Prospective Evaluation Strategy

Second Report in the Capstone Report Series

Marzia Azizi, Siya Kasat, Doris Luo, and Daija Yisrael

Georgetown University

Advisor: Prof. Jacobus Cilliers

May 4, 2026

Contents

List of Tables

Glossary

Abbreviation	Full Form / Definition
DiD	Difference-in-Differences
EE	Eligibility Engine
HRLM	Haryana Rural Livelihoods Mission
IA	Indus Action
IV	Instrumental Variable
IVRS	Interactive Voice Response System
near-PLD	A woman SHG member who narrowly missed the Eligibility Engine threshold and was therefore not matched to welfare schemes under the Lakhpati Didi initiative; used as the control group in this study
PLD	Potential Lakhpati Didi
RDD	Regression Discontinuity Design
SHG	Self-Help Group
SMS	Short Message Service
SRLM	State Rural Livelihoods Mission

1. Introduction

Given that Indus Action seeks to examine the effect of their intervention package on women's income growth and trajectories, this report, as the second in the capstone report package, provides a prospective framework to evaluate such the causal impact. Designing a credible evaluation strategy requires careful consideration of three features of this study's context. First, PLDs and near-PLDs were not randomly assigned — PLDs were pre-identified by the Haryana government based on composite income-related criteria, creating systematic pre-existing differences between the two groups that must be accounted for. Second, the intervention is compound, combining scheme matching, outreach, and market linkage support, which makes isolating the effect of any single component difficult. Third, the baseline balance check reveals significant differences between PLDs and near-PLDs on key capacity measures that further inform the choice of evaluation approach.

The structure of this report is as follows. Chapter 2 presents the balance check results. Chapter 3 demonstrates our prospective evaluation approach — matching plus Difference-in-Differences — and explains why this method is most viable given the study design. Chapter 4 lists the key assumptions that need to be held to use this approach, their plausibility in this context, and how they can be possibly examined contingent to future steps. Chapter 5 specifies the data and information that need to be collected in future rounds to implement this evaluation strategy. Chapter 6

2. Balance Check

We conducted a balance check of both the covariates and (potential) outcome variables at baseline before making proposal of possible evaluation approach, as the extent and nature of baseline imbalances will directly inform which identification strategies are viable.

2.1 Baseline Characteristics

Table 2.1. Baseline Covariate Balance

	Control		Treatment		t-test
Variable	N	Mean	N	Mean	p-value
Panel A: Individual characteristics					
Age	105	36.486	406	39.185	0.020*
		(9.297)		(10.813)	
Illiterate	105	0.390	406	0.431	0.454
		(0.490)		(0.496)	
Currently married	105	0.895	406	0.892	0.915
		(0.308)		(0.311)	
Is SHG leader†	105	0.476	406	0.303	0.001***
		(0.502)		(0.460)	
Panel B: Household characteristics					
Household size	105	5.752	406	5.739	0.958
		(2.098)		(2.410)	
Owns land	105	0.267	406	0.202	0.151
		(0.444)		(0.402)	
Has debt	105	0.343	406	0.300	0.403
		(0.477)		(0.459)	

Notes: Standard deviations are in parentheses. The p-values are from two-sided t-tests comparing means between control and treatment groups. * p<0.05, ** p<0.01, *** p<0.001. † Is a SHG president, secretary, or treasurer.

As shown in Table 2.1, treatment (PLDs) and control women (near-PLDs) differ on two key characteristics: on average, PLDs are on average about three years older (39.2 vs. 36.5 years, p=0.020), while near-PLDs are substantially more likely to hold SHG leadership positions like president, secretary, and treasurer (48% vs.30% p=0.001). Other characteristics including literacy, marital status, household size, land ownership, and debt levels show no statistically significant differences.

The SHG leadership imbalance is particularly consequential because leaders typically have stronger networks, better access to information, and greater entrepreneurial capacity — factors that may independently influence income trajectories (details discussed in section 2.2 and Table 2.3). This observable difference need to be addressed in any evaluation strategy.

