Kingdom of the Netherlands

unicef
for every child

EVALUATION

REPORT 3

UNICEF Impact Feasibility Assessment of PROSPECTS

Ethiopia 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 Mussarrat Youssuf, 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: Ethiopia Case Study', UNICEF: New York, 2025.

Cover photo: © UNICEF/UNI753297/Pouget
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 October 2024, Ethiopia hosts over 1,071,860 refugees and asylum seekers, predominantly from South Sudan, Somalia, and Eritrea, the majority residing in 24 refugee camps across five regional states, with an additional 79,571 living as urban refugees in Addis Ababa. Women and girls constitute 55.9% of the refugee population, and children account for 59.5%.

Furthermore, Ethiopia hosts approximately 3.3 million internally displaced persons (IDPs) and over 2.5 million IDP returnees, mainly due to ongoing conflict in northern Ethiopia, drought, and localized tensions. Recent policy developments, including the roll out of the Comprehensive Refugee Response Framework (CRRF) and the 2019 Refugee Proclamation, indicate growing political will and notable improvements toward refugee inclusion, resilience, self-reliance, and support for host communities. Nevertheless, gaps persist, particularly regarding formal and comprehensive integration of refugees into national systems such as social protection and healthcare.

The PROSPECTS Partnership was established to transform responses to forced displacement crises by integrating humanitarian and development efforts into a cohesive, multi-sectoral approach. Financed by the Government of the Netherlands, the initiative brings together the International Finance Corporation (IFC), the International Labour Organisation (ILO), the United Nations High Commissioner for Refugees (UNHCR), the United Nations Children's Fund (UNICEF), and the World Bank to move beyond short-term assistance responses. It aims to enhance the socio-economic inclusion of FDPs by expanding access to quality education, employment, and protection, while also strengthening the resilience of HCs.

In Ethiopia, UNICEF, ILO, and the World Bank, under the PROSPECTS partnership, are supporting a multi-pronged intervention aimed at strengthening the capacity of national systems to serve FDPs and host communities equitably, the Inclusive Social Protection System. The intervention focuses piloting inclusive cash-plus social protection models, improving public finance mechanisms and referral systems.[1,2] A strong emphasis is placed on capacity-building through technical assistance to key ministries and the development of gender-responsive, inclusive policy frameworks.

The integrated programme takes a multi-pronged approach targeting the same group of FDP and HC households in Ethiopia (specifically in Addis Ababa and Amhara regions). The goal is to strengthen household-level resilience and self-reliance by expanding both access to essential services and opportunities for economic empowerment.

1 UNICEF also supports a second component aiming to extend Community-Based Health Insurance (CBHI) to approximately 5,000 households. These include both FDPs and host communities living in the targeted regions. CBHI coverage will be supported via pooled financing mechanisms to cover premiums for vulnerable households. Alongside enrolment, the programme also envisions strengthening the referral systems so that beneficiaries can move beyond basic health centers and access secondary or tertiary care when needed. A joint feasibility study by UNICEF, ILO, and UNHCR is informing this component's design. Thus, it is important to keep in mind that, while the impact evaluation will focus on the cash-plus component of the intervention for practical reasons, CBHI can contribute to enhance the health impacts of the cash plus component.

2 Please refer to Stage 2 Impact Feasibility Assessment Report for a discussion on the intervention mapping. The mapping included clustering and reorganising the programs and activities into broader interventions using an evaluation lens.

The main component of this intervention focuses on **access to non-collateralized loans** through a pilot partnership between UNICEF and private banks. This cash-plus approach is more than just a financial intervention. It includes wraparound support through linkages with a broader set of social services such as education, child protection, nutrition, health, and WASH. The cash-plus model is designed to empower beneficiaries not only through liquidity or credit but also by improving their capability to use financial resources effectively— enabled by social service integration.

Box 1

Inclusive Social Protection System Selection

The Inclusive Social Protection System 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);[3] and 5) the type of programming, which assessed interventions based on ToC Integration and partners integration.

- **Country Context:** Ethiopia has scored high in terms of country context, with high political will, operational facility, prioritization based on knowledge gaps and national evaluation capacity.
- **Intervention Context:** The Inclusive Social Protection System was ranked as the top priority in Ethiopia (among four interventions included), as it met almost all assessment criteria, including those relating to scale and scalability, plans for future expansion, no existing impact evaluation, knowledge gains and type of programming based on partners integration.

The Inclusive Social Protection System is one of three interventions shortlisted as a priority for Stage 3 (alongside interventions in Jordan and Uganda) and one additional intervention as secondary priority (Kenya) for which impact evaluation plan is being developed.

The Inclusive Social Protection System was selected as a promising intervention for developing 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).

3 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

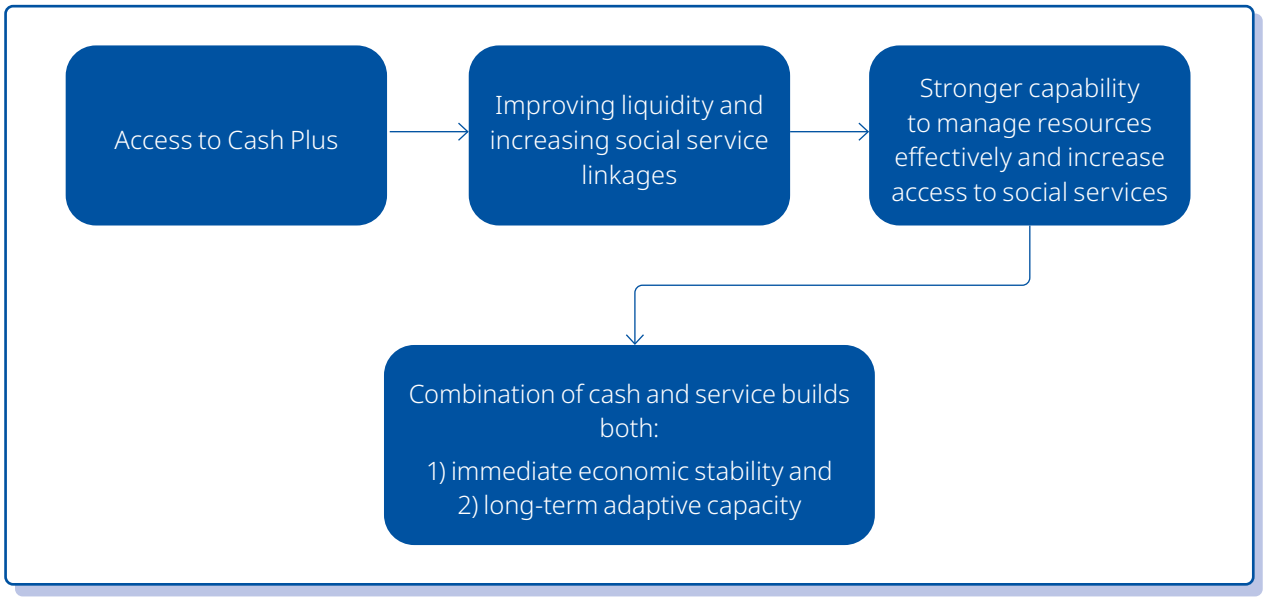
Impact Pathways and Key Outcomes of the Intervention

