Pre-analysis plan: Impacts of a livelihoods program with built-in spillovers
(Deviations in blue)

Sarah Janzen[1], Nicholas Magnan[2], Sudhindra Sharma[3], and William M. Thompson[4]

[1]Kansas State University, sajanzen@ksu.edu
[2]University of Georgia, nmagnan@uga.edu
[3]Nepa School of Social Sciences and Humanities and IDA Associates, sudhindra@ida.com.np
[4]IDinsight, willt78@uga.edu

November 23, 2022

Abstract

We will evaluate 2.5-3.5 year welfare impacts of a livelihoods program using an RCT in Nepal. The program targets women and employs self-help groups, livestock transfers, and trainings. We assigned three variations of the program: full benefits, no livestock, and no values-based training, which includes encouragement to "pay it forward" (PIF) by training and giving livestock to others. With this encouragement in mind, the study is designed to evaluate the impact of the program for two subpopulations: direct and PIF beneficiaries. We will consider seven welfare outcomes of interest: women's empowerment, financial inclusion, psychological well-being, assets, income, expenditures, and food security & nutrition.

Acknowledgements: The USAID Feed the Future BASIS Assets and Market Access Innovation Lab provided research funding for this project. The work was also undertaken as part of the CGIAR Research Program on Agriculture for Nutrition and Health (A4NH) Gender and Assets in Agriculture Project Phase 2 (GAAP2) led by the International Food Policy Research Institute (IFPRI) and funded by the Bill & Melinda Gates Foundation (BMGF), A4NH, and USAID. All errors are our own.

1 Introduction

The rural poor are often assumed to lack access to the productive assets and human and social capital required to be successful entrepreneurs. Productive asset transfer programs, which often include a training component, are one way non-governmental organizations (NGOs) and governments try to relax these constraints to facilitate permanent transitions out of persistent poverty. Rigorous impact evaluations of bundled asset transfer and training programs, particularly evaluations designed to measure spillover effects, are a recent development. In this paper we evaluate the short-term (1.5 year) welfare impacts of Heifer International's (HI) Smallholders in Livestock Value Chain Program (SLVC) in rural Nepal using a randomized controlled trial (RCT). The program targets women in poor rural communities, and provides a package of benefits that includes a livestock transfer (two doe goats and a shared breeding buck), technical training on improved animal management and entrepreneurship, self-help group (SHG) formation, and values-based training. Rather unique to the HI program, the values-based training encourages beneficiaries to "pay it forward" by sharing newly acquired technical skills and giving the first-born female offspring of their received goats to another individual in their community.

This analysis will contribute to the literature in three important ways. First, we add to a small but growing body of empirical evidence on the welfare impacts of productive asset transfer programs, especially those bundled with extensive training components. The SLVC is related to the graduation programs that have recently been evaluated, notably the six BRAC Graduation Program Consortium programs studied by Banerjee et al. (2015), and the BRAC Targeting the Ultra-poor Program (TUP) considered in Bandiera et al. (2017). Graduation programs take a holistic livelihoods approach to tackling the interrelated challenges faced by the poorest of the poor, bundling a productive asset grant with technical skills training, access to financial services, intensive monitoring, and short-term cash stipends to support consumption. Evidence suggests that after three to four years, graduation programs increase financial inclusion, mental health, assets, income, total expenditures, food security, and political awareness (Bandiera et al., 2017; Banerjee et al., 2015).

Second, our evaluation is carefully designed to estimate built-in spillover effects in targeted communities. The SLVC employs a distinct targeting and recruitment model characterized by a community-level intervention and encouragement to pay forward the benefits received. Rather than targeting the poorest of the poor (the approach taken by most graduation programs), HI recruits all households residing in a targeted (and usually central) neighborhood regardless of relative wealth or poverty. Through values-based training sessions, these directly targeted beneficiaries are then encouraged to establish second generation

SHGs among others in their community, with the intent that the benefits of the SLVC will scale rapidly. These built-in spillovers embodied in the encouragement to pay benefits forward are an essential feature of all livestock transfer and training programs implemented globally by HI. To our knowledge, no study has evaluated the impact of a program with this type of pay it forward (PIF) model. Measuring the strength and persistence of this element of the program design is crucial to understanding overall program impacts.

