

# Pre-Analysis Plan: Data from Observational Tool

## Refugee and Host Integration through the Safety Net: Evidence from Ethiopia<sup>1</sup> Effects of Team Diversity on Productivity and Prosocial Behaviors

### Authors and institutional affiliation:

Dennis Egger, Oxford University

Alfredo Manfredini Böhm, Tony Blair Institute for Global Change

Sandra V. Rozo, World Bank, Development Research Group

Thierry Hounsa, Oxford University

**Date of latest draft:** 23/06/2025

### Abstract

This study explores how team diversity impacts productivity, prosociality, and social cohesion in Ethiopia's Refugee and Host Integration through Safety Net program, a public works and livelihood program that integrated refugees into the national safety net program. Using a randomized control trial we identify the causal effects of i) intergroup contact between hosts and refugees and ii) equal vs. minority group composition, on a rich set of outcomes including unannounced observations and incentivized games on prosocial interactions and team dynamics during weekly public works activities. By capturing both revealed and stated preferences, we provide nuanced insights into how diversity and team-based interventions affect group functioning in low-resource, high-stakes environments.

**Keywords:** Forced Migration, Social Protection, Public Works, Intergroup Contact, Social Cohesion, Productivity

**JEL codes:** H4, J61, O1

---

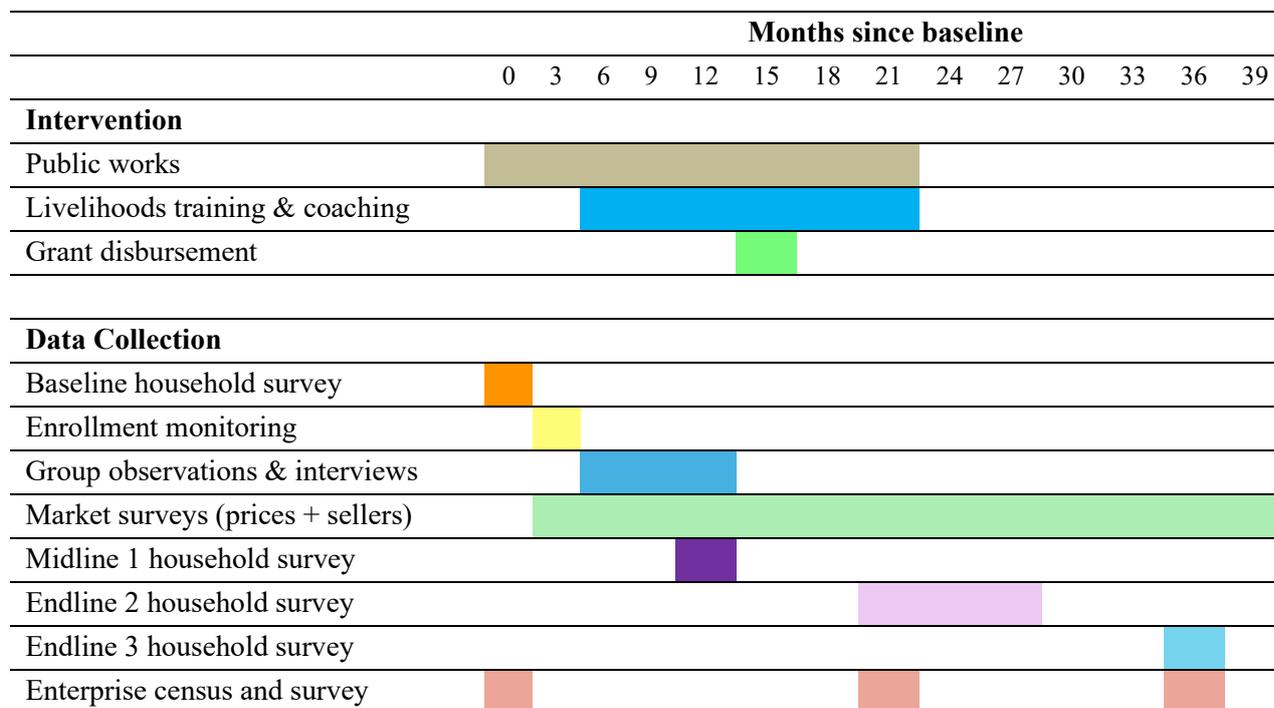
<sup>1</sup> We are grateful to Belay Mohammed Yimer for his support in preparing this document and as research assistants in this project. **Study pre-registration:** AER RCT Registry Number AEARCTR-0012288.

## Proposed timeline and relation to other analyses.

This document complements the midline PAP (Final Pre Analysis Plan Midline Data), which focuses on household-level outcomes, by providing granular group-level insights from unannounced observations and lab-in-the-field games during public works activities, and accompanying interviews. Particularly, this pre-analysis plan covers the data collected during stage one of a research agenda planned for two stages described below.

1. **STAGE 1 - Public works and training impact evaluation:** In the first 12 months, beneficiary households will be assigned to homogenous or mixed refugee-host groups and will start participating in remunerated public works, attending initial training and coaching in financial management and life-skills.
2. **STAGE 2 - Livelihood development impact evaluation:** Between months 12-24th, beneficiaries continue to attend training and coaching sessions, focusing increasingly on developing a business plan to improve their livelihood. Based on their business plans, beneficiaries receive a USD 600 livelihood grant, disbursed between the 18<sup>th</sup> and 24<sup>th</sup> month.

**Figure 1. Timeline of the RHISN Program**



## Analyses and progress to date

To date, no principal investigator has accessed the data. All data handling has been limited to quality control procedures conducted by research assistants and the survey firm (blinded to treatment). These procedures did not involve any analysis or estimation of treatment effects, nor any regressions or descriptive statistics stratified by treatment status.

## 1. INTRODUCTION

**Motivation.** Forced displacement, affecting over 120 million people globally (UNHCR, 2024), poses a critical challenge for integrating refugees into host communities, particularly in low-income countries. Successful integration requires fostering social cohesion and prosocial behaviors between refugees and hosts. Public works programs, where individuals work side by side, offer a unique platform for promoting such cohesion through everyday intergroup contact. However, the effects of workplace diversity on both social relations and productivity remain poorly understood. While mixing refugees and hosts may strengthen social bonds, it can also generate tensions or reduce efficiency if coordination breaks down. Understanding when and how diversity in public work teams promotes social cohesion *and* performance is essential to designing integration strategies that are both socially and economically effective.