The Inclusive Social Protection System intervention is explicitly designed to contribute to the outcomes of Pillar 3 (Protection and Social Protection) in the PROSPECTS 2.0 Theory of Change (ToC). It aims to strengthen protection and inclusion for forcibly displaced persons (FDPs) and host communities (HCs) by expanding access to essential social protection services and opportunities for financial support within national and community-based systems.

The Inclusive Social Protection System intervention is explicitly designed to contribute to the outcomes of Pillar 3 (Protection and Social Protection) in the PROSPECTS 2.0 Theory of Change (ToC). It aims to strengthen protection and inclusion for forcibly displaced persons (FDPs) and host communities (HCs) by expanding access to essential social protection services and opportunities for financial support within national and community-based systems.

The intervention generates its outcomes through an integrated approach that reinforces resilience, self-reliance, and dignity among target populations. It expands access to financial services via non-collateralized loans paired with extra support services—including referrals to education, child protection, nutrition, health, and WASH services. This cash-plus model empowers beneficiaries through improved liquidity or credit and enhanced capability to manage resources effectively through linkages with social services. The combination of financial support and service linkages is expected to build immediate economic stability and long-term adaptive capacity, especially when shocks (e.g., illness or climate events) occur (see Figure 1).

Figure 1. Impact Pathways



Importantly, this intervention is framed within broader national and international commitments, ensuring that its impact pathways align with Ethiopia's development frameworks and the centrality of protection in the PROSPECTS partnership's ToC. The approach directly supports Ethiopia's application of the Comprehensive Refugee Response Framework (CRRF) and the country's progressive Refugee Proclamation of 2019, which call for integrating refugees into national systems and equitable access to services. Rather than developing a separate project-specific theory of change, the intervention is anchored in the overarching PROSPECTS 2.0 Theory of Change. This ensures that improvements in education, protection and social protection for FDPs/HCs are aligned with higher-level outcomes of the partnership, and economic inclusion to high-level outcomes such as self-reliance, inclusion, and resilience. However, the ongoing forced displacement crisis—compounded by broader global and regional challenges like protracted conflicts, economic instability, and frequent climate shocks—creates an ever-evolving landscape in PROSPECTS countries that makes achieving these holistic outcomes difficult. Given these constraints, it is more realistic to frame expected results around near- to medium-term outcomes that are achievable within the current context, rather than expecting the full realization of those high-level impacts.

03

Evaluation Questions

The evaluation will assess the relevance, effectiveness and impact of the Inclusive Social Protection System intervention in expanding access to social protection and protection services for FDPs and HCs in Ethiopia. Specifically, it will examine the extent to which the intervention contributed to near- and medium-term outcomes related to inclusion in national systems and access to specialized protective services, aligned with PROSPECTS Pillar 3 outcomes. The evaluation will also explore how these results were achieved, which groups benefited, and the potential for sustainability.

Key evaluation questions include:

IMPACT

- To what extent has the integrated intervention improved household resilience among FDPs and HCs?
- Has the intervention increased self-reliance as evidenced by sustained income generation and reduced dependence on humanitarian aid?
- Have overall well-being and quality of life indicators improved among beneficiary households, including mental health, education participation, and food security?
- Were there differential impacts of the intervention on FDPs and host communities, and how did contextual factors (e.g., legal status, gender, location) influence outcomes?

EFFECTIVENESS

Financial Inclusion

- To what extent did non-collateralized loans lead to the initiation or expansion of income-generating activities?
- What proportion of households report an increase in income, savings, or productive asset ownership as a result of access to loans?
- How did access to financial services influence household coping strategies during financial or livelihood shocks?

Integrated Access to Social Services

- How effectively did the Cash Plus model—through the combination of financial services and linked referrals—facilitate access to health, education, child protection, nutrition, and WASH services for FDPs and HCs?
- Did households accessing multiple services report better well-being and resilience outcomes than those receiving a single intervention?
- Were there improvements in children's school attendance, nutrition status, or protection outcomes attributable to linked service provision?

Systems Strengthening and Inclusion

- How functional and timely were referral systems in connecting FDPs and HCs to necessary services?
- To what extent were programme beneficiaries integrated into national or local systems (e.g. education enrolment, ID systems)?
- To what extent has the intervention strengthened institutional mechanisms (e.g., case management, shared registries, policy linkages) to support long-term integration of FDPs and HCs?

This evaluation plan focuses on the impact and effectiveness criteria of the OECD-DAC criteria, while also examining aspects of sustainability, particularly as they relate to institutional uptake and alignment with national frameworks (e.g., CRRF, the Refugee Proclamation, and social protection policy).

