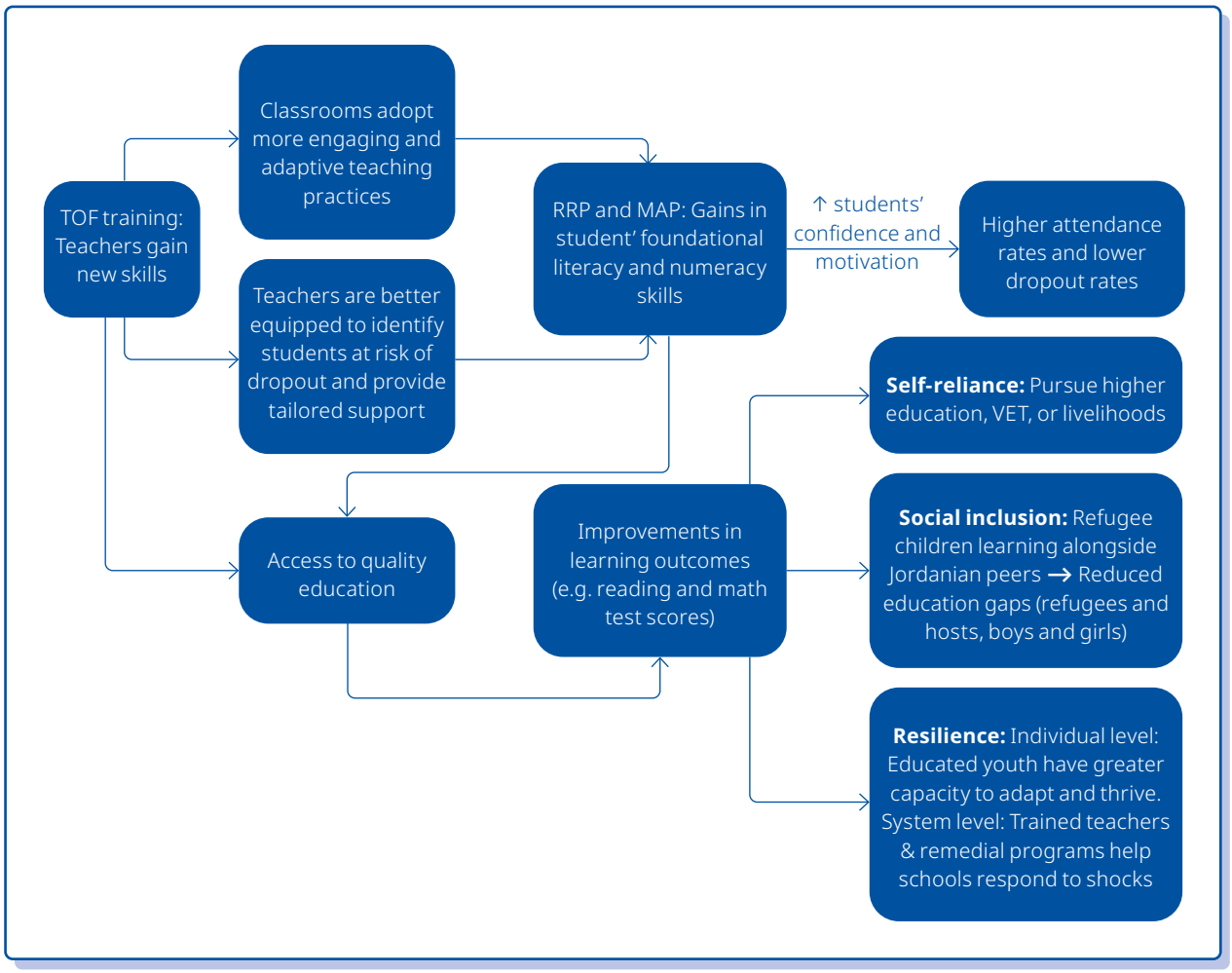
1 It is important to note that evidence gaps identified in the rapid review were limited to PROSPECTS countries and forcibly displaced populations. A detailed explanation of this decision can be found in the Section 2.1 "Selection Criteria" of the Stage 1 Impact Feasibility Assessment Report.

02

Impact Pathways and Key Outcomes of the Intervention

The ST4LR components operate along a clear results chain to achieve PROSPECTS' educational goals. Figure 1 below outlines the impact pathway, aligning short-term and intermediate outcomes with the broader PROSPECTS Theory of Change:

Figure 1. Impact Pathways



Through the TOF training, teachers gain new skills in inclusive, learner-centred instruction and digital pedagogy. Classrooms adopt more engaging and adaptive teaching practices (e.g. remedial techniques, gender-sensitive approaches) to support all learners. Teachers are better equipped to identify students at risk of dropout (e.g. those falling behind in reading) and provide tailored support.

Students gain foundational literacy and numeracy skills as a direct result of better teaching and targeted remediation. For instance, children in the RRP improve their reading fluency and comprehension, and those in MAP grasp critical math concepts appropriate for their grade. These academic improvements boost students' confidence and motivation, making them more likely to attend school regularly. In the short term, we expect higher attendance and lower dropout rates in the intervention schools, reflecting improved enrolment and retention. Parents, seeing their children progress, may also be more inclined to keep them in school.

By Year 2 and beyond, as ST4LR scales to more schools, a larger number of vulnerable children, including Syrian refugees and disadvantaged Jordanian students, are accessing education that is not only available but of higher quality. "Access to quality education" here means that more children are in school and benefiting from effective teaching and remedial support, rather than sitting in overcrowded classes with unmet learning needs. We anticipate measurable improvements in learning outcomes (e.g. reading and math test scores) across the target schools, narrowing the gap between refugee and host-community students. Schools become more inclusive environments, evidenced by better retention of groups historically prone to dropout (such as refugee children, girls, and children with learning difficulties). This directly supports PROSPECTS' education outcome of expanded quality education access for displaced and host community children.

Education is a foundational driver of longer-term socio-economic impacts. By keeping children in school and improving their basic skills, ST4LR contributes to future self-reliance – students who attain functional literacy/numeracy and complete basic education will be better positioned to pursue higher education, vocational training, or livelihoods. This underpins their ability to support themselves eventually. The intervention also fosters social inclusion: refugee children learning alongside Jordanian peers with improved outcomes helps integrate them into the national education system and community. Over time, reduced education gaps between refugees and hosts (and between boys and girls) indicate greater inclusion. Lastly, building resilience is an implicit goal: at the individual level (educated youth have greater capacity to adapt and thrive amid challenges) and at the system level (schools with trained teachers and remedial programs can more effectively respond to learning crises like COVID-19). These intermediate impacts align with PROSPECTS' higher-level vision of refugee and host communities that are more self-reliant, inclusive, and resilient. While such impacts fully materialise over many years, the evaluation will look for early signs – for example, improved transition rates to secondary school or non-formal education, and qualitative evidence of enhanced student aspirations and well-being – as proxies for these long-term goals.