Third, our evaluation includes three treatments designed to test the impacts of different program components. To our knowledge, previous studies in this area have not attempted to disaggregate the impacts of a bundled treatment, although this paper is contemporaneous with Banerjee et al. (2018), which seeks to unpack components of a graduation program in Ghana (specifically, they look at benefits from a stand-alone livestock transfer and additional benefits from access to a savings account and deposit collection). In the first treatment arm of our study, beneficiaries received a full treatment (FT) package that included a livestock transfer, skills-based technical training and values-based non-technical training. In the second treatment arm, beneficiaries received skills-based technical training and values-based non-technical training, but *not* goats (NG). In the third treatment arm beneficiaries received a livestock transfer and skills-based technical training, but *not* values-based non-technical training (NVT). Because encouragement to pay forward benefits is the primary element of the values-based training, the third treatment arm allows analysis of the PIF mechanism.

The core results from the short-term impact analysis using 2016 data were published in the 2018 AEA Papers and Proceedings (Janzen et al., 2018), and a longer working paper is available from the authors upon request. This pre-analysis plan describes our plans for analysis of the 2017 and 2018 impact analysis (2.5 and 3.5 year impacts). We note that at the time of writing this plan we worked extensively with the baseline and 2016 data but have not seen the 2017 or 2018 data. Although a research assistant has been checking, cleaning, and assembling the data, the authors have not discussed it with her at all.

The rest of the plan is structured as follows. In the following section, we describe Hl's livelihoods program in Nepal. Section 3 describes our experimental design, section 4 follows with a summary of the data collected for the evaluation. Our empirical approach is described in section 5, and a detailed description of outcome variables is presented in section 6. Section 7 contains our planned approach to benefit/cost analysis.

2 HI's livelihoods program in Nepal

Asset transfers, particularly livestock, have been provided to poor households by NGOs working in poor areas since at least 1944 when HI sent 17 cows from Arkansas to Puerto Rico. Since then HI has expanded its reach to over 125 countries. Numerous NGOs and even governments have also since embraced livestock transfer and training programs as a strategy for fighting poverty in rural areas (World Vision, BRAC, Save the Children, Oxfam, and the Government of Rwanda are a few examples).

The intervention we evaluate replicates HI's flagship program in Nepal, the SLVC. The intervention targets women and provides a package of benefits that includes the formation of women's self-help and savings groups, technical trainings on improved animal management and entrepreneurship, a productive asset transfer (in this case goats), and values-based trainings during which beneficiaries are taught to value and practice paying it forward.

The process is as follows: After identifying a location for project implementation, HI recruits an original group of direct beneficiaries. Direct beneficiary groups typically consist of most or all of the households in a given neighborhood (*tole*). As a rule, HI considers all the households in a targeted area to be objectively poor and therefore eligible for the program, allowing for the possibility that a considerable range of relative wealth and poverty exist within a group. Once selected, direct beneficiaries are organized into a self-help group (SHG). Over a period of several months all SHG members participate in a series of trainings. Trainings include (1) technical training on improved animal management, fodder/forage development, entrepreneurship, human and animal nutrition, and home gardening, and (2) HI's values-based training on topics of accountability, sharing and caring, sustainability, self-reliance, income management, environmental stewardship, spirituality, self-help group management, gender justice, and encouragement to pay it forward. The trainings culminate with the beneficiaries receiving a transfer of livestock which includes two doe goats for each beneficiary, as well as a shared buck of improved stock for the SHG to facilitate a breeding program.

A unique component of HI's approach is that it encourages members to pay benefits forward by recruiting additional community members into the program, giving a gift of livestock (of equal value to what was received), and passing down all technical knowledge that was gained through participation in the program. HI facilitates values-based empowerment training for both direct and PIF beneficiaries (albeit separately and at different points in time), while all other paid forward trainings are implemented by direct beneficiaries with minimal support from HI. As such, what might typically be thought of as a spillover effect is in fact an essential component of the overall program design (thus, "built-in" spillover

effect). In Nepal, the SLVC follows an innovation to the basic HI PIF model, in which each direct beneficiary SHG is tasked with recruiting up to five PIF SHGs, with the goal of full saturation and complete adoption of improved practices and technologies within a community over a short time frame.