04 Evaluation Design

Key Considerations for Clarification Before Finalizing Evaluation Design

This section draws on currently available information (e.g. PROSPECTS Ethiopia plans), but important details of the Cash Plus pilot remain unclear without further consultation with the country team. Outstanding information gaps – particularly around how the intervention will be rolled out, who it will reach, and on what schedule – need clarification. The following critical questions should be addressed before finalising the impact evaluation design for the Inclusive Social Protection "Cash Plus" component in Ethiopia:

- **Implementation phasing and timeline:** What is the planned schedule for rolling out the Cash Plus component across target areas and populations? Will the intervention launch concurrently in all intended regions (Addis Ababa, Amhara, and Somali) or be introduced in stages (e.g. a pilot phase in one region or a subset of communities before scaling up)? How long will each phase of implementation last, and what is the overall timeframe for reaching all targeted communities?
  - Clarifying the phasing will determine whether a staggered rollout (needed for certain evaluation designs) is feasible, and will help align the evaluation timeline with program operations.
- **Duration of support and service delivery:** How long will each participating household receive support under the Cash Plus pilot, and what are the expected timelines for delivering the financial and social service components? For instance, is the loan a one-time disbursement with a fixed repayment period (e.g. 6 or 12 months), or will beneficiaries have access to multiple loan cycles over an extended period? Over what period will linked social services (referrals to education, health, child protection, etc.) be provided – are these referrals concentrated at the start of the loan, ongoing throughout the loan term, or available for some time after disbursement?
  - Understanding the intended duration and intensity of support for each beneficiary is crucial for planning when outcomes (economic improvements, wellbeing changes, etc.) are expected to materialize and when to schedule midline or endline measurements.
- **Targeting criteria for FDPs and host communities:** Who exactly is eligible to participate in the Cash Plus program, and how are beneficiaries selected among forcibly displaced persons (FDPs) and host communities? What are the inclusion criteria or vulnerability markers used to identify qualifying households. Is there a quota or balanced ratio envisioned between FDP and host community participants in each location, or will coverage be needs-based regardless of status?
  - Clear criteria for how FDPs and hosts are targeted will inform the evaluation's sampling strategy and ensure that analysis can account for any systematic differences in participant characteristics. It will also be important to clarify whether certain sub-groups (e.g. recent arrivals vs. protracted refugees, or specific ethnic communities in host areas) are prioritized, as this could affect both implementation and the generalizability of findings.
- **Eligibility and access to financial services:** What conditions or prerequisites must individuals meet to receive the non-collateralized loans, and are there any regulatory or practical barriers for certain groups? Clarification is needed on how the partnership with private banks will ensure refugees and IDPs can open accounts or borrow.

- Any eligibility restrictions or necessary preparatory steps should be identified, as these factors could influence uptake of the intervention and thus affect how the evaluation is designed

- **Integration of financial and social components:** How will the loan provision be operationally coordinated with the social service referrals in the Cash Plus model, and to what extent are these components delivered as an integrated package versus separate streams?
  - It is important to clarify whether the same beneficiaries who receive loans are automatically enrolled or actively guided into complementary services (and if so, which services are most commonly linked – e.g. entrepreneurship training, family counseling, child education support), or if the "plus" services are more ad-hoc referrals provided upon request/need.
  - The evaluation design will need to account for the degree of integration, since the theory of change assumes that the financial and social elements together drive outcomes. If the two components are not well-coordinated or if referral uptake is low, the impact evaluation may need to measure the utilization of services and consider differential effects for those who receive both the financial and social support versus only one part.

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

Finally, an important conceptual question for this intervention remains unaddressed in this evaluation: To what extent do the outcomes generated by non-collateralized loans differ (are superior or inferior) to those generated through direct grants or asset transfers and how do these modalities compare in their impact-efficiency? The current implementation plan of the interventions does not permit direct testing of this; the question represents an important potential research avenue. Future evaluations or follow-up studies might explore if a loan-based approach alone can achieve similar long-term impacts in refugee settings, provided conditions and resources permit such comparative analysis.

Overall Design

The evaluation will employ a mixed-methods, quasi-experimental design Ethiopia. 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, which includes access to non-collateralized loans, and integration with social services (education, child protection, WASH, and nutrition). 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 Addis Ababa and Amhara.

Quantitative Evaluation Designs

This section outlines the proposed evaluation strategy for the inclusive social protection intervention currently being rolled out in urban displacement-affected settings in Ethiopia. The evaluation seeks to establish whether the intervention achieves its intended effects on key outcomes such as resilience, access to services, and psychosocial wellbeing. We propose two broad scenarios to accommodate the operational realities of this evaluation. In the first scenario, we assume that randomisation at the community level is possible, allowing us to implement a rigorous Randomized Controlled Trial (RCT) design. In the second scenario, recognising that justifying the exclusion of certain communities from receiving the cash plus intervention might be challenging, we propose two quasi-experimental alternative designs that rely on statistical matching.

The primary design under the randomisation scenario is a cluster-level RCT. However, recognising the ethical and operational difficulties of maintaining a clear comparison group in settings such as the ones presented here, we also propose two viable alternative designs: a Propensity Score Matching (PSM) DiD and a retrospective matching design. 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.

Each of these approaches is described below in detail, with particular attention to their strengths, assumptions, and feasibility in the Ethiopian context.

Primary Design: Randomized Controlled Trial (RCT)

If random assignment at the community level is operationally feasible in Ethiopia, the primary evaluation strategy to consider is a cluster Randomized Controlled Trial (RCT). In this design, eligible communities across the target regions (Addis Ababa, Amhara, and Somali) would be randomly divided into intervention and control groups, allowing for a direct comparison of outcomes between communities that receive the Cash Plus program and those that do not (initially) receive it. This parallel cluster RCT would provide a robust counterfactual and the clearest attribution of impact, as randomization ensures that – on average – the communities in both groups are equivalent in observed and unobserved characteristics prior to the intervention.

Communities (for example, kebeles or wards hosting FDP and host populations) would be the unit of randomization. All communities selected for the pilot would be randomly assigned – likely stratified by region and other key factors (such as urban vs. rural, or proportion of refugees in the community) – into one of two arms:

- **Immediate Cash Plus arm (Treatment):** Communities in this arm begin receiving the full Cash Plus intervention from the start of the evaluation period. Households in these communities gain access to the non-collateralized loans through the UNICEF–private bank partnership, along with the intended package of social service linkages.
- **Delayed or No-Intervention arm (Control):** Communities in this arm do not implement the new Cash Plus components during the evaluation period. They continue under status quo conditions (or any existing support unrelated to the Cash Plus pilot), serving as a comparison group. For ethical and practical reasons, these control communities could be slated to receive the Cash Plus program after the evaluation is completed (a "wait-list control"), ensuring they are not permanently denied the benefits but rather delayed in receiving them.

Randomizing at the community/grographic (e.g. woreda) level (as opposed to individual households) is preferred here to prevent spillovers and contamination – it reduces the risk that control-group individuals access the program indirectly (e.g. by moving to or borrowing in a nearby treatment community). Stratified randomization will help ensure balance between the two groups (so that, for instance, each region has a fair mix of treatment and control communities, and one arm does not by chance contain more refugee-dense communities than the other). This approach capitalizes on any flexibility in program rollout: if the implementing partners are willing to stage the introduction of the pilot by community, the evaluation can leverage that to create a rigorous experimental setup.