2.2 Baseline Outcomes

Table 2.2. Baseline Outcome Balance

	Control		Treatment		t-test
Variable	N	Mean	N	Mean	p-value
Panel A: Individual Outcomes					
SHG Support Index	105	0.289	406	-0.075	0.001***
		(0.825)		(1.028)	
Women's Agency Index	105	1.785	406	1.803	0.660
		(0.383)		(0.366)	
Panel B: Scheme Outcomes					
Prior Scheme Putreach§	105	0.152	406	0.246	0.041*
		(0.361)		(0.431)	
Number of Schemes Prior Benefited	105	0.810	406	0.857	0.555
		(0.652)		(0.756)	
Panel C: Economic Outcomes					
Own enterprise†	98	0.847	314	0.678	0.001***
		(0.362)		(0.468)	
Enterprise Income‡	98	7,491	317	7,589	0.917
		(7,069)		(8,400)	
Business Profit‡	74	4,699	229	4,239	0.456
		(3,981)		(4,790)	
Entrepreneurial Confidence Index	105	0.231	406	-0.060	0.008**
		(0.840)		(1.030)	
Market Linkage Index	105	0.309	406	-0.080	0.000***
		(0.638)		(1.060)	

Notes: Standard deviations are in parentheses. The p-values are from two-sided t-tests comparing means between control and treatment groups. * p<0.05, ** p<0.01, *** p<0.001. § Ever received message about scheme eligibility (any government or NGO program). † A binary indicator derived from the number of enterprises variable: 0 = no enterprises reported, 1 = one or more enterprises reported. ‡ Reported only for women with active enterprises at baseline. Some women who reported zero enterprises nonetheless reported a non-zero enterprise income, which results in the discrepancy in N.

As shown in 2.2, there are four significant imbalances on outcome measures [1]. On average, near-PLDs report higher level of support from their SHGs, entrepreneurial confidence, and market linkage. near-PLDs also report higher prior exposure to scheme outreach: 25% of PLDs versus 15%

[1]Final outcome selection will be determined by IA in consultation with government stakeholders (see Section 6.4 for methodological considerations). We present the balance balance on these variables because they are possible candidates given IA's interest and their baseline imbalances have direct implications for evaluation strategy.

of near-PLDs had previously received messages about scheme eligibility from any government or NGO program (p=0.041) [2].

Table 2.3. SHG Leadership vs Capacity Measures

	SHG Member		SHG Leade		t-test
Variable	N	Mean	N	Mean	p-value
SHG support index	338	-0.066	173	0.129	0.037*
		(1.023)		(0.943)	
Entrepreneurial confidence	338	-0.065	173	0.127	0.040*
		(1.037)		(0.913)	
Market linkage index	338	-0.055	173	0.107	0.084
		(1.008)		(0.978)	

Notes: Standard deviations are in parentheses. The p-values are from two-sided t-tests comparing means between control and treatment groups. * p<0.05, ** p<0.01, *** p<0.001. † Is a SHG president, secretary, or treasurer.

The imbalances in capacity measures may be partly related to the SHG leadership imbalance documented in Section 2.1. As shown in Table 2.3, SHG leaders score higher than non-leaders on both the SHG support index (0.129 vs. -0.066, $p = 0.037$) and entrepreneurial confidence index (0.127 vs. -0.065, $p = 0.040$). The market linkage index shows a similar directional pattern, though the difference is not statistically significant (0.107 vs. -0.055, $p = 0.084$). Given that near-PLDs are substantially more likely to hold leadership positions than PLDs (48% vs. 30%), whether SHG leadership status contributes to the observed capacity differences between the two groups is worth investigating, though the extent to which it accounts for these differences cannot be determined from the available data.