Prior research has established positive welfare impacts of HI programs in a number of different contexts. The HI program in Zambia has been the subject of several such studies. In Zambia, the transfer includes a dairy cow, two draft cattle, or eight goats. Valued at approximately $2000, this transfer is five times as expensive as the SLVC in Nepal per direct beneficiary. Using difference-in-differences, Jodlowski et al. (2016) find evidence of increased consumption (expenditures) and dietary diversity. The analysis treats the intervention as a household-level one (rather than community-level), and includes PIF beneficiaries in the control group. Thus, they are unable to estimate impacts on PIF households and may find (likely downward) biased results for direct beneficiaries. Kafle, Winter-Nelson, and Goldsmith (2016) uses the same data and similar methodology to find the program increases livestock revenue and food expenditures and decreases subjective feelings of poverty. Phadera et al. (2017) leverages the panel nature of the Zambia data to investigate program impacts on resilience, and find direct beneficiaries were less likely to fall into poverty and more likely to be food secure. These latter two studies treat PIF households as if they were treated rather than in the control group, and find some attenuated effects on PIF households. The counterfactual used is "eligible" households in two other communities, yet the analysis is conducted as if treatment was assigned at the individual level casting doubt on statistical inference. It should also be noted that all three Zambia studies are based on 105 treated households receiving various livestock transfers (31 received a dairy cow, 20 received draft animals, and 54 received goats).

In Rwanda, Rawlins et al. (2014) used cross sectional observational data to investigate the impact of a HI program on child anthropometrics and consumption of animal source foods. They evaluate both a dairy cow transfer (valued at $3000) and a meat goat transfer (of unspecified value). Using regression controlling on observables including program eligibility, they find the dairy cow transfer increased milk consumption and the goat transfer increased meat consumption. Using regression and propensity score matching, they also find some evidence of impacts of both transfer types on child anthropometrics. This study was also based on a relatively small sample, with 155 households receiving livestock (78 received a cow and 77 received a goat) and a similarly sized comparison group.

The SLVC in Nepal has also been the subject of previous studies. Miller et al. (2014) and Darrouzet-Nardi et al. (2016) use a matched pair randomized control trial comparing three communities assigned to the SLVC program to three controls. Miller et al. (2014)

finds significant improvement in terms of height for age but not weight for age or days of school missed. Darrouzet-Nardi et al. (2016) find no overall effect of dietary diversity or consumption of animal sourced foods, but do find strong effects in the in the Middle Hills region (which consisted of one of the three matched pairs). In both of these studies, the authors did not account for the clustered level of the intervention, casting doubt on statistical inference. These two studies were based on 201 and 181 households receiving (the same) treatment, respectively.

Compared to the above studies, this paper is the most rigorous and comprehensive evaluation of a HI livelihoods program to date. The program evaluated here also shares several similarities with the graduation programs featured in the more rigorous and wide-ranging evaluations conducted by Banerjee et al. (2015) and Bandiera et al. (2017), but is distinct in several important ways. Like graduation programs, the SLVC bundles a one-time productive asset transfer with technical skills training on management of the transferred asset, some basic health and/or life skills training, and access to financial services. However, overall graduation program costs are significantly larger: the per-beneficiary cost across the six interventions in Banerjee et al. (2015) is up to thirteen times larger than the standard SLVC program cost per *direct* beneficiary. The value of the transferred assets are much lower in the SLVC, which explains much of the difference in cost. In addition, the SLVC does not conduct frequent follow-up home visits from program officers that are typical in graduation programs. Instead, HI relies on self-help groups (SHGs) to ensure beneficiaries internalize and implement the skills acquired through the intervention. While beneficiaries do receive small amounts of cash for specific purposes like building improved goat shelters and fodder/forage production, they do not receive a regular cash stipend for consumption support (graduation program recipients receive cash transfers for approximately one year). Finally, graduation program beneficiaries are not expected to pass on benefits in any way and existing impact studies indicate no evidence of spillover or indirect effects.

3 Experimental design

To establish a causal relationship between the program and changes in outcomes, this study uses a cluster RCT. A cluster design was employed for two reasons. First, by design HI's intervention targets groups rather than individuals. Second, local programatic spillover effects are an integral part of the intervention though the PIF mechanism. As described below, we will seek to estimate both direct and PIF effects.

Nepal comprises 75 districts. When this study began, districts were subdivided into Village Development Committees (VDCs). As of March 10, 2017 the VDC was dissolved

and replaced by the *gaunpalika*, or rural municipality. In this paper we will continue to use the VDC, as this was the geographical unit our design uses. A VDC can be thought of as groupings of villages within a district. Every VDC is split into nine wards, and each ward contains multiple *toles*, or neighborhoods. A typical *tole* in the study area has approximately twenty to thirty households; a typical ward has roughly 150 households.

Nepal-based HI staff first identified 60 VDCs in which they had never worked, but that would be good candidates for an asset transfer and training program. Before assigning treatments, HI also identified a central ward and targeted *tole* within that ward for each of the 60 selected VDCs. The expectation was that if assigned to treatment, everyone residing in the targeted *tole* would be targeted by the program, and therefore likely to enroll as a direct beneficiary. Through this process, HI pre-identified all targeted direct beneficiaries (but not necessarily actual beneficiaries). Following treatment assignment, HI put these targeted direct beneficiaries into SHGs in treated VDCs but not control VDCs. In this way, the individuals in the control arm are directly comparable with those in the treatment arms.