The cluster RCT implementation would follow two main phases aligned with key outcome measurements:

Baseline (Pre-intervention): All selected communities (treatment and control) would be surveyed before the intervention starts. This baseline establishes pre-program conditions (e.g., household income, resilience, employment, service access), confirming comparability between groups and improving analytical precision.

Phase 1 – Intervention Rollout (Year 1): Immediately after baseline, treatment communities receive the Cash Plus intervention (loans and social service referrals), while control communities maintain the status quo. Midline data collected after approximately one year would allow estimation of short-term impacts (e.g., household income, business activities, service uptake) by comparing treatment and control groups.

Endline (Year 2): After midline data collection, the intervention could be introduced to control communities using a wait-list design. By the end of Year 2, all communities will have received the Cash Plus program (treatment group for two years, control group for one year). The endline survey at this point would measure outcomes for both groups, providing insight into whether initial gains observed at midline are sustained over time and allowing analysis of the program's broader effectiveness. Alternatively, if feasible, control communities could remain untreated throughout the evaluation period, providing a longer-term, pure comparison at the endline.

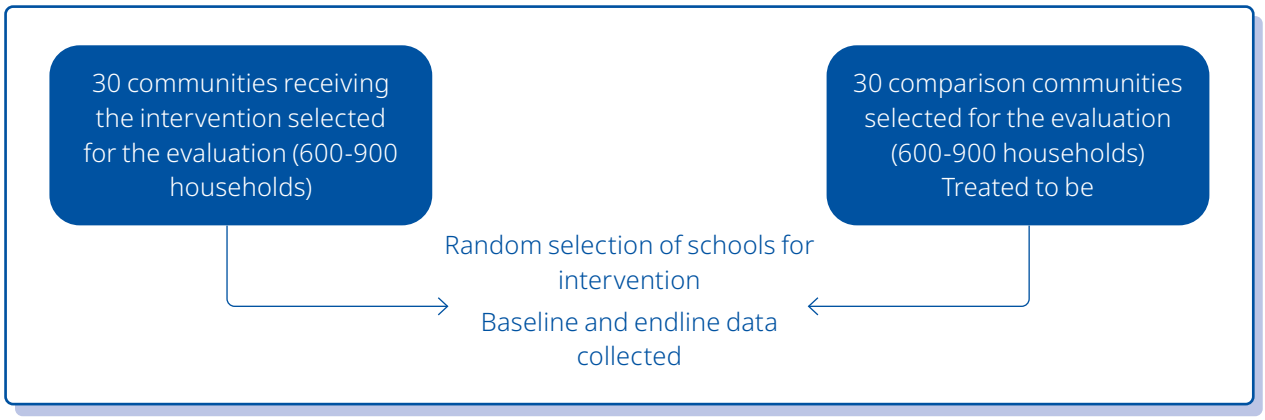
Throughout these phases, careful monitoring of program implementation is needed to ensure that the treatment communities indeed receive the intervention as planned and that control communities do not inadvertently receive similar benefits. The fidelity of the rollout is crucial: if some control areas were to receive loans through other channels or if participants migrate between communities, it could dilute the measured impact. Therefore, the implementation will include safeguards such as distinct geographic boundaries for participating communities and communication protocols so that partner banks and service providers only extend the Cash Plus offerings to designated treatment clusters during Phase 1.

Sample Size and Power

The cluster RCT approach will inform the required sample size in terms of number of communities and households. In general, to detect moderate impacts on household-level outcomes, we would aim for a substantial number of clusters per arm to improve precision. For illustrative purposes, suppose the pilot could engage on the order of 60 communities in total. These might be split into roughly 30 treatment and 30 control communities. If we plan to survey, say, 20–30 households per community, this yields approximately 600–900 households in each arm, which should be sufficient to detect an effect size of practical interest.[4]

4 These numbers should be refined with formal power calculations once more details are known about outcome variability and ICC in the target population. Key parameters for the power analysis include the expected impact, the ICC, and the baseline-to-endline correlation of outcomes. If the ICC is low and outcomes are volatile, fewer clusters might suffice; if ICC is high or anticipated impacts are small, more clusters or households would be needed.

Figure 2. Randomized Controlled Trial



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

In case the random assignment proposed in the preferred method is not feasible, but pre-intervention (baseline) data can be collected, a Propensity Score Matching with difference-in-differences framework will be applied. This method estimates the probability of selection into the intervention based on observable baseline characteristics, then matches Cash Plus participant units to non-participant units with similar propensity scores. After matching, a DiD estimator is applied to the matched pairs/groups.

Under PSM-DiD, a propensity score (the probability of receiving the intervention) is estimated for each unit (e.g. each community or household) based on baseline observable variables (such as demographics, poverty level, prior service access, etc.). For each treated unit (in an intervention community), one or more comparators are selected from the non-treated pool with a similar propensity score. This matching on the propensity score aims to achieve balance between treated and comparison groups on all observed covariates. Once the matching is done, the evaluation compares the before-and-after changes in outcomes between these matched groups, applying the DiD estimator to the matched sample. This yields an impact estimate analogous to the primary design, under the key assumption that conditional on the observed covariates (and thus propensity score), the treatment and matched control would have followed parallel outcome trends absent the program.

The PSM with DiD design is well-suited for situations where community selection was not random but baseline data are available. For instance, if certain neighbourhoods were chosen first due to perceived need, we can still evaluate impact by finding other neighbourhoods with similar needs that didn't yet get the program (or were ineligible) and using baseline data to match them. The strength of this design is that it maximizes the use of baseline information and can approximate the counterfactual in a statistically rigorous way. However, it relies on the assumption that all important differences between the intervention and comparison groups are observed and included in the propensity model. If some unobserved factor influenced both the likelihood of receiving the intervention and the outcomes (violating "selection on observables"), it could bias results. We will mitigate this risk by including a rich set of covariates in the propensity model (household demographics, location characteristics, etc.) and by conducting sensitivity analyses (e.g. checking robustness to unobserved confounding via simulation of potential hidden bias).