**Study Focus:** This Pre-Analysis Plan focuses on rich observational and behavioral data collected during public works (PW) activities as part of the Refugee and Host Integration through the Safety Net program (RHISN), and accompanying interviews. RHISN is a pioneering initiative by the Ethiopian government to integrate camp-based refugees into the national social safety net. The data provide a unique opportunity to examine how team diversity and inter-group contact influence real-time measures of productivity, prosocial behaviors, and social cohesion. These data enable causal analysis of how randomized team assignments (mixed vs. homogeneous refugee-host groups) shape economic and social outcomes. By observing how individuals interact, how tasks are assigned and performed, and how beneficiaries perceive and engage with both in- and out-group members, we can uncover the mechanisms through which social protection programs may foster (or hinder) inclusion, efficiency, and group solidarity.

**Research Questions.** This study leverages the randomized design of the RHISN program to examine how team diversity and intergroup contact shape outcomes in public works settings. We address two primary questions: (1) How does random assignment to mixed (refugee-host) versus homogeneous teams affect productivity? (2) Does intergroup contact in mixed teams enhance prosocial behaviors and social cohesion? By addressing these questions, the study aims to deepen understanding of how diversity operates in low-income, high-stakes environments, and to inform the design of social protection programs that can both enhance livelihoods and foster social integration.

**Randomization.** To identify the causal effects of team diversity and intergroup contact on productivity and social cohesion, we leverage the experimental design of the RHISN program, which involved the random assignment of individuals into public works groups. Individuals were first assigned either to homogeneous teams (composed entirely of refugees or entirely of hosts) or to mixed teams, where both groups worked together under identical program conditions. Within the set of mixed teams, assignment was further randomized along the relative composition of refugees and hosts (equal groups of minority refugee groups). Some groups were formed with an equal share of refugees and hosts, while others were constructed such that refugees were in minority. This two-tiered randomization enables us to isolate both the direct effects of intergroup contact and the heterogeneous effects of group composition on outcomes of interest.

The observations survey serves as the main measurement tool for this study. It captures both observed behaviors (through structured enumerator assessments of group and individual productivity, interactions, team dynamics, and productivity) and elicited behaviors (through incentivized games measuring cooperation, productivity, and altruism). In addition, we conducted individual interviews with group members, eliciting attitudes as well as their individual evaluation of team dynamics, and of

their team members. These data allow us to examine both outcome differences across team types and the mechanisms through which team diversity may influence social and performance dynamics on the ground.

***Contributions relative to existing work.*** This study contributes to two important strands of literature. First, we build on previous work exploring how intergroup contact can reduce prejudice and foster inclusion (Allport 1954; Rao 2019; Scacco and Warren 2018; Mousa 2020; Lowe 2021). While much of this literature has focused on interactions in educational or sports-based contexts, our study examines co-working relationships between refugees and host community members in public work teams. The RHISN program's scale and randomized design allow us to causally identify how intense contact during collective labor affects trust, cooperation, and informal group dynamics. Moreover, by varying the intensity of exposure we can assess not only whether contact improves cohesion, but how its effects depend on contact intensity.

Second, this study contributes to the limited but growing body of work on how ethnic or social diversity affects productivity in teams. While much of the evidence comes from lab or classroom experiments, Hjort (2014) provides rare causal evidence in a labor market setting by examining the productivity effects of tribal diversity in Kenyan flower firms. Our study extends this line of inquiry by focusing on public works teams in Ethiopia, with a novel focus on how diversity shapes performance, task assignment, cooperation, communication, leadership emergence, and peer evaluation in real-world environments. The unique structure of the RHISN observations data (including enumerator observations, bilateral rankings, and behavioral games) enables us to uncover the mechanisms behind diversity's impact on both performance and social integration. Additionally, our design allows us to separately study the role of on-the-job contact versus after-work interactions, which remain underexplored in the literature.

Together, these contributions provide new evidence on how forced displacement and social protection intersect with workplace diversity, offering insights relevant to both academic debates on intergroup relations and practical efforts to design inclusive public employment programs.

***Policy relevance.*** The RHISN program is among the first large-scale initiatives to bring refugees and host communities together in shared public works projects. While the program aims to support livelihoods and build resilience, it also serves as a platform for fostering social cohesion through structured intergroup contact.

This study provides actionable evidence for governments and development partners on whether (and under what conditions) diverse teams can work effectively and harmoniously. By rigorously evaluating the impact of team composition on both productivity and social relations, we inform policies about group formation in public employment schemes. Importantly, we assess not only whether diversity works, but also how specific design features (such as the intensity of outgroup contact) shape outcomes.

## 2. DATA: THE OBSERVATIONAL SURVEY

**Sample size:** The Observation data cover a total of 133 public works (PW) groups, all located in Dabat, and comprising both refugees from Dabat/Alemwach refugee camp as well as host community members living in the city. On average, each group includes approximately 27 beneficiaries, resulting in an overall sample of more than 3,500 individuals.

Group visits were unannounced, so that public work groups and beneficiaries would not know *ex-ante* when the observations surveys would be conducted. Enumerators first collected information on group composition to assess compliance, before conducting the observations survey.

The observations survey collected both quantitative and qualitative data, in the form of extended qualitative field notes. This PAP primarily concerns the quantitative data collected. Qualitative field notes will be digitized and analyzed separately, using appropriate qualitative methods.

**The observations survey:** The observations survey comprises three key components. First is the structured observation of a public work session that is conducted in a natural setting by a team of two enumerators. One enumerator uses a structured quantitative tool to assess group- and individual-level productivity, member interactions, subgroup dynamics, and related factors. The second enumerator captures qualitative insights, focusing on aspects such as task engagement, motivation, teamwork, interpersonal interactions.

The second component involves three behavioural ‘lab-in-the-field’ games: i) the Engineering Challenge, which assesses productivity and cooperation, ii) the Deer–Partridge Game, which evaluates trust and social cohesion, and iii) the Dictator Game, which assesses altruistic preferences within and across teams. Each game is facilitated by a pair of enumerators: one leads the activity and takes qualitative notes, while the other completes a structured quantitative questionnaire on SurveyCTO.

Finally, the third component consists of individual-level interviews with both a group leader and all team members. This survey collects individual perceptions of randomly selected teammates’ and the overall group’s productivity, interaction, trust, empathy, motivation, etc.

For each public work group, the survey is conducted over a three-day period. Prior to the first day of observation, enumerators meet with the group leader to verify the membership list against the original official assignment. This short exercise aimed to update the member list to reflect the current active members by removing dropouts and adding any new ones. The updated list is then pre-loaded into SurveyCTO so that enumerators have up-to-date group membership information during the observation and other data collection activities. The second and third days correspond to the data collection.