Direct beneficiaries are encouraged to pay benefits forward to individuals residing outside the *tole* but within the targeted ward. Although indirect PIF effects are anticipated within each central ward, and may spillover beyond ward-level administrative boundaries, we do not anticipate contamination of control VDCs. To an extent, the isolation of rural communities in Nepal provides a natural impediment to such contamination. This is especially true in the Middle Hills (home to sixty percent of our sample), where lower population density, rugged terrain, poor roads, and inferior cellular connectivity cause communities to be especially cut off. Nevertheless, communities are linked by family and commercial ties. Fewer natural barriers against contamination exist in the Terai, the densely populated plain along the Indian border. Apart from naturally occurring geographic and social barriers to contamination, we also buffered treated wards from each other and from control VDCs by selecting a central ward within a VDC to be the targeted ward. In this way, we ensure an additional degree of isolation and further reduce the prospect of unintentional spillovers across VDCs that could bias results.

To improve balance across treatment and control VDCs (and between the various treatment VDCs) we stratified by geography and caste/ethnic composition. First, we divided the sample of VDCs into four regional clusters containing 15, 15, 10, and 20 VDCs. Using administrative data, we ordered VDCs by caste/ethnic compositions to further subdivide regional clusters into strata bins of four VDC when possible.[1] We then randomly assigned

[1]We first calculated the proportion of residents in each VDC from each of 38 caste/ethnic groups. Then, within each regional cluster we ordered VDCs by the most prevalent caste/ethnic group, then second most prevalent caste/ethnic group, and so on through the ninth most prevalent caste/ethnic group. Finally, we ordered VDCs within regional clusters based on rank prevalence of caste/ethnicity.

treatment within bins.[2] Within each stratification bin, we randomly assigned the 60 VDCs to one of three treatment arms or pure control.

All three treatments share some common features. First, HI facilitates the formation of women's SHGs, so all beneficiaries are expected to acquire some level of social capital through group membership and participation. Group members are then encouraged to contribute to group savings accounts with a goal of increasing financial inclusion. All beneficiaries are trained on a variety of technical topics including nutrition, home gardening, fodder and forage production, and improved animal management. In addition, all beneficiaries are provided a small amount of cash support for home gardens (approximately $5), fodder/forage production (approximately $10), and goat shed improvement (approximately $40). Finally, all treatment VDCs receive access to a community animal health worker. We refer to these common features as the basic intervention.

In order to unpack the benefits of various program components, two additional programmatic elements vary across treatment arms: a productive asset transfer and additional values-based trainings. The productive asset transfer included two doe goats to each individual beneficiary, as well as a shared buck of improved breeding stock for the self-help group. The values-based trainings cover the HI "Cornerstones" not included in the basic intervention[3]: accountability; sharing and caring; sustainability and self-reliance; improving the environment; income; genuine need and justice; gender and focus on the family[4]; full participation; training, education, and communication; and spirituality. Perhaps most importantly, the values-based training encourages beneficiaries to pay benefits forward by providing technical training and giving the first two female offspring of their received livestock to another poor individual in their community.

The treatment arms can be described as follows:

1. *Full Treatment* (FT): basic intervention, values-based training, and livestock.
2. *No Goats* (NG): Identical to FT, but without the productive asset transfer.
3. *No Values-based Training* (NVT): Identical to FT, but without values-based training.

A fourth arm was randomly selected as pure control. Table 1 summarizes the elements of each treatment arm.

[2]Eleven bins had four VDCs, two bins had three VDCs and one had two VDCs. To ensure 15 VDCs in each treatment arm and the control arm we randomly re-alloced treatment in two VDCs.

[3]Improved animal management and nutrition are also HI Cornerstones, but are included as part of the basic intervention

[4]Notably, both men and women are encouraged to participate in gender and justice training.

Many of the welfare impacts we consider could be directly affected by either type of training or the asset transfer. For example, women's empowerment could increase as a result of interactions in the group, values-based trainings, technical skills trainings, and ownership over transferred assets or any resulting income. Similarly, income could increase as a result of any of these program components in concert or independently. Our experimental design allows us to differentiate between program components.