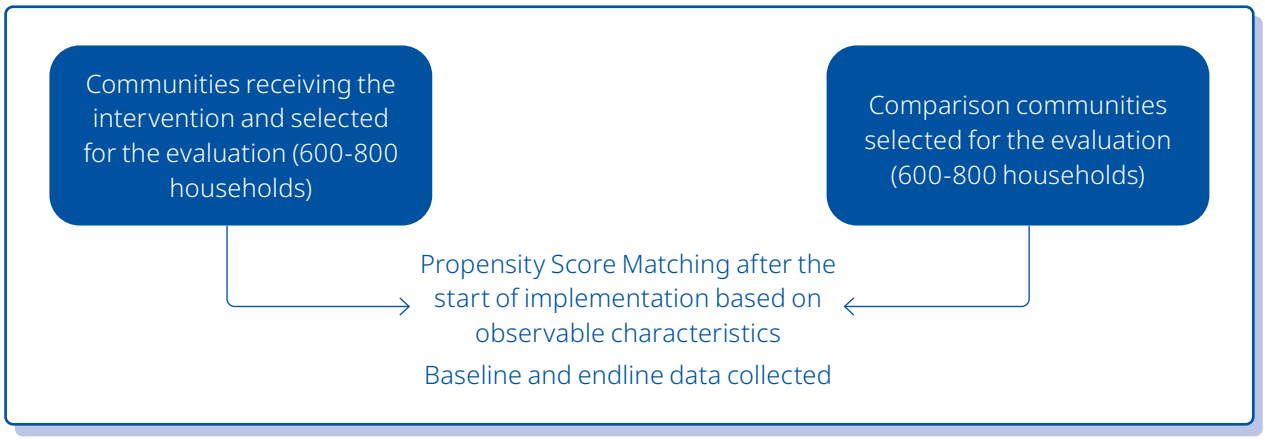
Another requirement is a sufficient overlap in characteristics – there must be enough "common support" where propensity scores of treated and untreated units overlap, so that each treated unit finds a comparable control. Provided this overlap and quality baseline data (from surveys or administrative sources), the PSM-DiD can yield impact estimates with moderate causal strength. It lacks the guaranteed balance of true randomization but still controls for a wide array of factors and uses DiD to net out time-invariant differences.

Sample Size and Power

Because this design may introduce some inefficiency (not all sampled units can be matched, and matching may not perfectly pair every treated unit), slightly

larger samples are ideal. We anticipate needing on the order of 600–800 households per group to achieve similar power as the primary design. This accounts for potential loss of sample in the matching process and a possibly higher variance if matching is imperfect. Nonetheless, with the total target population in the thousands, obtaining around 600 plus matched pairs is feasible. The DiD element still contributes to precision by controlling for baseline outcome levels. If the actual available sample is smaller, we may need to focus on the most critical outcomes or use stratified matching to improve precision. Overall, PSM-DiD is a strong alternative when a randomized control is not obtainable, ensuring the evaluation remains on solid footing in terms of internal validity.

Figure 3. Propensity Score Matching with Difference in Differences



Alternative Design 2: Retrospective Matching

For communities where the intervention has already begun without any baseline data or defined control group, a retrospective matched design will be used. This is essentially an ex post evaluation approach: after the intervention rollout, we intentionally reconstruct the counterfactual using whatever information is available about pre-intervention conditions.

In a retrospective matching scenario, the evaluation would sample households from program areas that have already been treated and from other areas not covered by the program, then collect data at a single point in time (post-intervention) for both groups. Because no baseline surveys were done, we rely on a combination of recall data and existing records to establish pre-intervention characteristics. For example, surveys might include retrospective questions asking households to recall key status indicators from before the program (such as their income level last year, or whether they were using certain services before). Administrative data or secondary sources (like health facility records or previous studies in those areas) may also provide baseline proxies. Using this reconstructed baseline information, we then match treated and comparison units post hoc – similar to PSM, we can match on recalled pre-intervention metrics or on inherently time-invariant traits. Once matched groups are formed, outcome differences between the intervention and comparison group (at endline) are analysed, possibly with regression adjustments to account for any remaining differences.

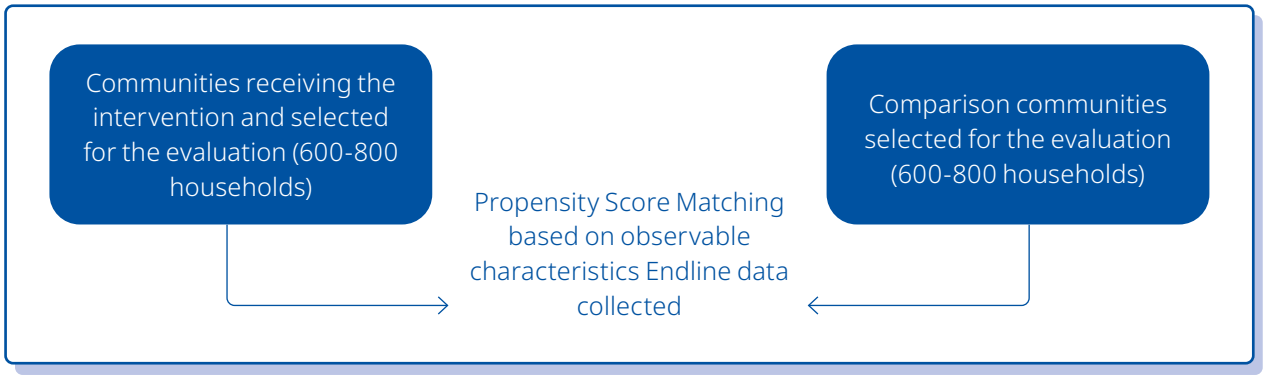
This last resort review can still provide valuable insights, especially if combined with qualitative evidence and if the differences in outcomes are large enough to be convincing. We will bolster credibility by: (a) using multiple matching techniques (e.g. matching on several recalled indicators, or using statistical controls in regression models), (b) conducting sensitivity checks (such as bounding exercises to see how large unmeasured bias would have to be to nullify the results), and (c) integrating findings with the process evaluation to see if outcome differences align with what stakeholders observed on the ground.

Sample Size and Power

In the absence of baseline control for variance, achieving adequate sample size is even more important. We will target a larger sample per group (as possible, 600–900 households in treatment group and control group) to compensate for the noisier data.

More observations help average out recall errors and provide greater degrees of freedom to control for differences. The evaluation team will thus emphasize collecting high-quality data and perhaps focus on more objective indicators (e.g., assets owned, children in school as verified by records, etc.) that can be recalled with less error or triangulated. With these measures, the retrospective matching design can yield useful estimates of program impact in areas that would otherwise be unevaluable, ensuring the evaluation covers all implementation settings.