The impact pathway assumes that training teachers will translate into changed classroom practices, that students will attend the remedial sessions, and that external factors (e.g. economic shocks or policy changes) will not severely counteract the gains. It also assumes the Ministry of Education (MoE) will support and sustain these programs. These assumptions will be monitored throughout implementation. If the pathway holds, ST4LR is expected to deliver tangible improvements in education outcomes that set the stage for broader socio-economic benefits in line with the PROSPECTS Theory of Change.

03

Evaluation Questions

The evaluation will focus on impact-focused questions to determine whether ST4LR achieves its intended outcomes and contributes to higher-level objectives. Key questions include:

How has the ST4LR programme affected student attendance, retention, and grade progression in target schools?

- Are dropout rates lower in intervention schools compared to non-intervention schools?
- Do students participating in remedial and teaching interventions remain in school longer?

What is the impact of the ST4LR programme on student learning outcomes in foundational literacy and numeracy?

- To what extent do RRP and MAP improve students' reading and math competencies?
- Do these academic improvements translate into increased student engagement or confidence?

How has the Teacher of the Future (TOF) training improved classroom teaching quality and inclusive practices?

- Are teachers better able to identify and support at-risk learners?
- Are learner-centred, gender-responsive, and digital pedagogies more widely used?

To what extent does the programme impact differ between Syrian refugee children and Jordanian host community students?

- Are refugee students benefiting equally in terms of learning gains and retention?
- What factors (e.g., displacement experience, language, support at home) explain any disparities?

To what extent does the programme impact differ between girls and boys in participating schools?

- Are girls achieving comparable or better outcomes than boys in learning and retention?
- How do gender norms, school climate, or teaching practices mediate these differences?

What school- and household-level factors moderate the effectiveness of ST4LR interventions?

- How do variations in school resources, teacher characteristics, or family support affect outcomes?
- What contextual conditions enable or constrain the programme's impact?

What early signs are that ST4LR contributes to long-term goals such as school transition, aspirations, or future educational participation?

- Are students more likely to transition to higher levels of education?
- Do students (especially girls and refugees) express stronger educational or vocational aspirations?

These questions will be answered through a mix of quantitative and qualitative data, emphasising identifying causal impacts and understanding how and for whom the program works. All questions align with the PROSPECTS theory of change and the impact pathways described in the previous section – focusing on whether ST4LR "moves the needle" on keeping children in quality education, and what that implies for broader development goals

04 Evaluation Design

Key Considerations for Clarification Before Finalizing Evaluation Design

This section is based on information currently available from PROSPECTS documents, particularly the Jordan PROSPECTS 2.0 MACP. Communication with the country office during the preparation of this report was limited; therefore, the evaluation design presented here relies primarily on a desk review. Consequently, important information gaps remain, particularly regarding target groups, the implementation plan, and the timeline. Addressing these gaps will require detailed discussions with the country team. The following, non-exhaustive, list of critical questions highlight key areas needing clarification before finalizing the actual evaluation design:

- **Intervention sequencing:** What is the specific sequence or relationship between the teacher training (TOF) and the remedial programs (RRP/MAP)? Are these components significantly overlapping, or is there a distinct transition between the two?
- **Intervention length:** What is the planned duration of each intervention component (teacher training and remedial programs)? Is it anticipated that interventions span multiple academic years, or are shorter intensive periods planned?
- **Target grades and groups:** Which specific grades or student cohorts are prioritized for each intervention component? Are there distinct target groups for teacher training versus remedial interventions? Are there specific considerations regarding the inclusion of IDPs versus host community populations in these target groups? What are the proportion of IDPs and host community populations benefiting from the programme?
- **Geographic spread:** What is the expected geographic distribution of the intervention? Is implementation concentrated in particular regions or broadly distributed to adequately represent diverse contexts?
- **Integration of transition-to-work or out-of-school programs:** Is it feasible to coordinate or co-locate the ST4LR intervention with existing or planned interventions targeting out-of-school youth or transition-to-work programs? If so, how might this integration be structured effectively?
- **Innovative or behavioural components:** Are there plans or opportunities to introduce additional innovative or behavioural strategies to enhance engagement, motivation, and effectiveness of the intervention? Could specific behavioural insights or innovations from similar contexts be adapted here?

Addressing these questions collaboratively between the evaluation team and country office will help ensure the evaluation design is appropriately tailored, practical, and capable of capturing meaningful impacts of the ST4LR intervention.

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 assess the ST4LR programme in Jordan rigorously. This design integrates quantitative impact evaluation with qualitative inquiry to capture both the magnitude of effects and the mechanisms behind observed changes.

A mixed-methods approach is particularly suited to the intervention's systems-oriented and multi-component structure, such as the one presented here. Quantitative methods will provide evidence on "what works" through statistical impact estimation, while qualitative methods—such as interviews, focus groups, and case studies—will illuminate "how and why it works" by examining contextual factors, implementation fidelity, and pathways to change.

This triangulated design improves internal validity and supports external applicability, making it well-suited for understanding outcomes across diverse displacement-affected populations in Jordan.

Quantitative Evaluation Designs

To rigorously assess impact, the evaluation will employ an experimental or quasi-experimental design that leverages the program's phased rollout. The ideal approach is a Stepped Wedge Cluster Randomised Trial (SW-CRT), wherein the 150 schools are introduced to the intervention in staggered phases by random assignment. This design takes advantage of the real-world scheduling of ST4LR to treat it as a built-in experiment.

If random phasing proves infeasible (e.g. if school selection was predetermined by MoE without randomisation), the evaluation will adopt quasi-experimental strategies – namely propensity score matching with difference-in-differences (PSM-DiD) or retrospective matching – to construct a credible comparison group. Each option balances internal validity, feasibility, and ethical considerations, ensuring the evaluation can adapt to operational realities while maintaining rigorous standards.

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 sections below describe each design option, including its methodology, requirements, and applicability. We also provide a summary comparison of designs. Throughout, the priority is to ensure the evaluation yields unbiased, policy-relevant estimates of ST4LR's impact on retention and learning.