Overall, across its three-day structure, the Observations Survey combined objective observation, randomized assessment, and behavioral experimentation to provide an overview of group functioning in public works settings. Day 1 focused on verifying group structure and capturing session-wide dynamics through direct observations. Days 2 and 3 introduced randomized individual productivity measurement and behavioral games designed to elicit social preferences and group interaction styles. Enumerators played a central role throughout, not only recording factual data but also interpreting group atmosphere, cohesion, and motivation through systematic observation. The resulting dataset offers an unusually rich foundation for analyzing both the technical and social aspects of public works implementation.

## **First Day: Initial Observations, Group Dynamics and Randomized Observations**

The first day of observations focused on capturing information on group composition and task environment. Enumerators arrived at the PW sites, and recorded metadata including their arrival time (relative to session start), group ID, and GPS coordinates.

Enumerators then verify group composition by comparing actual participants to official lists. This process allows the team to record deviations such as absenteeism, substitutions, or unexpected participants. To enable structured observation of team behavior and interactions by enumerators, team members were then issued numbered jerseys (to be worn over their work uniforms), and the correspondence between numbers and group members was recorded during the initial roll-call. Numbers assigned remained constant across the three days, and were verified each day.

The tasks observed varied by PW groups; common activities included tree planting, and streets cleaning. Enumerators document the main task, tools in use, and give a subjective rating (measured through a Likert scale) of productivity, effort level, and group organization.

Enumerators also assessed team dynamics. They rated how well teams worked together, noted signs of conflict or cooperation, and provided structured observations of motivation, leadership, and social interactions within the group. These assessments were complemented by open-ended notes, enabling richer interpretation of group behavior.

Next, we introduce a more structured and randomized approach to observation at the individual and group levels. From each group, enumerators randomly select a subset of beneficiaries. Each selected beneficiary is observed while performing a task over a four-minute timed window. Then the enumerator records objective measures such as: number of trees planted, number of wheelbarrows filled, and square meters of street cleaned. In addition to physical outputs, enumerators noted how individuals worked, whether they showed leadership, interacted with others, appeared motivated, or were idle. Enumerators also capture group interactions: who beneficiaries talked to, helped, ignored, or socialized with.

## **Second Day: Behavioral Games and Structured Interaction**

We measure behavioral preferences and social interactions through a series of carefully designed games. These games were administered in small sub-group settings and aimed to elicit cooperation, trust, leadership, and altruism, under controlled, incentivized conditions. The main games included:

- **The Engineering Game:** For this task, each public works team was subdivided randomly into a maximum of 5 subgroups/teams. A collaborative task where groups were instructed to build a tower using simple materials. Enumerators recorded whether the tower fell, its final height, and the team's ranking. They observed the role distribution within the team, i.e., who took leadership, who supported others, who remained passive, and who dominated conversations. These interactions offered a behavioral insight into latent group norms, such as cooperation and conflict resolution. The engineering game was incentivized by monetary rewards based on team performance. The team ranked first received 200 ETB, while the second-place team earned 175 ETB. Rewards then decreased in increments of 25 birr for each subsequent team, down to the last-ranked team. On average, the game lasted about 90 minutes.
- **Deer–Partridge Game (a stag hunt variation):** For this game, group members were paired at random with their teammates. Each participant played a series of rounds with different partners. The game modeled coordination under risk: participants could choose a safe, individual reward (partridge) or a riskier, high-reward option requiring mutual cooperation (deer). Each

participant played six rounds in total—three rounds under anonymous matching (partners not revealed) and three rounds under known-partner matching (partners revealed but no communication allowed), with the order of these two conditions randomized across groups. Incentives were tied directly to outcomes: a deer successfully hunted by both partners earned 4 days of food for each player, a partridge alone yielded 2 days, and mutual partridge yielded 1 day each; mismatched choices gave the partridge hunter 2 days and left the deer hunter with 0. At the end of the game, accumulated days of food were converted into cash payments, with each day worth a monetary amount. The entire exercise lasted around two hours.

- **Dictator Game:** This game was administered as part of the individual-level survey and conducted with the group leader and beneficiaries. In this game, respondents were given an endowment of 25 ETB (0.17 USD) and asked how much they would share with randomly selected group members (sometimes an in-group member, sometimes from an out-group). This will allow us to assess altruistic preferences within and across teams.

### Third Day: Individual Interviews

On the third day, the field teams conducted detailed individual interviews with selected beneficiaries across the sampled public works groups. The purpose of this activity was to collect information on individuals' perceptions of productivity and social cohesion. Within these interviews, enumerators also gathered data on bilateral tie strength and interpersonal assessments. Respondents were asked about eight randomly chosen teammates (balanced across in-group and out-group members) and provided information on whether they socialize outside work and whether they frequently collaborate at the worksite. The tool also incorporated self-assessments and peer-assessments of productivity and effort, measured through Likert scales and ranking exercises, and empathy and trust scales.

The interviews captured broad indicators of social cohesion and trust. Beneficiaries were asked whether they would want to share tea or coffee with selected peers, whether they would trust them to look after their house, or whether they would accept them as future group leaders. These questions were complemented by modules on subgroup dynamics (whether the larger team splits into smaller units, and how those subgroups are composed), interaction types (e.g., chatting, singing, quarrelling, or working together outside the programme), and frequency of meeting group members beyond public works (at home, religious institutions, or community events).

## 3. EMPIRICAL STRATEGY

### 3.1. Assessing Impacts of Team Diversity: Mixing and Non-Mixing

**Individual level data.** We will employ the following specification to estimate the effects of team diversity in public works on individual productivity, prosociality, and social cohesion:

$$y_{ig} = \beta_0 + \beta_1 \text{MixedGroup}_{g(i)} + \beta_2 \text{ShareMixedNeighbors}_{c(i)} + X_i' \Lambda_2 + \epsilon_{ig} \quad (1)$$

$y_{ig}$  is a beneficiary level outcome for beneficiary  $i$  working in public group  $g$ . Individuals live in neighborhoods/clusters, denoted by  $c$  (ketenas for hosts, and camp block for refugees).  $\text{MixedGroup}_g$  is a binary variable taking the value of one if the beneficiary was (randomly) assigned to a mixed group of refugees and hosts.  $\text{ShareMixedNeighbors}_{c(i)}$  is the share of RHISN eligible households living in

$i$ 's residential neighborhood working in a mixed public works group.<sup>2</sup>  $X_i'$  is a dichotomous variable for refugee/hosts).<sup>3</sup> <sup>4</sup>

$\beta_1$  identifies the Intent-to-Treat (ITT) impact of mixing refugees and hosts into a public work group on the outcomes of interest. Equation (1) will be estimated for the whole sample and for host and refugees samples separately. This will enable us to compare the effects of mixing relative to only refugee and only host homogenous groups. As treatment assignment to mixed groups was conducted at the individual level, we will use robust non-clustered standard errors for our primary analysis.<sup>5</sup>

$\beta_2$  estimates the spillover impacts within residential neighborhoods of mixing. We consider this analysis *secondary*, and will estimate these effects overall, and separately for mixed and non-mixed beneficiaries. The share of RHISN eligible households working in a mixed public works group is determined exogenously by the cross-randomization of neighborhoods into high saturation  $HiSat_{c(i)}$  (a higher and lower share participating in RHISN), and high-mixing  $HiMix_{c(i)}$  (a higher or lower share of participants assigned to mixed or non-mixed groups. In exploratory analyses, we will replace  $ShareMixedNeighbors$  with these indicators, and their interaction to test for differential impacts of saturation and mixing. Since assignment to high-mixing and high-saturation was at the neighborhood level, we will cluster standard errors at the neighborhood level when assessing spillover impacts ( $\beta_2$ ).