Notably, despite these differences in capacity measures, actual economic outcomes show no baseline difference: there is no statistically significant difference in enterprise income and business profit between PLDs and near-PLDs with active enterprises.

[2]This measures historical outreach before IA's intervention.

3. Prospective Evaluation Approach: Matching + Difference-in-Differences

3.1 Motivation

Two natural evaluation approaches — simple pre- and post- intervention comparison between the treatment and control group and regression discontinuity design (RDD) are not viable for this project. First, a simple comparison of post-intervention outcomes between PLDs (treated) and near-PLDs is not feasible because the treatment effect would be compounded by the pre-existing differences between these groups. As shown in Table 2.1, PLDs and near-PLDs differ significantly at baseline on SHG leadership (30.3% vs. 47.6%, $p < 0.001$), entrepreneurial confidence (-0.060 vs. 0.231, $p < 0.01$), and market linkage (-0.080 vs. 0.309, $p < 0.001$). Such differences may be driven by the sampling process where PLDs were pre-identified by the government based on a composite income-related criteria. Second, RDD is not viable either, as the PLD threshold cannot serve as a running variable for RDD. As mentioned before, PLDs were pre-selected by the government based on their income status and potential in a composite way which we don't have clear or concrete information on. Even if such data is available, government discretion involved in the final PLD designation would create imperfect compliance that violates RDD.

Another question would be: why cannot we use being PLD as an instrument for actual scheme uptake to estimate the causal effect of schemes themselves, separate from other intervention components? This instrument variable (IV) approach is not viable because its key assumption – the exclusion restriction that the instrument (PLD status) should impact the outcome(s) only through the endogenous treatment variable (scheme uptake) – is violated in the projects' setting. On the one hand, being a PLD can influence a woman's income via multiple pathways not just through scheme uptake, as IA's intervention package not only promote scheme awareness and application but also provide help on market linkage and financial literacy. On the other hand, a woman pre-identified by the government as having the potential to reach 1L annual income is likely to have characteristics that may generate higher income even without scheme uptake.

3.2 Matching + DiD

Given all the context and constraint described above, we propose a two-step identification strategy combining matching with DiD, which can remove the selection bias on observables and unobservables and isolate the effects of IA's intervention.

Firstly, matching can create balance on those observable characteristics measured in the baseline survey – including SHG status, baseline economic activity, entrepreneurial capacity, networks, etc – by comparing PLDs to near-PLDs who look similar on those characteristics.

After matching, DiD can be applied to differentiate out the time-invariant differences between the matched groups. It can address the unobserved features that cannot be not fully captured by the survey measures, such intrinsic entrepreneurial talent, unobserved social capitals, time-varying resilience, etc, by comparing changes over time rather than levels. The DiD estimator is calculated as: Δ_{PLDs} - $\Delta_{\text{matched near-PLDs}}$. In practice, DiD is implemented via regression:

$$Y_{it} = \beta_0 + \beta_1 (PLD_i \times \text{Post}_t) + \alpha_i + \gamma_t + \varepsilon_{it} \tag{3.1}$$

where β_1 captures the treatment effectacross all post-intervention periods, α_i denotes individual fixed effects that absorb all time-invariant unobservables, and γ_t denotes time fixed effects that absorb common trends affecting both groups in each wave.. The key identifying assumption is parallel trend, which will be discussed in detail in later subsections. It means, absent IA's intervention package, PLDs and matched near-PLDs would have experienced similar changes in outcomes over time. Under this assumption, the control group's trajectory represents a valid counterfactual for what would have happened to PLDs without the intervention.

3.3 Estimand: Conditional Averager Treatment Effect

By combining matching with DiD, the estimand will be the conditional average treatment effect (CATE) of IA's intervention package on the economic outcome on PLDs. It answers the question "What is the impact of targeting women as PLDs and providing them with IA's scheme support package, compared to similar women who do not receive such support?" Importantly, it does not measure the effect of receiving specific schemes on those who received them or the causal effect of schemes themselves. These questions can be explored through heterogeneity analysis by scheme types received, livelihood categories, or baseline capacity, which can provide insight into which women benefit most and through which mechanisms. But such analyses should be interpreted as descriptive rather than causal given the endogeneity of scheme uptake.