Figure 1 illustrates a timeline of relevant programatic activities and events. Project implementation began in mid to late 2014 (depending on location). All direct beneficiaries first formed SHGs (shortly after the baseline survey, as described below) and were encouraged to begin saving at this time. They also began training and built improved livestock shelters. Approximately six months later, between March and June 2015, these same direct beneficiaries received livestock if they were assigned to either the FT or NVT treatments. In late 2015 the second generation of beneficiaries, recruited by direct beneficiaries in their area, entered the program through the PIF mechanism, began to form groups, and participated in the various trainings. Notice that while we know when program activities for these beneficiaries began, it is difficult to know exactly when second generation PIF beneficiaries received livestock transfers because such transfers depend on livestock fertility, which is inherently random. In fact, the program is designed in such a way that PIF livestock transfers will be staggered, with some receiving livestock transfers within six months of enrolling in the program, while others will wait years before receiving a livestock transfer.[5]

For establishing hypotheses regarding mechanisms and the anticipated timing of impacts, we must carefully consider livestock fertility cycles. We assume a doe can reasonably be impregnated within any given four month window, a five month gestation period, and that offspring reach sexual maturity at around seven months (females) or an optimally marketable size at around ten months (males). Depending on breeding cycles and the availability of an improved buck, most direct beneficiaries would have been expected to impregnate their does between June and October of 2015, implying the members of a second generation of program goats were typically born near the end of that year and the beginning of 2016. Goats normally experience single births (although multiples are not uncommon), and the gender of the kid has important implications for impact. The program requires beneficiaries to donate their first two female offspring (once they have reached sexual maturity) to another beneficiary through the PIF mechanism. Male kids are sold for meat and not passed on to PIF beneficiaries.

Taken together these facts imply three noteworthy features of this study, all shown in

[5]SHGs figure out among themselves who will receive goats when among PIF beneficiaries based on their own criteria.

Figure 1: (1) the earliest PIF beneficiaries could possibly have received goats was in mid 2016, (2) the earliest possible goat sales (of transferred goat kids) for direct beneficiaries would have taken place after in late 2016 (after the data analyzed in Janzen et al. (2018) , and (3) the earliest possible goat sales (of transferred goat kids) for second generation PIF beneficiaries would have taken place in early 2018. These features are important for understanding mechanisms and impacts.

The intervention concluded in mid 2017. At this time, official HI program activities and monitoring ceased.

4 Data

4.1 Sample description

We collected baseline data from rural women eligible to participate in the program across three regions of rural Nepal in June-September 2014. Midline data was collected in June-July 2016, approximately 1.5 years after initial enrollment in the program and analyzed from 2016-2018 (Janzen et al., 2018). This analysis plan concerns the analysis of data collected in mid-2017 (endline 1) and mid-2018 (endline 2). Notably, the endline 1 survey took place shortly before official program activities concluded. The endline 2 survey then took place approximately 1 year after the "end of project," which is important for evaluating persistent effects. Figure 1 shows how the survey timeline fits with program implementation.

There are two types of respondents in the endline sample: targeted direct beneficiaries (in the central ward), and prospective PIF beneficiaries (also in the central ward). Specifically, our sample of targeted direct beneficiaries consists of all households in each of the targeted *toles* (around 25 per ward). In addition, after removing households from the targeted *tole*, we originally selected a random sample of 15 potential PIF beneficiaries from a complete roster of all households in the same ward. Because of the aggressive nature of the PIF model, we expect that many (if not most) of these households had the opportunity to enter the program. Although no intervention took place in control VDCs, sampling in these VDCs occurred in exactly the same manner as in treatment VDCs: 25 individuals from pre-determined targeted *toles*, and 15 individuals from a complete roster of all other households in the central ward.

Our total baseline sample used in the impact analysis is 2,376 women, including 1,286 targeted for direct treatment, and 1,089 households from the central ward likely to enter the program through the PIF mechanism (i.e. PIF households). Shortly after HI delivered training and livestock to the original beneficiaries of the project, a devastating earthquake

struck Nepal. The earthquake greatly affected the 10 VDCs belonging to the Middle Hills stratification pool, and were therefore spread evenly across treatment groups and control. We made the decision to drop these from the RCT so that HI could provide earthquake relief in whatever manner they deemed appropriate. Following additional attrition not explicitly related to the earthquake, the remaining midline sample consisted of 50 VDCs and 1,828 households, including 1,031 from targeted *toles* and 797 PIF households from the central ward more broadly.