**Group level data.** To estimate the impact of mixing refugees and hosts on group- or sub-group level outcomes, we will employ the following specification:

$$y_g = \beta_0 + \beta_1 MixedGroup_g + \epsilon_g \quad (2)$$

Where  $y_g$  is a group level outcome of interest and  $\epsilon_g$  are heteroskedasticity-robust standard errors.

**Team level data.** Since some outcomes related to the Engineering game were conducted in teams with members randomly assigned to each one, we will additionally explore whether random variation within teams (in addition to entire public work groups) affects outcomes measured as part of the Engineering Challenge, i.e. if the extent of mixing within sub-groups impacts outcomes over and above the extent of mixing within the public work group overall.

### 3.2 Assessing Impacts of Mixing on Minority and Equalized Groups

To test whether the effect of group mixing varies with the relative composition of refugees and hosts, we use two separate indicators: one capturing whether the workgroup consisted of an equal number of refugees and hosts (*Equal*), and another indicating whether refugees were in the minority (*Minority*). In both Equations (1) and (2), we will replace the indicator for  $MixedGroup_g$  with two indicators for minority and equal refugee shares:  $Minority_{g(i)}$ ,  $Equal_{g(i)}$ . All other terms remain the same.

---

<sup>2</sup> Groups are formed at the kebele level, a geographic level higher than the cluster, and assignment to groups is stratified by cluster, so that each group has the same share of members from each cluster.

<sup>3</sup> We will explore additional controls selected using Machine Learning (DDML and LASSO) techniques (see section below).

<sup>4</sup> We will explore the possibility of accounting for systematic subjective differences across enumerators in measuring beneficiaries' productivity by including enumerator fixed effects. This is only possible if the same enumerator observed both mixed and non-mixed groups, which was not always possible given language constraints. We therefore consider this analysis exploratory.

<sup>5</sup> Following recent work by Abadie et al. (2022), we will explore the robustness of our findings to clustering at the group-level.

### 3.3 Assessing the Impact of After-Work Socializing

30% of all public works groups (cross-randomized by whether the group was mixed, refugee-only, or host only) were randomly selected to participate in after-work social cohesion activities. Social cohesion activities involve after-work teas, lunch, and cultural festivals. By June 2025 – the end of our data collection – work teams had attended six tea sessions and one lunch. In secondary analysis, we estimate the additional impact of after-work socializing on our primary outcomes using the following specifications for individual and group-level outcomes:

$$y_{ig} = \beta_0 + \beta_1 \text{SocialCohesionActivity}_{g(i)} + X_i' \Lambda_2 + \epsilon_{ig} \quad (3a)$$

$$y_g = \beta_0 + \beta_1 \text{SocialCohesionActivity}_g + \epsilon_g \quad (3b)$$

where  $\text{SocialCohesionActivity}_{g(i)}$  is an indicator variable for being selected to participate in social cohesion activities. We cluster standard errors at the group-level to reflect the level of randomization of social cohesion activities. While our main specification pools all groups, we will also test whether social cohesion activities have different impacts for refugees and hosts, and for mixed and non-mixed groups using a fully interacted specification between social cohesion activities and mixing.

### 3.4 Peer effects, spillovers, and bilateral assessments

Groups vary not only in their composition of refugees and hosts, but also along other dimensions (e.g. gender, individual productivity, education, etc.). Due to random assignment, these differences, too, are exogenous. In exploratory analyses (potentially to be specified in an additional pre-analysis plan), we will further explore peer effects along other important dimensions following specifications and approaches in the peer effects literature. Additionally, we have rich information on within-group bilateral connections: We know whether they were relatives or friends at baseline from our baseline survey, and whether and how they interact both within and outside groups. The rich bilateral/dyadic structure of this data will open other avenues of enquiry, including spillovers within social networks, or dyadic regressions using similarity measures on dimensions other than refugee status, etc. At this stage, we consider these exploratory.

## 4. OUTCOMES

We attempt to estimate Equation (1) for all beneficiaries level outcomes and equation (2) for all group outcomes. For the outcomes for which it is not possible to estimate equations (1) or (2), we will specify the corresponding equation in front of the outcome. Additionally, for variables collected in Likert scales, we will present standardized values (using the control group mean and standard deviation) and for indexes we will re-code all variables so that positive outcomes take higher values. We will represent monetary values in USD PPP terms.