Figure 4. Retrospective Matching



Summary of Research Designs

Each of the three approaches balances methodological rigor with field feasibility. The evaluation is designed to be adaptive – implementing the strongest possible design where conditions allow but prepared to shift to alternative methods in other contexts. Table 1 summarizes the core features of the primary and alternative designs, highlighting their requirements and best-use scenarios:

Table 1. Evaluation Design Options

	Design A: RCT	Design B: PSM DiD	Design C: Retrospective Matching
Randomization	Yes – cluster-level (community level)	No	No
Baseline Data	Yes	Yes (strongly preferred)	No
Matching Timing	Prospective – comparison groups defined before implementation	After implementation	After implementation
Matching Method	Random assignment ensures balance	Statistical matching (Propensity Score)	Statistical matching on reconstructed baseline
Causal Inference Strength	**High** – Strong internal validity due to randomization	**Moderate** – Good control of observables; some residual bias risk	**Low–Moderate** – Indicative results; higher uncertainty
Best Use Case	Randomization is feasible and ethical, and baseline data can be collected.	Programme started but baseline data exist	Programme already implemented without baseline

05 Key outcome indicators

A suite of key outcome indicators has been defined to measure the Inclusive Social Protection System's performance. These indicators align with the intervention's impact pathways (Section 2) and address the evaluation questions (Section 3), ensuring we capture changes across all result areas – from health and economic outcomes to protection and inclusion.

Table 2 below presents a simplified evaluation matrix, grouping indicators by thematic focus. For each category, it lists the core indicators (with some disaggregation where applicable). This comprehensive indicator set covers short-term outputs (e.g. service uptake), medium-term outcomes (e.g. behaviour changes, well-being improvements), and longer-term impacts (e.g. resilience, social inclusion).

Table 2. Simplified Evaluation Matrix

Level	Statement	Indicators
	Improved resilience	% increase in food security scores; % of households reporting reduced reliance on negative coping strategies (e.g. borrowing, child labour)
	Improved self-reliance	Percentage increase in utilisation of basic services, including:
	Improved well-being	Education (school attendance among children) [5]
		WASH services (access to safe drinking water, latrines)
		Percentage of households reporting reduced reliance on negative coping strategies (e.g. borrowing, child labour)
		Self-reported household well-being and stability index
Outcome	Improved access to a wide range of basic services	% HH becoming eligible for/ accessing non-collateralized loans
		% HH receiving loans (disaggregated by sex, age, displacement status)
		# of service referrals issued, accepted, and completed
		% of referred cases that accessed services within XX weeks
		# of service providers integrated into referral network (health posts, schools, social workers)

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

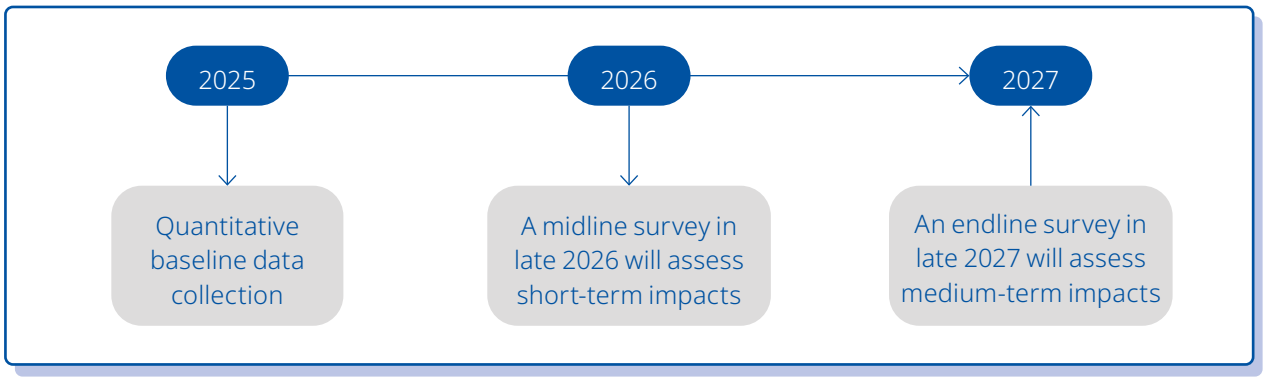
Output	FDPs/HCs receive loans via private bank partnerships	# of bank branches offering non-collateralized loans to FDPs/HCs
	FDPs/HCs linked to services through referral system	# of referral coordinators/social workers trained
		# of functional referral tracking tools deployed and used
		% of planned activities completed according to timeline and budget

06 Timeline

The evaluation will be conducted over a multi-year period (approximately 2025–2027) to capture both immediate effects and longer-term impacts. A phased implementation of the program is assumed, and the evaluation timeline is aligned accordingly in several key stages. A timeline is presented in Figure 5 and the details on the phases explored below.[6]

Baseline (2025 – Pre-Intervention): Before intervention rollout, a comprehensive baseline survey will be conducted in all selected treatment and control communities in Addis Ababa, Amhara, and Somali regions. This baseline establishes pre-intervention values for key indicators (e.g., income, employment, resilience, access to services), confirms comparability of groups, and informs the evaluation design. Qualitative assessments will also be conducted, capturing local contexts, stakeholder expectations, and potential implementation barriers. Enumerators will be trained and survey instruments piloted in this stage.

Figure 5. Evaluation Timeline



Phase 1 Implementation (2025–2026): Immediately following the baseline, the Cash Plus intervention (non-collateralized loans with social service referrals) begins in the randomly assigned treatment communities. Control communities receive no new intervention during this initial phase. Throughout this period, administrative data will track loan distribution and referral activities. A midline survey in late 2026 will assess short-term impacts, comparing outcomes between treatment and control communities to estimate the program's immediate effectiveness. Additionally, qualitative methods will be used to document early experiences, beneficiary perceptions, and contextual challenges.

Endline (2027 – Post-Intervention): After midline data collection, control communities may begin receiving the Cash Plus intervention as a delayed treatment (wait-list design). By the end of 2027, all communities will have received the program—treatment communities for approximately two years, and control communities for one year. The endline survey will be conducted at this point to measure outcomes for all communities. It will assess whether initial gains observed at midline persist and evaluate broader program impacts, such as household resilience, economic stability, access to services, and psychosocial wellbeing. This final survey will follow a panel design, re-interviewing households surveyed

6 The phases are presented based on the Primary Design and should be adjusted depending on feasibility of implementation.

at baseline. Complementary qualitative work—including interviews with key stakeholders and community discussions—will explore perceived impacts, unintended effects, and sustainability of changes.