3.4 Possible Matching Techniques

We recommend that IA test three matching approaches — inverse probability weighting, Mahalanobis-distance nearest-neighbour matching, and coarsened exact matching — and compare results across all three. All three approaches will be applied to the same set of baseline covariates. For each

approach, a post-matching balance check is conducted using metrics like standardised mean differences (SMDs), with a pre-defined threshold for acceptable balance (usually 0.10 if SMD). The five imbalanced variables identified at baseline — age, SHG leadership, SHG Support Index, Entrepreneurial Confidence Index, and Market Linkage Index — are the primary benchmarks for this check. IA should then compare the treatment effect estimates across the three approaches: consistent estimates strengthen confidence in the causal claims, while meaningful divergence would warrant investigation into whether residual imbalance in specific covariates is driving the difference. The approach that achieves the best post-matching balance and yields the most stable treatment effect estimate should be considered as the primary specification for the DiD analysis.

Inverse Propensity Weighting (IPW)

The first technique we suggest IA to test is inverse probability weighting (IPW), a propensity score-based method that addresses the selection bias arising from the government's non-random identification of PLDs by creating a pseudo-population to remove the measured baseline differences between PLDs and near-PLDs, making the comparison between the two groups as clean as possible before estimating the treatment effect.

It usually takes four steps. First is to estimate each woman's probability of being a PLD — her propensity score — based on their baseline characteristics, such as district and block, livelihood type, household size and assets, SHG leadership status, baseline scheme participation, baseline entrepreneurial confidence, baseline market linkage, and baseline SHG support. Next, weights are calculated as the inverse of the propensity score for PLDs and the inverse of one minus the propensity score for near-PLDs, so that women who look similar to the opposite group receive higher weights and exert greater influence on the comparison. Then, these weights are applied directly to the DiD regression — rather than constructing a matched sample — creating a weighted pseudo-population in which the observable differences between PLDs and near-PLDs should reflect treatment status alone. Finally, a post-IPW balance inspection is conducted using metrics like standardized mean differences (SMDs). In general, a SMD smaller than 0.1 indicates good balance.

Compared to the matching approaches described below, IPW retains all observations rather than discarding unmatched units, which can make more efficient use of the sample given the control group size of 105. However, IPW tends to be more variance-prone, especially in small samples, and is sensitive to poor overlap. That is, if some women have propensity scores very close to 0 or 1, which indicates that the two groups are very different on certain characteristics, then the resulting extreme weights can inflate variance and destabilise the estimates.

When implementing IPW, IA will need to pay attention to two aspects. First is whether to use stabilised weights— calculated by replacing the numerator with the marginal probability of PLD status rather than 1 — to keep weights within a more reasonable range and reduce variance. Second, IA should inspect the distribution of weights after estimation. If extreme weights are present, then truncation at the 1st and 99th percentiles may be needed.

Nearest-Neighbour Matching with Mahalanobis-Distance

The idea behind Nearest-neighbour matching (NNM) is to assigns each treated unit to one or more control units with the smallest distance in covariate space. The distance can be measured via different approaches. We recommend using the Mahalanobis distance, which a multivariate distance metric that accounts for the scale and correlation of covariates, given the mix of index variables, binary indicators, and continuous measures in the covariate set. The process is as follows: each PLD is matched to the nearest near-PLD(s) in covariate space using the Mahalanobis distance, with replacement, to accommodate the approximately 3.8:1 treatment-to-control ratio. After matching, the matched sample becomes the analytic dataset for the DiD regression.