<b>Group 1: Productivity</b>			
<b>ID</b>	<b>Outcome Family</b>	<b>Unit of analysis</b>	<b>Variables within each family</b>
1.1	Perceived Individual Productivity	Beneficiaries and Group Leaders	<p><b>Primary outcome:</b> 1.1.2 Productivity (peer assessment of productivity)</p> <p><b>Secondary outcome:</b> Perceived productivity index: Anderson index of 1.1.1, 1.1.2, 1.1.3, 1.1.4, 1.1.5, 1.1.6.</p> <p><b>Index components:</b> We have likert scales for the following outcomes:  1.1.1 Productivity (self-assessment)  1.1.2 Productivity (assessment of PW peers)  1.1.3 Productivity (group leader assessment of PW peers)  1.1.4 Work ethic (self assessment)  1.1.5 Work ethic (peer assessment)  1.1.6 Work ethic (group leader assessment)</p> <p>We will then analyse whether productivity/work ethic assessments of PW peers by beneficiaries vary between in- and outgroup members in peers:  1.1.x.1 assessment of in-group peers (all beneficiaries)  1.1.x.2 assessment of out-group peers (sample: all beneficiaries/GL)  <i>For eqs. 1/2, use in-group outcome in non-mixed groups. For eq. 3, use out-group outcome everywhere.</i>  1.1.x.3 assessment of refugee peers (sample: refugees in refugee-only and mixed groups)  1.1.x.4 assessment of host peers (sample: hosts in host-only or mixed groups)</p>
1.2	Enumerator Observed Individual Productivity	Beneficiaries	<p><b>Primary outcome:</b> Standardized productivity index (Standardized variable to measure deviation from the non-mixed mean for each task using 1.2.1, 1.2.2 and 1.2.3)</p> <p><b>Secondary outcomes:</b>  1.2.1 Number of trees planted (Observation)  1.2.2 Number of wheelbarrows full of dirt shovelled (Observation)  1.2.3 Number of square meters cleaned (Observation)  1.2.4 Productivity (Enumerator observation)  1.2.5 Work ethic (Enumerator observation)</p>
1.3	Measured Performance in Engineering Game	Teams	<p><b>Primary Outcome:</b> 1.3.2 Engineering game (Tower height)</p> <p><b>Secondary outcomes:</b>  1.3.1 Engineering game (tower did not fall)  1.3.2 Engineering game (tower height)  1.3.3 Engineering game (share taking leadership)  1.3.4 Engineering game (share not participating or just watching)  1.3.5 Engineering game (share supporting the leader)  1.3.6 Engineering game (share working independently)  1.3.7 Engineering game (share actively speaking during game)  1.3.8 Engineering game (team's productivity - observation)  1.3.9 Engineering game (team working hard - observation)  1.3.10 Engineering game (team motivation - observation)  1.3.11 Engineering game (team's level of conflict - observation)</p>
1.4	Group Productivity	Public Work Groups	<p><b>Primary Outcome:</b> Standardized productivity index (Standardized variable to measure deviation from the mean for each task using 1.4.1-1.4.4)</p> <p><b>Secondary outcomes:</b>  1.4.1 Number of trees planted (Observation)  1.4.2 Number of wheelbarrows full of dirt shovelled (Observation)</p>

			1.4.3 Number of square meters cleaned (Observation) 1.4.4 Number of pieces scooped up (Observation) 1.4.5 Group Productivity (Observations, beneficiaries and GL assessment)
--	--	--	--

<b>Group 2: Prosocial Behaviors and Attitudes</b>			
<b>ID</b>	<b>Outcome Family</b>	<b>Unit</b>	<b>Variables within each family</b>
2.1	Altruism	Beneficiaries and Group Leaders (GL)	<p><b>Primary outcome:</b> 2.1.2 Dictator game: share of donations to other group members (incl. beneficiaries and GL)</p> <p><b>Secondary outcomes:</b></p> <p>2.1.1 Dictator game: indicator variable for any donations to other group members (sample: all beneficiaries and GL)</p> <p>2.1.1.1 Dictator game: indicator variable for any donations to in-group peers (sample: all beneficiaries/GL)</p> <p>2.1.1.2 Dictator game: indicator variable for any donations to out-group peers (sample: all beneficiaries/GL) <i>For eqs. 1/2, use in-group donations in non-mixed groups. For eq. 3, use out-group donations everywhere.</i></p> <p>2.1.1.3 Dictator game: indicator variable for any donations to refugee peers (sample: refugees in refugee-only and mixed groups only)</p> <p>2.1.1.4 Dictator game: indicator variable for any donations to host peers (sample: hosts in host-only or mixed groups)</p> <p>2.1.2 Dictator game: share of donations to other group members (sample: all beneficiaries and GL)</p> <p>2.1.2.1 Dictator game: share of donations to in-group peers (sample: all beneficiaries/GL)</p> <p>2.1.2.2 Dictator game: share of donations to out-group peers (sample: all beneficiaries/GL) <i>For eqs. 1/2, use in-group donations in non-mixed groups. For eq. 3, use out-group donations everywhere.</i></p> <p>2.1.2.3 Dictator game: share of donations to refugee peers (sample: refugees in refugee-only and mixed groups only)</p> <p>2.1.2.4 Dictator game: share of donations to host peers (sample: hosts in host-only or mixed groups)</p>
2.2	Trust	Beneficiaries, Group Leaders (GL), and enumerator	<p><b>Primary outcome:</b> 2.2.1 Trust peers to look after house (sample: all beneficiaries and GL)</p> <p><b>Secondary outcomes:</b></p> <p>2.2.1 Trust peers to look after house (sample: all beneficiaries and GL)</p> <p>2.2.1.1 Trust in-group peers with house (sample: all beneficiaries/GL)</p> <p>2.2.1.2 Trust out-group peers with house (sample: all beneficiaries/GL) <i>For eqs. 1/2, use in-group trust in non-mixed groups. For eq. 3, use out-group trust everywhere.</i></p> <p>2.2.1.3 Trust refugee peers (sample: refugees in refugee-only and mixed groups)</p> <p>2.2.1.4 Trust host peers (sample: hosts in host-only or mixed groups)</p> <p>2.2.2 Trust other people in general (beneficiaries and GL)</p> <p>2.2.3 Trust other people in this group (beneficiaries and GL)</p> <p>2.2.4 Difference between trust people in the group and trust people in general (beneficiaries and GL)</p> <p>2.2.5 People in this group appear to trust each other (Observation)</p>

2.3	Empathy	Beneficiaries and GL	<p><b>Primary Outcome:</b> 2.3 Empathy index: Anderson index of 2.3.1, 2.3.2, 2.3.3, 2.3.4, 2.3.5, 2.3.6. (sample: all beneficiaries and GL)</p> <p><b>Secondary Outcomes:</b>  2.3a Affective empathy index: Anderson index of 2.3.1, 2.3.2, 2.3.3  2.3b Cognitive empathy index: Anderson index of 2.3.4, 2.3.5, 2.3.6</p> <p><b>Index components:</b> When thinking about members of this public works group...</p> 2.3.1 I often feel sorry for people who are less fortunate than me 2.3.2 When I see someone being taken advantage of, I feel kind of protective toward them 2.3.3 Being in a tense emotional situation scares me 2.3.4 I try to look at everybody's side of a disagreement 2.3.5 I try to put myself in others shoes 2.3.6 I try to understand my friends better by imagining how things look from their perspective
2.4	Cooperation	Pairs, Beneficiaries, and Enumerators	<p><b>Primary outcome:</b> 2.4.1 Deer - Partridge Game: total payoff for both players (sample: all rounds)</p> <p><b>Secondary Outcomes:</b>  2.4.1 Deer - Partridge Game: total payoff for both players (sample: all rounds)  2.4.2 Deer - Partridge Game: payoff for player (sample: all rounds)  2.4.3 Indicator for choosing cooperative strategy</p> 2.4.4 Share who make decisions by focusing on themselves 2.4.5 Share who make decisions by considering their partner/common good 2.4.6 People help each other with their work (enumerator observation) <p><b>Secondary analyses:</b> To get a sense of how cooperation varies between the in- and the outgroup, and changes over time as the game unfolds, we will also analyse the game-specific outcomes (2.4.1/2/3) along the following dimensions:</p> <p><b>By in- vs. out-group</b>  2.4.x.1 outcome in games with in-group partners (all beneficiaries)  2.4.x.2 outcome in games with out-group partners (sample: all beneficiaries/GL)  <i>For eqs. 1/2, use in-group outcome in non-mixed groups. For eq. 3, use out-group outcome everywhere.</i>  2.4.x.3 outcome in games with refugee peers (sample: refugees in refugee-only and mixed groups)  2.4.x.4 outcome in games with host peers (sample: hosts in host-only or mixed groups)</p> <p><b>By known vs. unknown partner</b>  2.4.x.5 outcome in games with known partner  2.4.x.6 outcome in games with unknown partner</p> <p><b>By round</b>  2.4.x.7 outcome in each round of the game to test for learning/deepening of cooperation over time.</p> <p>We may also investigate interactions between these, i.e. whether the trajectory of trust in games differs between in- and out-groups, or whether knowing the partner within the group affects differences between in- and outgroup outcomes.</p>