Our own simulations demonstrate power gains can be achieved using a larger within cluster sample even given lower baseline correlation due to missing baseline data.[6] To make up for the loss of power due to fewer clusters and reduced sample size following the earthquake, we added additional PIF households to the sample for endlines 1 and 2.[7] Specifically, we removed households already in the sample, and then selected a random sample of 30 *additional* PIF beneficiaries from the complete roster of all households in the central ward at baseline (the exact same lists used at baseline). We refer to these households as the expanded PIF sample. Following additional sampling and attrition, the total endline 1 sample consisted of 50 VDCs and 3,222 households, including 1,034 from targeted *toles* and 2,188 PIF households from elesewhere the central wards. After attrition, the remaining endline 2 sample consisted of 50 VDCs and 3,111 households, including 1,012 from targeted *toles* and 2,099 PIF households from elsewhere in the central wards.

4.2 Network data

During baseline data collection we took photos of all respondents in treatment VDCs, which were then used to make composite photo directories. Approximately three months later we used these directories to ask respondents in all treatment VDCs (but not in control VDCs) about their connections to one another. For each other surveyed individual in the same VDC, respondents were asked who is in their close family, who is a more distant relative, who is a friend, who is a geographic neighbor, who is a acquaintance, who is a familiar face, who is someone they talk to about family issues, who is someone they talk to about livestock, who do they trust livestock information from, who do they trust financial information from, who is economically better off than they are, and who is economically worse off they they are. We will use this data to explore recruitment mechanisms into Heifer SHGs.

[6] This holds except when the intra-cluster correlation and the amount of baseline correlation not captured by strata fixed effects are both high.

[7] Our original sample of targeted direct beneficiaries includes everyone from that sample frame, so we were unable to expand the direct beneficiary sample.

4.3 Balance

We are confident that treatment was randomly assigned, as there was no possibility for HI to re-assign treatment after our randomization. However, imbalance by chance is a distinct possibility. To test for balance across treatments we regress $y_{hv}^{t=0}$, a demographic characteristic or outcome index for household h residing in VDC v as measured at baseline ($t = 0$), onto treatment status. Specifically, we estimate the regression below separately for the subsamples of direct and PIF beneficiaries:

$$y_{hv}^{t=0} = \beta_0 + \beta_1 T_{hv}^{FT} + \beta_2 T_{hv}^{NG} + \beta_3 T_{hv}^{NVT} + \varepsilon_{hv}. \tag{1}$$

In equation 1, T^{FT}, T^{NG}, and T^{NVT} are dummy variables for a household receiving the "full treatment" package, the "no goats" package, and "no values-based training" package, respectively, and ε_{hv} is an idiosyncratic error term clustered at the VDC level. Because magnitude is important, we will also report normalized differences. We will report balance for all primary outcome indices and any control variables used in the analysis.

4.4 Attrition

To assess if and how attrition might affect our results, we will first regress attrition on treatment status for direct and PIF beneficiaries separately at time t. This is the same specification as in 1, where the dependent variable becomes $attrit_{hv}^{t}$, a dummy variable taking a value of one for any household missing from the sample at time t. Again we cluster standard errors at the VDC level.

$$attrit_{hv}^{t} = \beta_0 + \beta_1 T_{hv}^{FT} + \beta_2 T_{hv}^{NG} + \beta_3 T_{hv}^{NVT} + \varepsilon_{hv}. \tag{2}$$

We will also test if attrition is correlated with household characteristics by regressing attrition status, $attrit_{hv}^{t}$, on a vector of demographic and outcome index values at baseline ($\mathbf{y_{hv}^{t=0}}$). Again, we will do this at endlines 1 and 2 for direct and PIF beneficiaries separately:

$$attrit_{hv}^{t} = \alpha + \boldsymbol{X_{hv}^{t=0}}\boldsymbol{\beta} + \varepsilon_{hv}. \tag{3}$$

Finally, we will conduct balance tests as in equation 1, but restricting the sample to returning direct and PIF households at endlines 1 and 2 allowing us to observe systematic attrition.

Instead of running these three separate analysis, we ran a single regression that included

treatment status, demographic and outcome variables at baseline, and the interactions thereof. This allows us to test whether attrition varies by treatment status (it does not) and whether differential attrition by treatment status is correlated to other variables (it is not more so than expected by chance). Even though these tests do not suggest attrition affects our results, we report Lee bounds Lee (2009) in appendix B of the paper.

4.5 Questions with limited variation

Questions for which 95 percent of control observations have the same value at time t will be omitted from the analysis. If doing so makes it impossible to calculate a proposed indicator, the indicator will be not be calculated.