Compared to IPW, NNM with Mahalanobis distance does not require a propensity score model, avoiding the additional layer of model dependence introduced by first estimating treatment probabilities. However, it is more sensitive to the choice of tuning parameters — namely, the number of matches and the caliper — and may discard more observations than IPW, which retains the full sample through weighting. Increasing the number of matches per treated unit can reduce variance but may also introduce bias from lower-quality matches. IA can start with a number of 4 as a reasonable default, test sensitivity by also running from 1 to 8, and then select the specification that achieves the best post-matching balance. Caliper defines the maximum Mahalanobis distance beyond which a potential match is rejected, preventing poor-quality matches from entering the analytic sample. To choose an appropriate caliper, IA should run the matching without a caliper first, inspect the post-matching SMDs for the five imbalanced baseline variables, and tighten the caliper iteratively until all SMDs fall below 0.10.

Coarsened Exact Matching (CEM)

Coarsened exact matching (CEM) is a monotonic imbalance bounding technique which bounds the maximum imbalance between the treated and control groups through ex ante choice. The approach works in three steps: (1) temporarily coarsened covariates into substantively meaningful bins, (2) exactly matched each PLD to near-PLDs who fall within the same bins, and (3) discard any PLD or near-PLD for whom no match exists from the analytic sample. As the coarsening becomes finer, i.e., the bins become more narrow, the bound on the maximum imbalance becomes tighter.

Once matching is complete, CEM returns a set of weights that are carried forward into the DiD regression.

To implement CEM, IA will need to specify cutpoints that define the coarsening for each covariate. That is, IA need to ensure the bins that are substantively meaningful given the context of the study rather than relying solely on automated binning algorithms.

CEM is particularly suited to this study as it can balance the nonlinearities and interactions between covariates, which is important given the complex relationships between SHG participation, livelihood type, and income in this context.

3.5 Statistical Power and Minimum Detectable Effect

Before midline data collection, we recommend IA to conduct a power analysis to determine the minimum detectable effect (MDE) — the smallest effect size the study can reliably identify — and confirm the sample is adequately sized to detect economically meaningful changes in the primary outcome(s). Power analysis is particularly important in this study for two reasons. First, the control group is small — 105 near-PLDs at baseline — and matching will reduce this further by discarding unmatched units or downweighting poorly matched ones depending on the technique chosen. Second, the primary outcome — enterprise income and/or household income — has high variance.

We recommend that IA conduct a formal power simulation prior to midline data collection, in three steps. First, using the baseline dataset, apply each of the three matching approaches and record the number of treatment and control observations retained in each matched sample. Second, using the baseline distribution of the primary outcome variable — particularly its standard deviation — simulate the MDE for each matched sample under the planned DiD specification, varying assumptions about within-person income correlation across waves to test sensitivity. Third, compare the MDEs across the three matching approaches and assess whether they correspond to effect sizes that are economically meaningful and realistic given the intervention's scope. This comparison will be one input — alongside post-matching balance — into IA's eventual selection of a primary matching specification.

4. Key Assumptions / Conditions

As briefly touched on in prior sections, making rigorous, credible causal inference from the proposed matching plus DiD approach requires some core assumptions to hold. Understanding these assumptions is helpful for Indus Action to assess whether the proposed evaluation strategy is worth pursuing and to identify threats that could undermine causal interpretation. So, this subsection outlines those assumptions, assesses their plausibility in this project's context, and discusses how they may be examined contingent to current and future data collection.

4.1 (Conditional) Parallel Trend

DiD primarily requires the conditional parallel trends to hold. That is, in the absence of IA's intervention package, the PLDs and matched near-PLDs would have experienced similar changes in outcomes over time. This assumption is plausible in this project context for several reasons. On the one hand, near-PLDs were selected for similar income potential as PLDs, creating comparable economic trajectories. On the other hand, matching creates balance on observable factors of differential trends.

However, the parallel trend assumption could be violated in the following scenarios: (1) PLDs are targeted by other (government) programs that may impact their income and enterprise activity during the intervention period, or (2) PLDs possess systematically different income growth trajectories, compared to similar non-PLD women, based on observed characteristics.