Group 3: Social Cohesion			
ID	Outcome Family	Unit	Variables within each family
3.1	Attitudes towards PW peers	Beneficiaries / GL	<p><b>Primary outcome:</b> 3.1 Anderson index across all out-group attitudes and assessments (3.1.x), all respondents, and all peers within the public works group.</p> <p><b>Secondary outcomes:</b> We will then analyse impacts on in- vs. out-group peers in the same public work group, as well as for refugees and hosts separately:</p> <p>3.1a Attitude index (all) towards <i>out-group</i> peers (3.1.x.2)  3.1b Attitude index (all) towards <i>in-group</i> peers (3.1.x.1)  3.1c Attitude index (all) towards refugee peers (3.1.x.3)  3.1d Attitude index (all) towards host peers (3.1.x.4)</p> <p>To shed light on which index components move most, we will also separate the out-group attitudes into two separate indices (which we will analyze overall, and by in- vs. outgroup.</p> <p>3.1x Integration attitudes (index components 3.1.1-6)  3.1y Friendship and respect (index components 3.1.6-7)</p> <p><b>Index components:</b> We collect likert scales on the following attitudes between a respondent, and randomly selected peers within the PW group (both in- and out-group members in mixed groups):  3.1.1 ... would like to have peer as a co-worker outside of PW  3.1.2 ... would hire peer as an employee  3.1.3 ... would accept peer as a local police officer  3.1.4 ... would have tea/coffee with peer outside PW  3.1.5 ... would accept this person as the next group leader  3.1.6 ... how friendly their relationship is with their peer  3.1.7 ... whether they are respected by others</p> <p>For each of these attitudes/assessments of peers, we can include all assessments (in- and out-group), or only a subset of assessments to investigate whether impacts depend on the in- vs. outgroup:</p> <p style="padding-left: 40px;">3.1.x.1 assessment of in-group peers (all beneficiaries)  3.1.x.2 assessment of out-group partners (sample: all beneficiaries/GL)  <i>For eqs. 1/2, use in-group assessment in non-mixed groups. For eq. 3, use out-group assessment everywhere.</i>  3.1.x.3 assessment of refugee peers (sample: refugees in refugee-only and mixed groups)  3.1.x.4 assessment of host peers (sample: hosts in host-only or mixed groups)</p>
3.2	Interactions with peers outside of PW	Beneficiaries / GL	<p><b>Primary outcome:</b> 3.2 Number of distinct types of interactions with PW peers outside public works (3.2.x), all respondents, and all peers within the public works group.</p> <p><b>Secondary outcomes:</b> We will then analyse impacts on in- vs. out-group peers in the same public work group, as well as for refugees and hosts separately:</p> <p>3.2a Number of interactions outside PW with <i>out-group</i> peers (3.2.x.2)  3.2b Number of interactions outside PW with <i>in-group</i> peers (3.2.x.1)</p>

			<p>3.2c Number of interactions outside PW with refugee peers (3.2.x.3)  3.2d Number of interactions outside PW with host peers (3.2.x.4)</p> <p>To shed light on which index components move most, we will also separate interactions into social and economic interactions (which we will analyze overall, and by in- vs. outgroup.</p> <p>3.2a Number of social interactions with PW peers (including ‘co-living’, ‘romantic relationship’, ‘had a conversation of 15min or more’, ‘shared food/drinks’, ‘asked for a favor’, ‘gave a favor’)</p> <p>3.2b Number of economic interactions with PW peers (including ‘working together’, ‘sent money’, ‘received money’)</p> <p><b>Index components:</b> We collect likert scales on the following attitudes between a respondent, and randomly selected peers within the PW group (both in- and out-group members in mixed groups):</p> <p>3.2.1 ... whether they know them outside PW  3.2.2 ... how they have interacted outside PW (select multiple):”</p> <ul style="list-style-type: none"> <li>co-living</li> <li>romantic relationship</li> <li>working together</li> <li>had a conversation of 15min or more</li> <li>shared food/drinks</li> <li>sent money</li> <li>asked for a favor</li> <li>gave a favor</li> </ul> <p>For each of these interactions with peers, we can include all interactions (in- and out-group), or only a subset of pairs to investigate whether impacts depend on the in- vs. outgroup:</p> <ul style="list-style-type: none"> <li>3.2.x.1 interactions with in-group peers (all beneficiaries)</li> <li>3.2.x.2 interactions with out-group partners (sample: all beneficiaries/GL)</li> </ul> <p><i>For eqs. 1/2, use in-group interactions in non-mixed groups. For eq. 3, use out-group interactions everywhere.</i></p> <ul style="list-style-type: none"> <li>3.2.x.3 interactions with refugee peers (sample: refugees in refugee-only and mixed groups)</li> <li>3.2.x.4 interactions with host peers (sample: hosts in host-only or mixed groups)</li> </ul>
3.3	Quality of Interactions during PW (Observations)	PW Group	<p><b>Primary outcome:</b> 3.3 Number of positive interactions among group members (index of 3.3.1.1 - 3.3.1.6)</p> <p><b>Secondary outcomes:</b> We will then analyse impacts on in- vs. out-group peers in the same public work group, as well as for refugees and hosts separately:</p> <p>3.3a Number of positive interactions with <i>out-group</i> peers (3.3.x.2)  3.3b Number of positive interactions with <i>in-group</i> peers (3.3.x.1)  3.3c Number of positive interactions with refugee peers (3.3.x.3)  3.3d Number of positive interactions with host peers (3.3.x.4)</p> <p><b>Other secondary outcomes:</b></p> <p>3.3.2 Number of negative interactions (index of 3.3.1.7-9)  3.3.3 How would you describe the overall interactions within the group?</p>