Primary Design: Stepped Wedge Cluster Randomized Trial (SW-CRT)

Under the SW-CRT design, the rollout of ST4LR to the 150 schools is randomized in phases, so that all schools eventually receive the intervention but at different times. In practice, this would mean randomly assigning which schools are in the first wave (Year 1) versus the second wave (Year 2) of implementation. The phased rollout already planned (50 schools in Year 1, 100 in Year 2) can be used as the framework for the wedge – for example, if those 50 were selected randomly or can be treated as such. Each school (cluster) acts as its own control prior to receiving the program, and schools that have not yet received the program serve as controls for those that have, at any given time. Implementation: All 150 target schools would be included in the study. A baseline measurement is conducted before any school starts the program. Then:

- **Phase 1 (Year 1):** 50 schools (Group A) implement TOF (teacher training) while the remaining 100 schools (Group B) have not yet received ST4LR and thus act as a control group. We measure outcomes at the end of Year 1 (midline) in both groups.

- **Phase 2 (Year 2):** The program is extended to the remaining 100 schools (Group B now receives TOF in Year 2, while Group A continues into Year 2 with both TOF and the addition of RRP/MAP). By the end of Year 2, all 150 schools have had the intervention for at least some time (Group A for two years, Group B for one year). A midline measurement is conducted at that point.
- **Phase 3 (Year 3):** The RRP/MAP program is extended to the remaining 100 schools (Group B now receives TOF+ RRP/MAP in Year 3, while Group A continues into Year 3 with both TOF and RRP/MAP). By the end of Year 3, all 150 schools have had the intervention for at least some time (Group A for three years, Group B for two year). An endline measurement is conducted at that point.

This two-period stepped wedge ensures every school gets the intervention by the study's end, addressing ethical concerns about denying benefits. Impact is estimated by comparing outcomes between groups over time: for instance, after Year 1, Group A (treated) vs Group B (control) differences indicate the effect of one year of TOF; after Year 2, additional gains in Group A vs Group B can shed light on the effect of having RRP/MAP (since Group A had them in Year 2, Group B did not until Year 3). A combined analysis uses data from all periods to estimate the overall average treatment effect while controlling for time trends.

A SW-CRT has high internal validity due to randomised timing – any systematic differences between schools are balanced on expectation, and temporal effects (like nationwide improvements or shocks) can be accounted for since each wave acts as control at some point. This design leverages the phased scaling as an evaluation asset, enabling rigorous measurement of impacts across diverse schools and over time. It also increases statistical power by effectively using both between-group and within-school (before/after) comparisons. Importantly, all schools eventually benefit from the program, which helps stakeholder buy-in.

Implementing an SW-CRT requires coordination with MoE to allow random (or at least unbiased) assignment to early vs later rollout. If the initial 50 schools were selected purposively (e.g. highest-need), true randomisation may not have been used in Year 1. In that case, the evaluation could still approximate a stepped wedge by randomly assigning the sequence for remaining schools or by analytically adjusting for any baseline differences. Logistically, the design requires multiple rounds of data collection (baseline, midline, endline) and close tracking of when each school starts the intervention. We assume the phased approach is operationally manageable (50 schools in year 1, 100 in year 2) – this seems realistic given Year 1 has already been completed on a smaller scale. We also assume minimal contamination between clusters (e.g. teachers from Phase 1 schools are not transferring en masse to Phase 2 schools during the study, and program materials are not shared extensively before rollout).

The main challenges to this method are administrative and ethical. We must maintain the phased rollout schedule – any deviation (e.g. if MoE decides to accelerate delivery to all schools simultaneously due to political pressure) would undermine the design. Managing data collection in two waves requires resources and planning to avoid missing data or high attrition. Also, if the first 50 schools were not randomly chosen, we must be cautious in interpreting differences; however, the difference-in-differences analytic framework can help isolate the treatment effect assuming parallel trends (or by controlling for baseline levels). Lastly, we need to account for the layered nature of the intervention: since RRP/MAP comes after TOF in the same schools, isolating the effect of each component is complex. The evaluation will primarily estimate the overall effect of the ST4LR package as delivered in sequence, but the evaluation team will attempt exploratory analysis to see if additional gains occurred once RRP/MAP started (to infer their contribution).

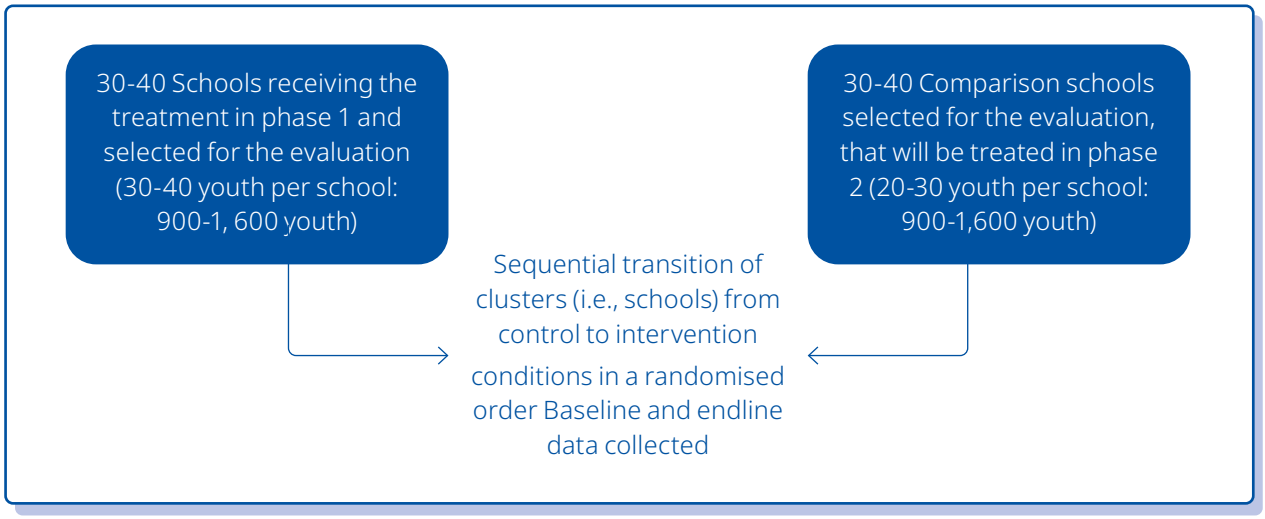
Sample Size and Power

With approximately 60–80 clusters, the sample size is robust. Each step will include around 30–40 schools per treatment arm, meeting the typical minimum of 30–40 clusters per group recommended for cluster trials. Even accounting for clustering effects, this design will still allow detection of relatively small program impacts. The evaluation team plans to assess approximately 30–40 students per school, generating roughly 900–1,600 student-level observations at each measurement round. This sample size maintains high statistical power, allowing detection of meaningful

effects such as a 10-percentage-point improvement in retention rates or around a 0.3 standard deviation increase in test scores with 80% power at a 5% significance level, assuming moderate intra-cluster correlation. Power is further enhanced by the longitudinal nature of the study, with baseline measurements serving as covariates and each school contributing repeated pre/post-intervention data. In summary, the SW-CRT with approximately 60–80 schools is sufficiently powered to reliably identify meaningful program effects on enrolment, attendance, and learning outcomes. The evaluation team should conduct formal sample size calculations using precise estimates of outcome variance and ICC from baseline or previous studies and adjust the student sample per school as necessary to ensure adequate statistical power.