The parallel trends assumption cannot be directly tested with a single baseline period. However, since IA plans to conduct multiple post-intervention midlines quarterly , they can examine treatment effect dynamics through an event study analysis — plotting the DiD estimates at each post-intervention wave separately rather than collapsing them into a single average effect. Consistent, gradually increasing effects across waves would lend credibility to the assumption that pre-intervention trends were parallel.

4.2 Spillover Effect

Both matching and DiD require the Stable Unit Treatment Value Assumption (SUTVA) to hold. That is, IA's intervention on PLDs should not influence the near-PLDs. This assumption could be violated if PLDs and near-PLDs have some interaction that creates spillover. For instance, PLDs

and near-PLDs in the same SHGs may share information about schemes and application processes. Besides, if PLDs and near-PLDs operate similar businesses in the same/close local markets, then the growth in PLDs' business may create competition effects that decrease the income of near-PLDs, or conversely, create spillover demand that benefits them.

The severity of spillovers can be assessed by, for example, keeping track of whether women share scheme information with other SHG members and whether they perceive increased business competition and comparing the estimated effects in high-treatment-concentration SHGs versus low-concentration SHGs.

4.3 Attrition

DiD requires that the same individuals are observed across all survey waves. If respondents drop out of the panel and attrition is random, then the loss of observations reduces statistical power but does not bias the treatment effect estimate. However, if attrition is differential — if PLDs and near-PLDs drop out for systematically different reasons — then the composition of the two groups will change over time in a way that cannot be captured by the baseline matching, which will risk biasing the DiD estimator.

IA should monitor attrition rates separately for PLDs and near-PLDs at each wave, and test whether attritors and non-attritors differ systematically on baseline characteristics. If differential attrition is detected, inverse probability weighting on survey participation (separate from the matching weights) or Lee bounds can be used to bound the treatment effect estimate under different assumptions about the missing data.

5. Key Elements to Collect in Future Rounds

If IA decides to implement the proposed matching plus DiD approach, making rigorous causal inference from such approach requires the collection of certain data and information in future survey rounds. This section discusses the necessary data collection from three aspects: outcome (to ensure high-quality impact measurement) , intervention delivery and uptake (to examine the actual implementation status), and panel maintenance (to ensure validity of analysis).

Outcome

Once priority outcomes are finalized through the Theory of Change process described in Section 6.2, they should be measured consistently across all survey waves using the same questions and recall periods as at baseline. Measurement consistency does not require exactly identical wording if baseline questions were problematic; rather, it requires capturing the same underlying construct in a comparable way. Detailed recommendations on survey instrument refinement — including addressing data quality challenges identified at baseline — are discussed in Report I and should be implemented before midline data collection.

Intervention Delivery and Uptake

In the survey waves post-intervention, IA needs to incorporate questions to examine the actual implementation status of the intervention package and to decompose its effect mechanism. Given the actual intervention design and implementation, some example questions could be: Have you received IVRS/SMS messages from Indus Action about government schemes? If yes, when and regarding which schemes? Have you attended an IA-organized scheme application camp? If yes, how many times, which block, when, and for which schemes? Have you received support for scheme applications from Citizen Service centers (CSC)? Not only can this question directly capture the actual outreach and uptake of IA's intervention package, but it can also enable analysis of which intervention components drive effects and help distinguish between that intervention doesn't work versus that women don't take up the intervention. After collecting such data, IA can compare it with the intervention delivery record to identify the caveat during implementation and address them as the intervention progresses.

Future survey rounds also need to incorporate questions to track actual scheme uptake after the intervention rollout. Some examples could be: Which government schemes did you apply for from baseline? For each scheme applied: Were you approved? If yes, what benefits did you

receive and when? If no, why were you rejected? For schemes you didn't apply for despite being informed of eligibility, what are the barriers that prevented your application? If possible, IA can obtain government administrative data on scheme applications and approval to verify self-reported scheme receipt and analyze approval rates and reasons for rejections.