			<p><b>Specifying index components:</b></p> <p>3.3.1.1 Interactions during PW: Laughing with others doing public works (=1)</p> <p>3.3.1.2 Interactions during PW: Singing, alone or with others (=1)</p> <p>3.3.1.3 Interactions during PW: Talking with others doing public works</p> <p>3.3.1.4 Interactions during PW: Talking to the group leader</p> <p>3.3.1.5 Interactions during PW: Helping others in the group with their work</p> <p>3.3.1.6 Interactions during PW: Positive physical touch, shaking hands, pat on the shoulder, hugging, etc.</p> <p>3.3.1.7 Interactions during PW: Ignoring some people in the group</p> <p>3.3.1.8 Interactions during PW: Verbal conflict: Arguing, loud talking</p> <p>3.3.1.9 Interactions during PW: Physical conflict, pushing, shoving, etc.</p> <p>For each of these interactions with peers, we can include all interactions (in- and out-group), or only a subset of pairs to investigate whether impacts depend on the in- vs. outgroup:</p> <p style="padding-left: 40px;">3.3.x.1 interactions with in-group peers (all beneficiaries)</p> <p style="padding-left: 40px;">3.3.x.2 interactions with out-group partners (sample: all beneficiaries/GL)</p> <p style="padding-left: 40px;"><i>For eqs. 1/2, use in-group interactions in non-mixed groups. For eq. 3, use out-group interactions everywhere.</i></p> <p style="padding-left: 40px;">3.3.x.3 interactions with refugee peers (sample: refugees in refugee-only and mixed groups)</p> <p style="padding-left: 40px;">3.3.x.4 interactions with host peers (sample: hosts in host-only or mixed groups)</p>
3.4	Discrimination (Observation)	Beneficiaries	<p><b>Primary Outcome:</b> 3.4.2 Share of refugees who get to use the available equipment (refugee-only and mixed groups only for eqs. 1/2, mixed groups only for eq. 3)</p> <p><b>Secondary outcomes:</b></p> <p>3.4.1 Share of hosts who get to use the available equipment (refugee-only and mixed groups only for eqs 1/2, mixed groups only for eq. 3)</p> <p>3.4.2 Share of refugees who get to use the available equipment (refugee-only and mixed groups only for eqs. 1/2, mixed groups only for eq. 3)</p>
3.5	Segregation (observation)	PW Groups	<p><b>Primary Outcome:</b> 3.5.1.3 Visible difference between beneficiaries by nationality (enumerator observation)</p> <p>We will also descriptively (outside the regression framework) analyse subgroup composition within mixed groups: Do subgroups that are observed at public works reflect homophily, or do they look as if they were randomly composed (i.e. reflecting the average share of in- vs. out-group members within the group as a whole. For equation 3, we use the deviation of subgroup refugee shares from the average refugee share in the overall group as a primary outcome.</p> <p><b>Secondary outcomes:</b></p> <p>3.5.1 Any visible difference between beneficiaries by demographics</p> <p style="padding-left: 20px;">3.5.1.1 Visible difference between beneficiaries: gender</p> <p style="padding-left: 20px;">3.5.1.2 Visible difference between beneficiaries: language</p> <p style="padding-left: 20px;">3.5.1.3 Visible difference between beneficiaries: nationality</p> <p>3.5.2. Any visible/observed subgroups by demographics</p> <p style="padding-left: 20px;">3.5.2.1 Subgroups by: Gender</p> <p style="padding-left: 20px;">3.5.2.2 Subgroups by: language</p> <p style="padding-left: 20px;">3.5.2.3 Subgroups by: nationality</p> <p>3.5.3. Indicator for different tasks across subgroups</p>

			<p>3.5.3.1 Different tasks by subgroups by: Gender</p> <p>3.5.3.2 Different tasks by subgroups by: language</p> <p>3.5.3.3 Different tasks by subgroups by: nationality</p> <p>3.5.4 Indicator for any differences in productivity observed across groups</p> <p>3.5.4.1 Differences in productivity observed by: Gender</p> <p>3.5.4.2 Differences in productivity observed by: language</p> <p>3.5.4.3 Differences in productivity observed by: nationality</p>
--	--	--	---

<b>Group 4: Team Dynamics</b>			
<b>ID</b>	<b>Outcome Family</b>	<b>Unit</b>	<b>Variables within each family</b>
4.1	Attendance and work satisfaction	Beneficiaries / GL	<p><b>Primary outcome:</b> 4.1.1 Reservation wage</p> <p><b>Secondary outcomes:</b></p> <p>4.1.1 Reservation wage</p> <p>4.1.2 Number of days beneficiary was present for PW over the last 5 days of PW (GL and self assessment)</p> <p>4.1.3 Like this work as PW group leader (GL only)</p> <p>4.1.4 Like working on PW/RHISN Projects (only beneficiaries)</p> <p>4.1.5 Happy coming to Public Works</p> <p>4.1.6 Stressful to be a group leader (GL only)</p> <p>4.1.7 Group Motivation (Observation, Beneficiaries, GL)</p> <p>4.1.8 Indicator for hard to motivate the group (GL)</p>
4.2	Team cohesion	Beneficiary/ GL	<p><b>Primary outcome:</b> Team cohesion index (Anderson index of 4.2.1, 4.2.2, 4.2.3, 4.2.4, 4.2.5, 4.2.6, 4.2.7, 4.2.8, 4.2.9, 4.2.10)</p> <p><b>Secondary outcomes:</b></p> <p>4.2.1 Team socialization scale</p> <p>4.2.2 Team harmony scale</p> <p>4.2.3 Collective vs. individual efforts</p> <p>4.2.4 Easy-going vs. conflict-prone</p> <p>4.2.5 Group contact index (Anderson index of 4.1.5.x)</p> <p><b>Index components:</b></p> <p>4.2.5.1 Contact within group: Perceived as equal</p> <p>4.2.5.2 Contact within group: Involuntary/voluntary</p> <p>4.2.5.3 Contact within group: Superficial/Intimate</p> <p>4.2.5.4 Contact within group: Forced/Natural</p> <p>4.2.5.5 Contact within group: Unpleasant/pleasant</p> <p>4.2.5.6 Contact within group: Competitive/cooperative</p>
4.3	Leadership	Beneficiary and enumerator observations, GL assessment)	<p><b>Primary Outcome:</b> 4.3.1 Indicator for beneficiary taking a leadership position in the group: Leading a sub-group or the full group (=1, Observation)</p> <p><b>Secondary outcomes:</b> We will then analyze leadership positions by in- vs. outgroups</p> <p>4.3.1 Indicator for beneficiary taking a leadership position in the group: Leading a sub-group or the full group (observation)</p> <p>4.3.1.1 Refugees taking leadership role (sample: refugees in refugee-only and mixed groups)</p> <p>4.3.1.2 Hosts taking leadership role (sample: hosts in host-only or mixed groups)</p> <p>4.3.2 Indicator for whether the GL says they take a leadership role in the group or subgroup</p> <p>4.3.3 Number of distinct actions taken by the GL to motivate the group (GL, observations)</p>