Figure 2 provides an illustration of the Stepped Wedge Cluster Randomized Trial (SW-CRT) sampling design and its phases.

Figure 2. Stepped Wedge Cluster Randomized Trial (SW-CRT)



Cluster 1: TOF				
Cluster 1: TOF+RRP/MAP				
Cluster 2: TOF Treated to be				
Cluster 2: TOF+RRP/MAP Treated to be				
	Pre-trial	**Phase 1**	**Phase 2**	**Phase 3**

Control Condition
Experimental Condition

Alternative Design 1: Propensity Score Matching (PSM) with Difference in Differences (DiD)

In the absence of a randomised rollout, such as when intervention schools were purposively selected based on vulnerability criteria, a quasi-experimental design combining Propensity Score Matching (PSM) with Difference-in-Differences (DiD) provides a credible method for estimating program impact.

This approach is particularly appropriate when interventions have commenced in selected areas and baseline data are available for treated and untreated schools. For instance, if approximately 30–40 schools were chosen for Year 1 implementation due to specific vulnerability indicators (e.g., high dropout rates), a comparison group can be identified by selecting schools with similar baseline characteristics that did not initially receive the ST4LR intervention. These may include schools scheduled for later phases or other similarly vulnerable schools outside the immediate intervention group.

The approach involves three key steps:

- **Baseline Data Collection and Propensity Score Estimation:** Baseline surveys and administrative data collection will occur in intervention schools and a pool of potential comparison schools. A propensity score—representing the probability of receiving the intervention—will be estimated based on key characteristics (e.g., enrolment rates, dropout rates, school demographics, and regional attributes). Each treated school is then matched with one or more control schools possessing similar propensity scores, ensuring comparability on observable characteristics.
- **Difference-in-Differences Estimation:** Post-matching, DiD analysis will be employed to measure changes in key outcomes (attendance, enrolment, learning scores) from baseline to endline between matched treated and control schools. This method effectively controls for pre-existing differences and external trends affecting both groups equally, assuming parallel trends in the absence of intervention.
- **Timing and Cohort Design:** Data collection must be timed carefully to ensure the availability of pre- and post-intervention data points for both treated and comparison schools. Ideally, comparison schools would include those scheduled for future intervention phases, measured at baseline prior to receiving the intervention, or external matched schools that will not receive intervention within the evaluation period.

This PSM-DiD approach allows the evaluation to simulate counterfactual conditions without randomisation, leveraging existing variations in rollout timing and rich baseline data to estimate causal effects.

However, two key limitations must be addressed:

- **Unobserved Confounding:** If unmeasured factors influence treatment assignment and outcomes, estimates may be biased. To mitigate this, the propensity model will incorporate a broad set of covariates (e.g., poverty indicators, teacher-pupil ratios, baseline scores), and sensitivity analyses will be conducted to assess robustness.
- **Loss of Sample Due to Matching:** Some treated units may lack suitable matches, reducing the effective sample size. To account for this, the evaluation will sample more comparison schools than strictly needed (e.g., 75–100 for matching to 50 treated schools), excluding poor matches and ensuring common support.

When implemented rigorously, this design can yield credible impact estimates and enable meaningful insights into program effectiveness, particularly in contexts where purposive targeting precludes experimental designs.

Sample Size and Power

Considering resource constraints, the evaluation will strategically target a smaller yet sufficient number of schools, approximately 30–40 treated schools matched with 50–60 comparison schools at baseline. Within each school, sampling around 30–40 students provides approximately 900–1,600 observations in the treatment group and 1,500–2,400 in the comparison group at each measurement point.

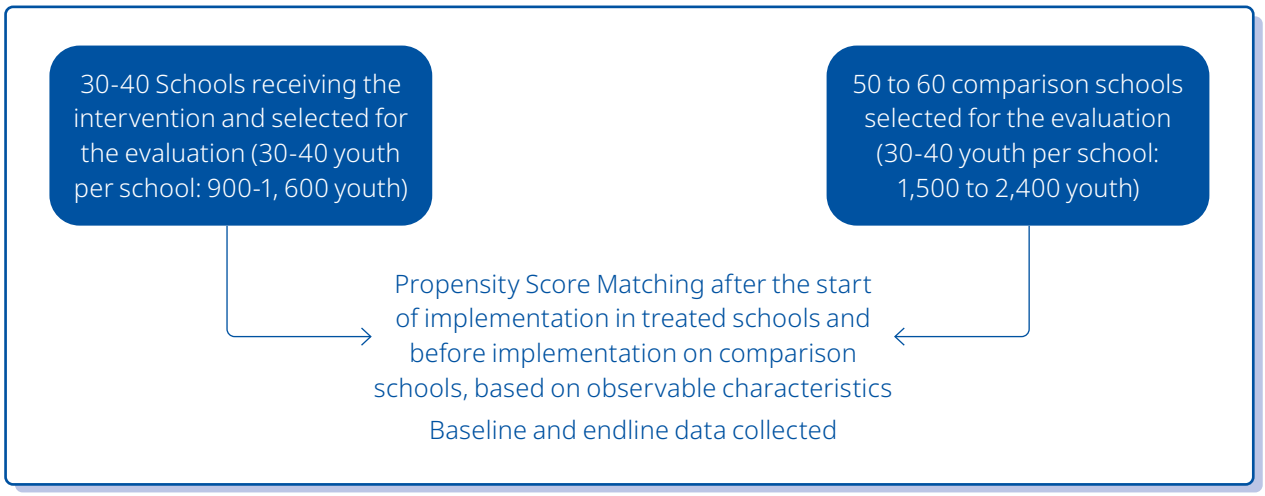
This sampling framework ensures robust statistical power (power of 80% and significance level of 0.05) to detect small-to moderate effects (approximately 0.25–0.30 standard deviations for continuous outcomes, such as literacy and numeracy scores). Key assumptions for these estimations include:

- **Significance level (α):** 0.05
- **Power (1–β):** 0.80
- **Intra-class correlation (ICC):** 0.10–0.20, consistent with similar education studies
- **Average cluster size:** Approximately 35 students per school
- **Matching ratio:** Approximately 1:1.5 to 1:2 (treated to control)

This adjusted design balances logistical feasibility and methodological rigor, ensuring sufficient power to detect meaningful intervention effects. Sensitivity analyses will further refine power calculations once empirical ICC and variance data become available from baseline data collection.