5 Empirical approach

Our main research questions are: (i) what are the long-term welfare impacts of a livestock transfer and training program? (ii) are all program components necessary for achieving impact? (iii) within a treated ward, are benefits effectively passed on to subsequent generations of beneficiaries? and (iv) which package of benefits results in the most cost-effective improvements to household and individual well-being? We present our empirical approach for addressing questions (i)-(iii) in this section. Section 7 presents our analysis of question (iv).

5.1 Recruitment and retention

For understanding welfare impacts, it is helpful to analyze recruitment and retention in the program. We will first estimate the following equation where m_{hv}^t is a dummy variable for stated HI SHG membership on assigned treatment status at time t.

$$m_{hv}^t = \beta_0 + \beta_1 T_{hv}^{FT} + \beta_2 T_{hv}^{NG} + \beta_3 T_{hv}^{NVT} + \varepsilon_{hv}. \tag{4}$$

Membership is determined by both recruitment into the program and retention until time t. Assuming the participation reported in control VDCs is an error (there are many NGO-operated groups one could belong to), we can adjust membership rates as the estimated coefficient for each treatment dummy, netting out the reported control group membership. We note that doing so would likely lead to conservative estimates of recruitment rates, as respondents in treatment VDCs are probably less likely to say they are in a HI SHGs than

respondents in control VDCs because they are likely aware of the HI intervention and SHG activities. We can calculate less conservative estimates as treatment arm means. Because households may continue joining, or drop out of, SHGs between rounds of data collection we will report membership at both endline 1 and endline 2.

Instead of using regressions, we opted to report what percentage of respondents in a group (targeted or non-targeted) claimed to be an original group or pay-it-forward (PIF) group member. This is because there was much more cross over between what group a respondent was targeted for, and what type of group they ended up joining. We found it much easier to describe, compare, and graph membership trends by using sample means as opposed to regression output.

PIF recruitment is a particularly unique and important aspect of the HI intervention. It is intended to rapidly scale out benefits, greatly reducing per-beneficiary costs. Thus, we are interested in understanding how the pay-it-forward mechanism works in practice, including how direct beneficiaries select and recruit PIF beneficiaries. This will help us understand the characteristics of actual PIF beneficiaries (compliers) under current practice and the avenues through which direct beneficiaries recruit and/or potentially exclude individuals from the PIF process. To do this we will expand the regression in 4 using interactions between baseline descriptive statistics and treatment status.

We will also test how pre-existing social linkages affect recruitment of PIFs. Using social network data we can typify the type and directionality of each relationship between individuals in the original PIF sample and those in the OG sample and use this data to predict recruitment. We can also use measures of sameness captured in other ways using demographic variables. We do not know which individuals from the OG were instrumental in recruiting a given PIF (recruiting is supposed to be a group effort), therefore we will use PIF connections to the group as a whole to predict recruitment.

Because our analysis of recruitment channels is exploratory rather than confirmatory and attempts to describe mechanisms rather than outcomes, we will do not go into further detail in this PAP.

Instead of running the regressions described above, we conducted t-tests between members and non-members along a number of baseline variables, including network variables (has at least one targeted friend, has at least one targeted acquaintance, number of targeted friends, and number of targeted acquaintances.)

5.2 Main specification

To analyze the welfare impacts of a productive asset transfer and training program, we estimate the intent to treat (ITT) effects for each of the three treatment groups relative to a common control, noting that ITT will be somewhat conservative since takeup rates were not 100%. To analyze whether treatment effects reach subsequent generations of beneficiaries, we estimate these effects separately for two subsamples: direct and PIF beneficiaries. In the analysis of direct treatment effects, the sample consists of those pre-selected for direct benefits (including those in control areas). In the analysis of PIF treatment effects, the sample consists of all other individuals in the central ward (including those in control areas). We note that PIF effects could arise through technical training conducted by direct beneficiaries, values-based training conducted by HI staff, asset transfers from direct beneficiaries, or through observation and replication. If households observe and replicate the behavior of direct beneficiaries, they may benefit indirectly from trainings, even if they do not identify as a second generation program beneficiary.[8] By estimating ITT effects we can capture benefits coming through all of these channels.