This structured timeline allows the evaluation to rigorously measure short-term impacts (2026) and longer-term outcomes (2027), aligning measurement closely with implementation realities while capturing meaningful trajectories of change.

07

Data Collection Methods

To support this design, the evaluation will utilize a combination of data sources and collection methods, each aligned with the units of analysis and the mixed-methods approach:

Household and Individual Surveys: Structured surveys will serve as the backbone of the quantitative data collection at baseline and endline (and any midline, if conducted). A detailed household questionnaire will be used to capture demographic information, socio-economic status, consumption and income, access to services, and shocks. Within each sampled household, an individual questionnaire (or roster sections) will be administered to target specific respondents; By combining household and individual modules, coverage of all components' indicators will be ensured. The baseline survey in 2025 will be administered to a large sample of both program beneficiaries and comparison group members (matched non-beneficiaries in the same communities or in similar communities not yet reached). The endline survey in 2028 will aim to re-interview all baseline households and individuals. Survey implementation will be carried out according to high standards: enumerators will be trained for work in sensitive contexts (such as refugee camps and urban poor neighbourhoods), instruments will be translated into local languages, and digital data collection will be used to support quality control.

Administrative and Program Monitoring Data: The evaluation will make extensive use of program administrative records to complement survey data and reduce respondent burden. Program's Management Information System (MIS) will supply data on the delivery of benefits: lists of beneficiaries receiving non-collateralized loans, amounts and timing of payments. These administrative data streams will be collected continuously by the implementers and shared with the evaluation team at intervals. Such data is invaluable for monitoring implementation fidelity – checking if the intended services reach the target population in the planned doses. It also enables cross-checking survey responses (for instance, confirming whether a household that says it received a transfer indeed appears in the payment records). In addition to programme MIS data, the evaluation may draw on national or regional statistics (e.g. health facility utilization rates, labour market statistics in urban areas) to understand broader trends during the period and to contextualize the findings, if available.

Referral and Case Management Data: Since the intervention links to child protection and other services, any referral logs or case management databases can be useful. For instance, if social workers refer beneficiaries to protection services, they might record the referral and outcome (service received or not). These records can indicate whether the integrated approach is operational (are people actually being referred and served?)

Qualitative Methods (KIIs, FGDs): A robust qualitative data collection will run in parallel with the quantitative collection, aimed at understanding the why and how behind the numbers. KIIs will be conducted with stakeholders at multiple levels: programme administrators, local implementers, and community leaders. These KIIs will explore topics such as how the interventions were implemented on the ground, any operational challenges, perceptions of the program's effectiveness, and coordination between agencies. FGDs will be held with groups of beneficiaries and non-beneficiaries. For example, separate FGDs may

be organized with refugee community members about their experience and discuss any changes in social dynamics or service access. The FGDs will probe issues like barriers to participation (such as cultural or information barriers for refugees accessing insurance), satisfaction with services (e.g. the quality of healthcare received), and perceived impacts (do participants feel more financially secure or healthier).

The qualitative data collection will be timed to align with key phases – e.g. some KIIs/FGDs at mid-term (to assess implementation progress and community feedback) and more at endline (to assess outcomes and overall experiences). All qualitative sessions will be conducted by trained facilitators in the local language, with transcripts translated for analysis. Notes on contextual factors (such as local economic conditions, security issues, or cultural norms) will also be documented during field visits, as these factors are critical to understanding variations in program performance across locations.

08 Analysis Methods

RCT Design: For areas implementing the RCT, the analysis will estimate the average treatment effect by comparing outcomes across randomized community clusters with control units. The core analytical model will follow a simple OLS specification, comparing outcome changes in treatment arms relative to control communities.

The preferred regression model is:

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

Where:

- Y_{it} is the outcome for household or individual i at time t,
- $Treatment_i$ indicates assignment to treatment status,
- 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., community or enumeration area) to account for intra-cluster correlation. Where multiple follow-up points are available (e.g., baseline, interim, 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 communities 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: Qualitative data (KIIs and FGDs transcripts) will be analyzed using a thematic analysis approach. A coding framework reflecting the key evaluation questions and emergent themes (e.g., themes such as "implementation fidelity," "barriers to access," "perceived benefits – health," "perceived benefits – income," "social cohesion," "unintended effects," etc.) will be developed by the evaluation team. Transcripts will be coded line-by-line according to these themes using Nvivo. Through this systematic coding, patterns and insights will be extracted. Commonalities and divergences in perspectives will be sought by the analysts; for instance, whether multiple respondents independently mention the same facilitating factor or bottleneck will be explored. Particular importance will be placed on the search for explanatory mechanisms—the qualitative data will be mined to understand how the programmes led to changes. For example, if mixed results are observed in the employment

program, FGDs might point to local job market saturation as a limiting factor. Direct quotes from participants may be used in the report to illustrate key findings, allowing voice to be added to the statistics.

Integration and Triangulation: Data triangulation will also be carried out by cross-verifying findings from the surveys, qualitative narratives, and administrative records. If qualitative feedback suggests an important outcome or issue that was not fully captured by the quantitative metrics, this will be noted as an area for program learning. Conversely, if a significant quantitative change is observed (e.g., a drop in some hardship indicator), the qualitative data will be examined to see if participants and stakeholders also perceived this change and how they explain it. This integrative analysis will enrich the evaluation conclusions and help formulate actionable recommendations.

09

Estimated Resources for Data Collection

This section will be further refined in collaboration with UNICEF's Ethiopia Country Office, as additional information becomes available regarding the number of participating communities, total beneficiary targets, and operational coverage the components of intervention. Details on the scale and geographic rollout of the cash plus components, including the number of participating cities or eligible individuals, are not yet confirmed. The full cost estimate for data collection will depend on final sample sizes, unit of analysis (household or individual), agreed number of survey waves, and geographic dispersion. Once KIIs with Country Office staff are complete, a more precise costing—covering quantitative surveys, qualitative fieldwork, and administrative data acquisition—will be provided.