Figure 3 below provides an illustration of the Propensity Score Matching with Difference in Differences sampling design and data collection phases.

Figure 3. Propensity Score Matching with Difference in Differences



Alternative Design 2: Retrospective Matching

In cases where neither randomisation nor baseline data collection is feasible—such as when ST4LR is rolled out rapidly to all target schools before the evaluation begins—a retrospective matched design may be used as a last-resort option. This design constitutes an ex post impact evaluation that relies on post-intervention data and reconstructed baseline information.

This approach would apply if, by early 2025, all 150 schools have already received the ST4LR intervention and no prospective baseline was conducted, or if specific components (e.g. TOF training) were introduced universally before an evaluation could be established. A prospective control group or baseline cannot be created in such scenarios. The counterfactual must instead be reconstructed retrospectively. This design can also be applied selectively—for instance, if TOF were implemented everywhere. However, RRP and MAP are newly introduced, a retrospective evaluation could still be conducted for RRP/MAP by comparing the first 50 schools to others yet to implement these components.

The evaluation would involve a single endline survey, covering both:

- Treated schools (those that received ST4LR), and
- Comparison schools (those that did not receive the intervention or specific components at the time of the survey).

Since no baseline survey exists, pre-intervention conditions would be reconstructed using two data sources:

- **Recall data:** Respondents such as school principals, teachers, students, or parents may be asked to report key indicators retrospectively (e.g. past enrolment levels, regularity of attendance, or teaching practices prior to training).
- **Administrative or secondary data:** Existing records—such as school registers, Ministry of Education data on past enrolment, attendance, or exam results—would be compiled to serve as proxy baseline data.

These data would be used to perform post-hoc matching, conceptually similar to Propensity Score Matching. Treated and non-treated schools would be matched on recalled baseline characteristics and on time-invariant factors such as location, school type, and gender composition. The objective is to approximate pre-intervention comparability between the two groups.

Once matched, outcomes at endline are compared across treated and control schools. The underlying assumption is that the matched comparison group approximates what would have happened to the treated group without the intervention. Analysis may involve simple comparisons of group means, supplemented by regression adjustment for observable differences.

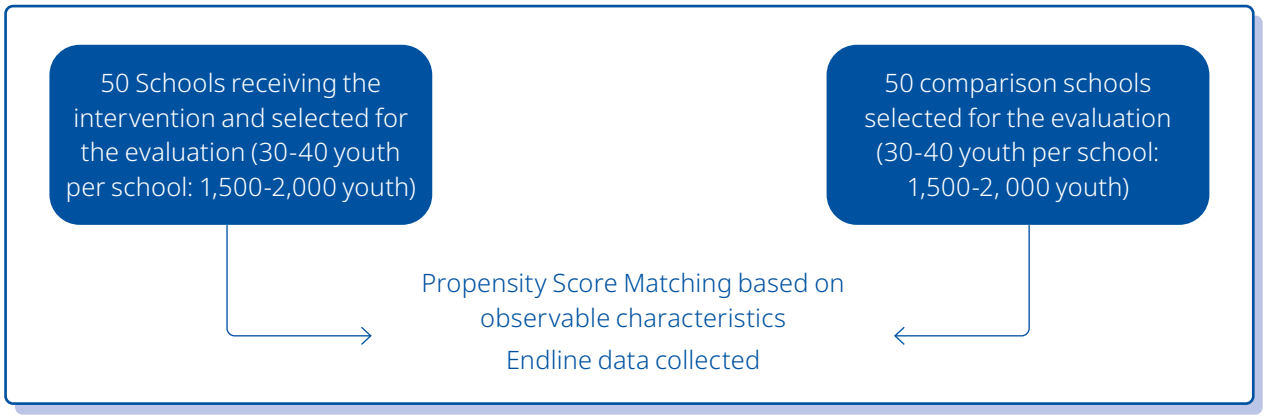
This design has lower internal validity than prospective approaches and depends on strong assumptions. In the absence of randomisation or direct baseline data, the validity of results hinges on the accuracy of recall and the quality of the matching process. Recall data are prone to memory errors and bias, especially where respondents are aware of the intervention and may consciously or unconsciously attribute improvements to it. Moreover, unobserved confounders may still bias results, as they cannot be controlled ex post.

As such, findings from a retrospective matched evaluation should be interpreted as indicative rather than conclusive evidence of impact.

Sample Size and Power

Recognizing the noisier data in this design, the evaluation team will inflate the sample size to compensate. We would survey as many schools and students as feasible – for instance, all 50 treated schools and perhaps an equally large number of comparison schools. If resources permit, targeting around 30-40 students in each school of the treated group and the control group. This oversampling helps average out recall errors and allows more robust statistical controls. The data collection would occur in one main round.

Figure 4. Retrospective Matching



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	SW-CRT	PSM+DiD	Retrospective Matching
Randomization	Yes – cluster-level (schools randomly phased into treatment)	No (uses observed comparison)	No (program already implemented)
Baseline Data	Yes – required before rollout	Yes – strongly preferred (for matching & DiD)	Not collected prior (reconstructed via recall)
Timing of Comparison Setup	Prospective – comparison groups defined before implementation (planned rollout)	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	Difference-in-Differences	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	Phased rollout 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 track progress along the PROSPECTS ToC, we have defined a set of key outcome indicators. These align with PROSPECTS' ToC and results framework – covering short-term outcomes (education access and retention), intermediate outcomes (education quality and inclusion), and intermediate impacts (self-reliance, inclusion, resilience proxies). The table below outlines the main indicators by result level:

Table 2. Key Indicators by Outcome Level and Data Sources

Outcome Level	Key Indicators	Data Source & Collection
Short-Term Outcomes	Gross/net enrolment rate (by gender, nationality)[2]	School administrative records
	Attendance rate (% of days attended)	Attendance registers
	Annual dropout rate (or retention to next grade)[3]	Baseline/endline survey data (household or student report)
	% of at-risk students identified for support	School reporting forms (re: at-risk students)

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

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

Outcome Level	Key Indicators	Data Source & Collection
Intermediate Outcomes	% students reading at grade level	Learning assessments administered at baseline & endline (e.g. Reading and math test)
	% students meeting numeracy proficiency (target grade)	Classroom observations and teacher survey (midline & endline)
	Average reading fluency (words per minute) and math test score	Student assessments data from RRP/ MAP (program monitoring)
	Teacher quality observation score (TOF practices)[4]	Baseline/endline survey (for perception of inclusion)
	Inclusion index (score gap between subgroups)	
Intermediate Impact	Transition rate to secondary education or formal training[5]	School records tracking graduates (or MoE stats on secondary entry)
	Social inclusion perception (% who feel school is inclusive)	Endline survey of students/parents (aspirations, perceived change in opportunities)
	Proxy self-reliance: % of older adolescents progressing to vocational training or employed (if applicable)	Community focus groups (on attitudes to schooling, changes in behaviour)
	Instances of dropout due to economic shock (qualitative/recall)	Possibly Ministry or UNHCR data on refugee education progression

Note: A detailed Indicator Matrix will be developed in the inception phase, aligning with PROSPECTS and MoE indicators. This table illustrates the main types of measures the evaluation team will use.