Panel Maintenance

To ensure the validity of DiD, the same individuals at baseline should be surveyed in the subsequent midlines and endline. At each follow-up wave, enumerators need to re-interview all 511 baseline respondents, update contact information, document (if any) reasons for non-response distinguishing between refusal, migration, illness, and unreachable cases so that IA can assess whether attrition is random or differential across PLDs and near-PLDs. To minimise drop-out, IA should consider providing advance notice of survey timing and small participation incentives, calibrated carefully so as not to induce differential retention across groups. If attrition still occurs, approaches like inverse probability weighting or Lee bounds may be needed to address selection bias from differential attrition.

In addition, the follow-up surveys need to document the related contextual factors that may violate the parallel trend assumption. For instance, at the aggregate (state/district/block) level, if there is any socio-economic incident or policy changes that may affect livelihoods and economic growth, such as introduction of new agricultural policies, market disruptions, black swan events like COVID-19, etc. At the individual level, IA should pay attention to any pronounced patterns/changes in, for example, if respondents participate in other government programs beyond the 10 priority schemes (name of program, benefits received, timing), their SHG participation status (still active member, leadership role changes, SHG dissolved), their household composition. This information can facilitate the assessment of whether or not external factors impact PLDs and near-PLDs differently. If yes, they can be incorporated as time-varying controls in the DiD regression if needed.

6. Discussion

6.1 Limitations on Causal Inference

With appropriate data collection and implementation, the matching plus difference-in-differences strategy outlined in Section 5 can isolate and estimate the causal impact of Indus Action's intervention package on PLDs' economic outcomes. However, it cannot credibly isolate the causal effect of the schemes themselves or any specific intervention components. Given the compound setting of this project, outcome differences by scheme uptake can only be explored through heterogeneity analysis, which is descriptive not causal. This limitation is structural, not methodological, as the actual scheme uptake is endogenous.

6.2 Key Recommendations Moving Forward

To maximize the credibility and policy relevance of the impact evaluation, we recommend three strategic actions before midline data collection.

Finalize Theory of Change

A clearly articulated Theory of Change is a prerequisite for rigorous impact measurement. Concretely, we recommend three actions under this heading. First, research questions should be refined and anchored around the core causal question the study design is well-positioned to answer: what is the impact of being identified as a PLD and receiving IA's scheme support package, compared to similar women who did not receive such support?

Second, priority outcomes should be pre-specified before midline data collection. We suggest treating scheme access — the extent to which exposure to the intervention is associated with higher uptake of target HRLM schemes — and income outcomes — whether households exposed to the intervention exhibit differences in income from ongoing livelihood activities — as the two primary estimable outcomes. Secondary outcomes including entrepreneurial capability, women's agency, resilience to shocks, savings, and debt can capture mechanism and spillover effects.

Third, the channel of change needs to be explicitly defined. The study's sample spans diverse livelihood categories, yet its outcome framing is primarily in terms of entrepreneurship. Clarifying whether income growth is expected to operate through enterprise specifically, or through broader pathways such as daily wages, scheme transfers, or asset accumulation, is essential for both mea-

surement and attribution. Once finalized, outcomes should be pre-registered in a pre-analysis plan (PAP) before any midline data is collected, to avoid post hoc outcome selection.

Request Documentation of PLD Selection Criteria

If possible, IA should request documentation from the Haryana government on how PLDs were identified, ideally including specific thresholds, data sources, and the extent to which the process was formulaic versus discretionary. This matters for two reasons under the proposed evaluation strategy. First, any variables used in the selection process are likely confounders and should be incorporated into the matching covariate set to reduce residual imbalance. Second, understanding whether the selection criteria systematically favor women on particular income trajectories is important for assessing the plausibility of the parallel trends assumption.