10 Feasibility and Limitations

Feasibility Considerations

Ethiopia offers both opportunities and challenges for implementing a rigorous mixed-methods evaluation of the Inclusive Social Protection System under the PROSPECTS 2.0. On the one hand, the country exhibits strong political commitment to refugee inclusion—demonstrated by the roll-out of the CRRF, the progressive 2019 Refugee Proclamation. The Government's recognition of social protection as a national priority and openness to evidence-based policy also create an enabling environment for evaluation.

Furthermore, Ethiopia has an expanding base of national and regional data systems (including health and social protection MIS), and several local academic and research institutions with experience in impact evaluation, which can be leveraged for local partnerships and capacity building. These factors enhance the feasibility of longitudinal data collection and stakeholder engagement across implementation phases.

Operationally, urban sites such as Addis Ababa are relatively accessible and pose fewer logistical or security barriers. Likewise, in some regions like Somali and Amhara, previous programming has already established community entry points and working relationships with regional authorities, UNHCR, and civil society actors. This can facilitate the roll-out of the baseline, monitoring, and endline exercises in a coordinated manner.

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. We can switch to using a quasi-experimental method for that subset while still carrying out an RCT in another. This flexibility ensures that even under suboptimal conditions, we will collect usable data and generate findings. It de-risks the evaluation from a feasibility standpoint.

Potential Limitations and Mitigation

Security and Access Constraints

Several regions targeted by the intervention—particularly Amhara and parts of Somali—are prone to conflict, natural disasters, and displacement-related volatility. These conditions may hinder safe access for enumerators, affect community stability, and introduce delays or incomplete data collection.

Mitigation: The evaluation will coordinate closely with implementing agencies and regional authorities to monitor security conditions and develop site-specific protocols. Remote data collection (e.g., phone surveys) may be used if physical access becomes unsafe. Oversampling in more stable regions will also be considered to offset potential gaps.

Risk of Attrition in Panel Surveys

Given the high mobility of forcibly displaced populations and the fluidity of urban and camp-based settings, attrition between baseline and endline surveys is a significant risk. Attrition could bias impact estimates if those who remain differ systematically from those lost to follow-up.

Mitigation: At baseline, the team will collect detailed contact and locator information (including phone numbers, alternate contacts, and local leaders) to support follow-up. Where feasible, community-based enumerators will be used to maintain links. Statistical corrections (e.g., inverse probability weighting) will also be applied during analysis to address attrition bias.

Non-Randomised Intervention Allocation

While a Randomized Control Trial is proposed, the absence of random assignment to intervention arms could lead to selection bias. This is especially pertinent if areas receiving cash plus benefits support differ in unobserved ways from comparison sites.

Mitigation: Matching will be done rigorously using pre-intervention characteristics from administrative or survey data, and robustness checks will be performed. If the tentative randomized sub-sample are not feasible, two other quantitative evaluation designs are proposed.

Data Gaps and MIS Limitations

Administrative data systems, while improving, may still suffer from inconsistent recordkeeping or missing data, particularly on service uptake or disaggregation by refugee status, gender, or disability.

Mitigation: The evaluation will triangulate administrative data with independent household and individual surveys. Where possible, evaluators will support partners to improve data quality and ensure alignment with indicator frameworks through capacity strengthening.

Sociocultural Barriers to Participation

Social norms, especially around gender, disability, and refugee identity, may inhibit participation in surveys or focus groups—particularly among adolescent girls or women from conservative communities.

Mitigation: Enumerators will be trained on gender and cultural sensitivity. Female enumerators and facilitators will be used for certain subgroups, and FGDs will be stratified by gender, age, and refugee status. Interviews will be conducted in local languages, and sessions will ensure privacy and comfort for respondents.

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 under the Ethiopian National Research Ethics Review Committee (NRERC). 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

Participation in the study (surveys, interviews, etc.) will be fully voluntary. For respondents under 18 years old (since the intervention targets adolescents, some will be minors), the evaluation will obtain informed assent from the youth and consent from a parent or guardian. For those 18 and above, their own informed consent is sufficient. Consent forms will explain in the appropriate language the purpose of the evaluation, the procedures, potential risks and benefits, and the respondent's rights (including the right to refuse or withdraw at any time without any consequences on their involvement in the program). Special care will be taken to assure participants that declining to participate in the evaluation will not affect their access to the intervention services or any aid – to avoid any coercion.

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.

Privacy and Confidentiality

The evaluation will ensure strict confidentiality of respondents' data. All survey and interview responses will be anonymized – identified by codes rather than names. Personal identifiers collected for panel tracking will be stored securely and separately from survey responses. Data will be encrypted and access limited to the core evaluation team. In reporting, no names or identifying details will be revealed; any quotes used will be generic or attributed to anonymized descriptors (e.g., "Refugee woman, Somali region"). Prior to stakeholder workshops or dissemination, the evaluation will also seek consent if we plan to share any photos or video documentation (though primarily this evaluation uses written data). All data handling will comply with relevant data protection laws and UNICEF's data privacy standards

Ethical Approvals and Coordination

The evaluation will obtain ethical clearance from a recognized Institutional Review Board (IRB) in Ethiopia (e.g., Addis Ababa University College of Health Sciences Institutional Review Board) as well as from UNICEF's own Ethics Review Panel. The research protocols (survey instruments, consent scripts, etc.) will be submitted for review to ensure they meet ethical standards. Given that refugees are a potentially vulnerable population, the IRB will scrutinize plans to ensure no exploitation or undue intrusion into their lives. The evaluation will also coordinate with UNHCR for approval to conduct research in refugee settlements, following their guidelines for research with persons of concern.

Cultural and Gender Sensitivity

The evaluation team will be diverse and include members who speak the local languages and understand the cultural norms of both refugee communities (South Sudanese, Somalian, Eritrean.) and Ethiopian host communities. Instruments will be translated and back-translated to ensure meaning is preserved. Interviews and FGDs will be conducted in a gender-sensitive and inclusive manner—e.g., using female facilitators with women and adolescent girls, and ensuring accessible venues for persons with disabilities. Where necessary, sign language interpretation or community liaison support will be provided.

Fairness in Randomisation and Participation

In components involving random assignment (e.g., cash-plus RCT), fairness and transparency will be prioritized. Participants will be fully informed of selection processes, and where feasible, wait-list controls or phased inclusion will be used to ensure all eligible participants have access over time.

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 communities and participants in an accessible format (e.g., community 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