This mix of indicators will allow us to capture whether ST4LR is achieving its intended outcomes (like better retention and learning) and also to observe any movement towards the broader goals of inclusion and resilience. During analysis, these indicators will be used to quantify impact (for quantitative ones) and to provide narrative evidence (for qualitative ones).

Finally, the evaluation team will ensure continuous monitoring of output indicators (e.g. number of teachers trained, number of remedial sessions held) to verify that the intervention was delivered as planned – while not impact measures themselves, they are critical for interpreting the outcomes (for instance, if outcomes didn't improve, was it because only half the teachers completed the TOF training, etc.).

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

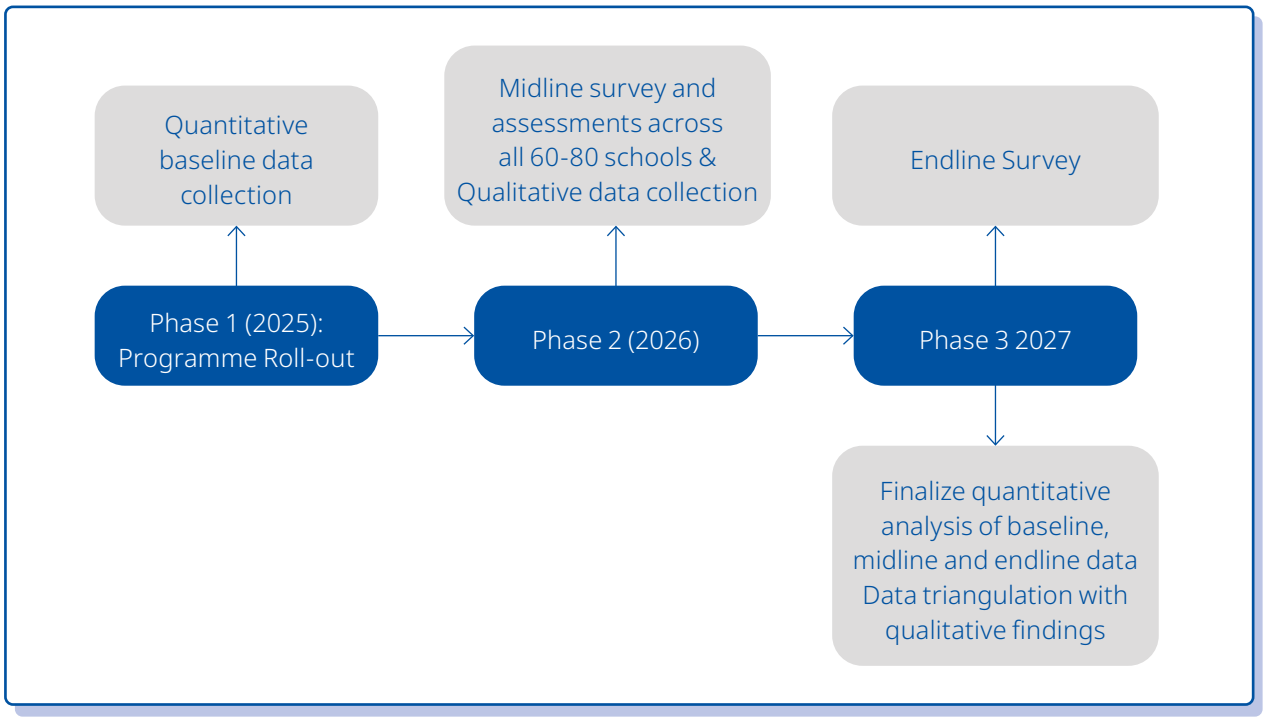
5 Results framework indicator 1a) Number of FDPs/HCs who completed their primary or secondary education program (formal and non-formal)

06

Timeline

The evaluation will span approximately three years (2025–2027), aligned with the phased implementation of ST4LR to ensure adequate time for observing program outcomes. Figure 5 provides an overview, with key activities and milestones detailed below.

Figure 5. Evaluation Timeline



Phase 1 (Year 1): At the beginning of Year 1 (early 2025), a baseline survey will be conducted across all sampled schools, capturing pre-intervention data on key indicators, including student literacy and numeracy levels, enrolment, attendance, and contextual information through qualitative interviews with principals, teachers, and parents. Following baseline data collection, approximately 30–40 schools (Group A) will implement TOF (teacher training), while another 30–40 schools (Group B) will remain as a control group during this phase.

Phase 2 (Year 2): At the beginning of Year 2 (early 2026), the midline survey and assessments will be conducted in all schools. At this point, Group A schools will have completed one year of TOF, and Group B schools will still be untreated, serving as a control. Immediately after midline data collection, Group B schools will begin TOF implementation, while Group A schools advance to implementing the Reading Recovery Program (RRP) and Math Accelerated Program (MAP).

Phase 3 (Year 3): At the beginning of Year 3 (early 2027), the endline survey and assessments will be conducted. By this time, Group A schools will have experienced two years of intervention (one year of TOF and one year of TOF+RRP/MAP), while Group B schools will have completed one year of TOF. Endline measurements will repeat the baseline and midline assessments, measuring student learning outcomes, enrolment, attendance, and teacher effectiveness. Additional qualitative data will also be collected to evaluate program perceptions and implementation fidelity.

This endline provides data to answer the primary impact questions – comparing Group A vs B to see the effect of the full package, and comparing within-group progress from baseline to endline. If a stepped wedge, all data (baseline, midline, endline) will be used in a combined analysis. The endline also serves as baseline for the RRP/MAP expansion in Group B if the program continues (though that will be outside our evaluation timeframe).

After data collection concludes, the evaluation team will conduct analyses comparing baseline, midline, and endline data to assess the overall program impact. Findings will be consolidated into a comprehensive evaluation report, supported by qualitative insights, and disseminated through stakeholder consultations and workshops planned for late 2027.