4.4	Participatory Governance	PW Group (Observations/ Beneficiary)	<p><b>Primary Outcome:</b> 4.4.2 Task assignment: Indicator for the group select the kind of work and tasks to do</p> <p><b>Secondary outcomes:</b>  4.4.1 Task assignment: GL chooses (Observations) (-)  4.4.2 GL Leadership style (Observations): democratic or facilitative  4.4.3 GL selection: the group voted (beneficiaries) (+)</p>
4.5	Absence of Conflict	PW Group (Observation)	<p><b>Primary outcome:</b> Existence of conflicts (4.5.1), rescaled to indicate absence of conflicts.</p> <p><b>Secondary outcomes:</b>  4.5.1 Was there any conflict between beneficiaries?  4.5.2 Number of beneficiaries involved in conflicts  4.5.3 Type of conflict: verbal  4.5.4 Type of conflict: physical  4.5.5 Source of conflict: work related  4.5.6 Source of conflict: Differences in values/religion  4.5.7 Source of conflict: Differences in ethnicity or language  4.5.8 Source of conflict: Political  4.5.9 Source of conflict: Gender-based  4.5.10 How was the conflict resolved: Not resolved  4.5.11 How was the conflict resolved: GL intervened  4.5.12 How was the conflict resolved: Someone else intervened  4.5.13. Number of beneficiaries who intervened in resolving the conflict?</p>

<b>Group 5: Gender Outcomes</b>			
5.1	Gender Norms (Observation)	PW Group	<p><b>Primary outcome:</b> Gender equality index (Anderson index of outcomes 5.1.1-5.1.11)</p> <p><b>Secondary outcomes:</b>  5.1.1 How are women working: intermingled with men (+)  5.1.2 How are women working: independently (+)  5.1.3 How are women working: leading the group, or subgroups (+)  5.1.4 The group leader appears to respect women in the group (+)  5.1.5 Men in the group appear to respect women (+)  5.1.6 Women doing different task than men (-)  5.1.7 Women work for less long than men (-)</p>
5.2	Gender Violence (Observations)	PW Group	<p><b>Primary Outcome:</b> 5.2.2 Number of observed incidences of harassment against women</p> <p><b>Secondary outcomes:</b>  5.2.1 Indicator for any observed incidence of women being harassed  5.2.2 Number of observed incidences of harassment against women  5.2.2.1 Number of observed incidences of verbal harassment  5.2.2.1 Number of observed incidences of physical harassment</p>

#### 4.1 Multiple outcomes and multiple hypothesis testing

Given the large number of outcomes collected and analysed in the surveys, there is a risk of false positives when interpreting statistical significance. To mitigate this concern, we organize variables into meaningful groups and outcome families. We will conduct inference with multiple testing adjustments for the primary outcomes within each group. Specifically, we will compute sharpened q-values. Following Anderson (2008), we report the lowest q-value at which each hypothesis is rejected. We will report both standard p-values and sharpened q-values for the empirical results.

## **4.2 Covariate adjustment**

To improve the robustness of our impact estimates from the data, we will implement Double/Debiased Machine Learning (DML), following the framework developed by Chernozhukov et al. (2018). This method combines flexible machine learning techniques with valid post-selection inference to correct for potential biases arising from high-dimensional covariates or model misspecification. By applying DDML to our estimation of the main equations, we aim to enhance both the precision and credibility of our results, particularly in the presence of complex, nonlinear relationships between outcomes and covariates. The covariate adjustment will be applied only for beneficiary-level outcomes and will potentially include covariates collected in the midline surveys as well as baseline surveys.

## **4.3 Exact tests of treatment effects**

Complementing the large-sample approach, we will also conduct Monte Carlo simulations to approximate exact tests of the treatment effect, following the framework introduced by Fisher (1935).

## **4.4 Heterogeneous impacts**

We will systematically assess the heterogeneous impacts of the intervention by i) hosts and refugees ii) whether the group leader is host/refugee iii) gender, (iv) age, (v) pre-existing social ties with refugees, vi) baseline prosocial behaviors toward out-groups, and (v) education.

We will also complement our primary empirical specifications exploring heterogeneous treatment effects using machine learning techniques such as causal forests (Wager & Athey, 2018) to flexibly estimate how the impact of team composition varies across other beneficiary and group characteristics. Some of these characteristics will come from the midline and baseline surveys.

We will also verify if our results on social cohesion and attitudes toward outgroups are sensitive to interactions with the Marlowe-Crown Social Desirability Scale which we collected for all beneficiaries and the group leader. This will enable us to test whether individuals' honest responses are affected by social norms.

## References

- Allport, Gordon Willard, Kenneth Clark, and Thomas Pettigrew, "The nature of prejudice," 1954.
- Hjort, Jonas, "Ethnic divisions and production in firms," *The Quarterly Journal of Economics*, 2014, 129 (4), 1899–1946.
- Lowe, Matt, "Types of contact: A field experiment on collaborative and adversarial caste integration," *American Economic Review*, 2021, 111 (6), 1807–1844.
- Mousa, Salma, "Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq," *Science*, 2020, 369 (6505), 866–870.
- Rao, Gautam, "Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools," *American Economic Review*, 2019, 109 (3), 774–809.
- Scacco, Alexandra and Shana S Warren, "Can social contact reduce prejudice and discrimination? Evidence from a field experiment in Nigeria," *American Political Science Review*, 2018, 112 (3), 654–677.
- Wager, Stefan, and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113, no. 523 (2018): 1228-1242.

## Administrative Information

**Funding:** We are grateful to the Forced Displacement Multi-Donor Trust Fund, Prospects, Knomad, IPA displaced livelihood initiative, and the Research Support Budget of the World Bank to provide financial support for this impact evaluation.

**Institutional Review Board (ethics approval):** This research project has been approved by Oxford's University Ethics Committee (approval number ECONCIA22-23-17).

**Declaration of interest:** The authors have no disclosures relevant to this study.