07

Data Collection Methods

A mixed-methods data collection approach will be used, combining quantitative surveys/assessments with qualitative fieldwork. This ensures we not only measure outcomes rigorously but also understand the context and implementation processes behind those numbers. Below we detail the data sources, units of analysis, and methods for both the quantitative and qualitative components.

Since the intervention is delivered at the school level (teachers and programs in each school), the school is the primary cluster for analysis. However, outcomes will often be measured at the student level (e.g. individual attendance, test scores) or teacher level (teaching practices), nested within schools. Within each selected school, the evaluation team will sample students for surveys and learning assessments. The evaluation team will focus on the grades that the programs serve. Typically, about 25–30 students per school will be sampled. All teachers who underwent the TOF training in each school (4 per school in year1 schools, similarly in year2) will be part of the evaluation for teacher-focused data.

Quantitative Methods: School Surveys and Administrative Data: The evaluation team gather school-level data such as enrolment figures, number of dropouts, attendance records, etc. This will involve both extracting data from MoE for each school (if accessible) and verifying/updating it via a short survey with school principals. For key indicators like enrolment and dropout, official records will be the primary source. The evaluation team will collect these at baseline, midline, and endline to track changes. In retrospective scenarios, we'll collect current and recalled previous-year numbers in one go.

Student Assessments: An important part of data collection is measuring learning outcomes. The evaluation team will administer standardized reading and math assessments to sampled students at baseline and endline. We may use an adapted EGRA (Early Grade Reading Assessment) tool to measure letter recognition, familiar word reading, and reading comprehension. For math, the evaluation team will include items on number sense, basic operations, and problem-solving. These tests will be designed in alignment with the curriculum and RRP/MAP content to sensitively capture improvements.

Student Survey: Alongside tests, students (especially older ones) may be given a short questionnaire covering demographics, attendance (cross-checking school records), study habits, how they feel about school (engagement, safety), and aspirations. This might be replaced by a parent survey.

Teacher Survey: A teacher survey will collect data on teacher background, attitudes, self-efficacy, and adoption of techniques (e.g. "Do you use group work regularly?", "How do you support students who lag behind?"). For TOF-trained teachers, we'll ask about their training experience and usage of skills. This data helps measure intermediate outcomes on teaching quality.

Qualitative Methods: Key Informant Interviews (KIIs) and Focus Group Discussions (FGDs): The evaluation team will interview a range of stakeholders. Ministry of Education officials to understand policy context, support and challenges in implementation, and perceived outcomes. School principals in treated schools to provide insight on how the program affected school operations, any changes in student behaviour or attendance, and how they tackled retention issues. Principals in non-program schools can discuss the challenges they face without such support (for comparison). Teachers: both those who participated in TOF and those who did not (for perspective). KIIs or FGDs with teachers will explore how their teaching has changed (or why not), the usefulness of training, how they implement remedial activities, and any anecdotes of student improvement.

08

Analysis Methods

The evaluation will employ rigorous analysis methods corresponding to the chosen design, aiming to isolate the causal impact of ST4LR on key outcomes while also extracting insights on implementation and context. The evaluation team will use a combination of quantitative statistical analysis and qualitative thematic analysis, followed by an integration of the two (mixed-methods triangulation). Below we outline the analysis plan for each component:

SW-CRT Design: For SW-CRT, the analysis will estimate the average treatment effect by comparing outcomes across randomized school clusters at different phases of the rollout. The core analytical model will follow a staggered Difference-in-Differences (DiD) specification, comparing outcome changes in treatment arms relative to control communities during the delayed phase.

The preferred 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,
- $Treatment_i$ indicates assignment to one of the three active treatment arms,
- $Post_t$ is a time dummy for endline (or later phases),
- $Treatment_i \times Post_t$ captures the DiD treatment effect,
- X_{i0} are pre-intervention covariates, including baseline values of the outcome where available.

Standard errors will be clustered at the level of random assignment (i.e., schools) to account for intra-cluster correlation. Where multiple follow-up points are available (e.g., baseline, midline, endline), fixed effects models will be used to leverage within-unit variation, improving statistical power and adjusting for unobserved time-invariant heterogeneity.

PSM with DiD: In case the random assignment proposed in the preferred method is not feasible but pre-intervention data is collected, a Propensity Score Matching with DiD framework will be applied. This method estimates the probability of treatment based on observable baseline characteristics, matches treatment units to control units with similar scores, and applies the DiD estimator to the matched pairs.

Balance tests will be conducted to ensure sufficient covariate alignment post-matching. The model specification will mirror the DiD format used in the primary design, but the sample will be restricted to matched pairs, and robustness checks will be used to assess the sensitivity of results to unobserved confounders.

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 data: KIIs and FGDs transcripts, observation notes will be analyzed using a thematic analysis approach. The evaluation team will develop a coding framework aligned to the evaluation questions and theory of change. Likely themes include: perceived changes in teaching practice, student engagement/motivation, barriers to attendance, parent attitudes, refugee inclusion experiences, unintended effects, implementation fidelity issues, suggestions for improvement, etc. Using software (e.g. NVivo or Atlas.ti), transcripts will be coded according to these themes. The evaluation team will then identify patterns and contrasts and pay attention to outlier perspectives. Qualitative data will be important to create case narrative summaries for KIIs and FGDs and illustrative quotes and stories can be used in the report to enrich interpretation of data.

The evaluation team will analyze data on how the program was delivered – any delays, material adequacy, etc. – to contextualize impact findings. If a certain region had delays, we might see less impact there (and qualitatively find complaints about late material delivery). The evaluation team will also specifically analyze qualitative data to answer the "how and why" questions: Why did enrolment improve? Why might learning not improve as much as hoped? These insights help in formulating recommendations.

Triangulation through Mixed Methods: Qualitative data will be deliberately collected in the same sites as quantitative to allow triangulation. For example, if a certain school's data shows huge improvement in attendance, we ensure to interview teachers/principal there to find out what drove it. Conversely, if data shows minimal change in some area, we investigate qualitatively what barriers existed. The evaluation team will take each evaluation question and compile evidence from quantitative and qualitative sides. For instance, for retention: Quantitative analysis indicates a +10% retention in treatment; Qualitative data points to "principals and students note dropout declined due to extra support". If both align, confidence in conclusion is high. If there's divergence (e.g. data shows modest test score gains, but teachers claim huge changes), we investigate further to explain the discrepancy (perhaps teachers referring to improvements in untested skills like confidence).

The evaluation team will use qualitative insights to explain any unexpected quantitative result. For example, if some schools did not improve learning despite training, maybe interviews reveal teacher absenteeism issues or differences in implementation.

The mixed-methods approach will allow us to present a nuanced narrative: not just whether the program worked, but how it worked and under what conditions. For example, we might find ST4LR greatly helped refugees in camp schools (quantitative) and qualitatively those schools had active community engagement, whereas in some urban schools the impact was smaller and teachers mentioned difficulties engaging parents – providing a possible explanation.

09

Estimated Resources for Data Collection

This section will be further refined in collaboration with UNICEF's Jordan Country Office, as it depends on the feasibility of each proposed quantitative evaluation design and the proposing data collection methods. At this stage, it is estimated that the extension of ST4LR will target approximately 100 new schools nationwide.

10

Feasibility and Limitations

Feasibility Considerations

Implementing this evaluation in Jordan is considered highly feasible, given the enabling environment and our adaptive design. Jordan's Ministry of Education has shown openness to innovation in education (e.g. collaboration with UNICEF on Learning Bridges and teacher training) and is likely supportive of evidence-based evaluation, especially under the PROSPECTS partnership. The evaluation team will work closely with MoE at central and field levels to secure buy-in – early consultations suggest interest in learning what works to improve refugee inclusion and retention.

Logistically, Jordan is a relatively compact country with good infrastructure, which eases field travel and school access. Security conditions are stable; unlike some contexts, we do not face active conflict or large security risks in the areas of intervention. This means longitudinal data collection can proceed without major safety constraints, beyond routine risk management.

We can leverage administrative data for sampling and cross-checking outcomes, which improves feasibility of measurement. Moreover, because UNICEF and partners have been working in these schools, there are established relationships with school principals and local education officers. This trust will help in organizing survey visits and getting truthful responses.

A key strength increasing feasibility is the adaptive design – we have built-in alternatives (PSM-DiD, retrospective matching) if the ideal scenario is not feasible. This means the evaluation won't be derailed if a baseline survey gets delayed in one region or if random assignment cannot be perfectly enforced. This flexibility ensures that even under suboptimal conditions, the evaluation team will collect usable data and generate findings. It de-risks the evaluation from a feasibility standpoint.

Potential Limitations and Mitigation

School Selection Bias:

The first 50 schools were chosen by MoE for Year 1, likely because they were among the most vulnerable. This means our intervention and control groups may differ at baseline (the 50 might have worse initial outcomes than the 100). This selection bias could confound results if not properly addressed.

Mitigation: The evaluation design explicitly accounts for this by collecting detailed baseline data and using matching/Difference-in-Differences to control for initial difference. The evaluation team will compare baseline indicators between these groups; if large imbalances remain, the evaluation team will incorporate those covariates in analysis or use weighting to re-balance.

Contamination and Information Spillover:

Since all schools are under the same education system, there is a risk that practices or benefits from ST4LR spill over to non-treated schools. For example, trained teachers in TOF schools might share techniques with colleagues in other schools (especially if they meet at regional meetings), or students in non-program schools might indirectly benefit from nationwide initiatives like Learning Bridges that run concurrently. If control schools improve due to such spillovers, it would dilute measured differences.

Mitigation: The evaluation team will attempt to geographically separate the phased rollout as much as possible – e.g., if feasible, Year 1 schools and Year 2 schools will be in different districts or far apart, to reduce direct contact. In analysis, the evaluation team will check for signs of contamination (for instance, if control schools also show adoption of certain practices, as per teacher survey, that could be contamination). If identified, we may treat it as a "partial treatment" effect and use instrumental variable analysis (e.g., use being officially in program as instrument for actual exposure to new practices) to still estimate impact. Also, since eventually all get the program by Year 3, contamination is mainly a concern for the interim period – we have kept that period relatively short (one year) to limit opportunities for diffusion.

Teacher mobility:

A trained teacher might transfer out of a school (common in Jordan's system annually). If a TOF-trained teacher leaves mid-year and is replaced by an untrained one, the impact in that school drops. Similarly, RRP/MAP might rely on certain enthusiastic teachers – if they are absent or over-burdened, the program effect suffers.

Mitigation: Adding the policy of training multiple teachers per school (4 in each) provides some redundancy – if one leaves, others remain. The evaluation team will document any such occurrences carefully during the study and factor that into interpreting results (for example, noting "X% of schools lost at least one trained teacher – a limitation on sustained impact").

Students Mobility:

Some students may leave the area (especially refugee families who might be resettled to a third country or move within Jordan). If many treatment-group students move and we can't follow them, we might undercount improvement (for example, a progressing student leaves and is replaced by a new struggling student – average score might look stagnated).

Mitigation: The evaluation team will try to follow up on why students left – distinguishing transfer to another school (which we might track via MoE if within Jordan) vs dropout vs leaving the country. The evaluation team will incorporate a conservative approach by considering those who left as dropouts for retention analysis (worst-case for program, which biases against finding impact – a safe side). The relatively short time frame (1-2 years) means mobility might not be extremely high; nonetheless, refugee resettlement is a possibility for a small number.

Economic or policy changes:

If in 2025 Jordan changes its policy on Syrian work permits or aid, families' socio-economic situations could shift, affecting schooling decisions (either positively or negatively). Likewise, a pandemic resurgence or new refugee influx would impact attendance system-wide.

Mitigation: In the analysis, the evaluation team will capture major contemporaneous events by including time fixed effects and possibly local economic indicators as controls (unemployment rate, etc.). The stepped wedge inherently accounts for common shocks by comparing trends.

11

Ethical Considerations

This evaluation will uphold the highest ethical standards, adhering to UNICEF's Procedures for Ethical Standards in Research, Evaluation, Data Collection and Analysis, and relevant national guidelines. Special attention will be paid to safeguarding the rights, safety, and dignity of all participants—especially vulnerable groups including children, women, persons with disabilities, and forcibly displaced populations.

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 labor). 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.

Fairness in Randomisation and Participation

In a randomized or phased design, a common ethical question is whether it's fair that some get the intervention later. Here, all schools are selected to receive ST4LR by Year 2 (TOF) and Year 3 (RRP/MAP), so no one is denied benefits – it's a timing difference agreed with MoE. The evaluation team will monitor the control schools to ensure they are not adversely affected by being in the study. They will continue with "business as usual" support (including any standard MoE or UNICEF programs they were already part of, like Learning Bridges) so they are not left neglected. If any urgent needs are observed in a control school (say a sudden spike in dropouts due to a crisis), UNICEF will intervene as part of its mandate, even if that somewhat complicates evaluation – human welfare comes first. The evaluation team and program team will maintain communication to uphold the best interest of the child principle.

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