We estimate the following equation at time t separately for the direct and PIF beneficiary subpopulations:

$$y_{hv}^{t} = \beta_0 + \beta_1 T_{hv}^{FT} + \beta_2 T_{hv}^{NG} + \beta_3 T_{hv}^{NVT} + \delta y_{hv}^{t=0} + \boldsymbol{X_{hv}'\gamma} + \boldsymbol{S_b'\rho} + \varepsilon_{hv} \tag{5}$$

In equation 5, y_{hv}^{t} is the outcome of interest for household h in ward v, measured at time t. As in equation 1, T_{hv}^{FT}, T_{hv}^{NG}, and T_{hv}^{NVT} are indicator variables for a household being in a VDC selected to receive the corresponding treatment. Control variables include the outcome of interest measured at baseline ($y_{hv}^{t=0}$), a vector of de-meaned covariates measured at baseline thought to affect outcomes including a full set of treatment interactions as described below ($\boldsymbol{X_{hv}}$), and strata bin dummies ($\boldsymbol{S_b}$).

For each subsample (direct or PIF) used in the estimation, β_1 represents the ITT effect on households in VDCs where direct beneficiaries were selected to receive the full treatment package (FT), β_2 identifies the same for the no-goats package (NG), and β_3 identifies the same for the no-values-based training treatment package (NVT). The counterfactual is targeted (direct or PIF) beneficiaries located in control VDCs.

New research by Goldsmith-Pinkham, Hull, and Kolesár (2022) finds that regressions containing multiple treatment variables and control variables, including strata bin dummies, can produce biased estimates. One of their suggested remedies is to run "one-treatment-at-a-time" regressions, which we do. The downside of this approach is that it is more difficult to

[8] This explains why LATE is not our preferred approach.

make comparisons across treatment arms. To make comparisons across treatment arms, we adopt another version of the one-treatment-at-a-time approach, which treats the BAP (FT) treatment as the control and either the BA (NVT) or BP (NG) as the treatment, excluding the other from the estimation sample along with the control.

For each outcome of interest we will calculate and report the minimum detectable effect.

We did not report minimum detectable effects. This is because it seems clear to us that where we fail to reject a null hypothesis (welfare outcomes), it is first and foremost because there is no evidence of an effect rather than because we are underpowered. The point estimates vary wildly across treatments and subgroups in ways that appear completely non-systematic.

5.2.1 Selecting control variables

To maximize power without p-hacking, we will employ machine learning to select the covariates to be included in $\boldsymbol{X_{hv}}$ for each regression. Specifically, we will employ the post-double-selection lasso (Least Absolute Shrinkage and Selection Operator) estimator developed by Belloni, Chernozhukov, and Hansen (2014) and implemented using the PDSLASSO command in Stata (Ahrens, Hansen, and Schaffer, 2018). All candidate control variables will be de-meaned, and a full set of treatment interactions will be included in the set of candidate controls. Current household size will always be included in the amelioration set to avoid the need for considering per capita outcomes. As indicated in equation 5, the outcome at baseline (when available) and strata bin dummies will always be controlled for.

Candidate control variables can be classified as either time invariant or potentially time-varying outcomes measured at baseline. Time invariant demographic variables include respondent and spouse age, literacy, and years of schooling, a dummy for no spouse, and the maximum years of schooling of the most educated household member. Because they are time invariant, we can use endline 1 data for these outcomes for households in the expanded PIF sample or other cases where a variable is missing at baseline. The latter two variables in the list are possibly time variant, but since they are not expected to be impacted by the intervention and relatively unlikely to change, we will impute endline 1 data for households with missing observations at baseline. For example, if the most educated household member was still in school at baseline —which we cannot observe for expanded PIF households —but not at endline 1, we cannot observe the exact number of years of schooling at baseline but can approximate it using current years of schooling and age. In the analysis of PIF impacts, we do not have baseline data for a large proportion of observations (see section 4.1 for an explanation of why). These eight characteristics are therefore the only control variables that are not missing for the expanded PIF sample.

We opted not use the expanded PIF sample. The gains from having baseline data seemed to overcome the gains from having a larger sample. While we do not have abnormally high imbalance across treatments, we have some, and want to be able to control for outcomes at baseline.

We will use all seven welfare indices and a subset of pre-selected subindices and mechanism variables measured at baseline as candidate time-varying control variables. We select 13 subindices that we believe are likely to be more predictive of outcomes than the index as a whole. These include land area, a productive assets index, total livestock herd size (TLU), total livestock income, total livestock investment, goat herd size, goat revenue, goat production practice index, decision-making over goats index (see section 6.8 for defining the latter two indices), income aspirations, social status aspirations, an ordered categorical "patience" variable, and an ordered categorical "planning horizon" variable. As suggested by Lin (2013) and Bloniarz et al. (2015) we de-mean all covariates and also include the interaction terms between treatment variables and de-meaned covariates as potential control variables. We thus have 28 covariates plus 84 covariates interacted with treatment dummies to use for the LASSO procedure.

We decided to use far fewer control variables, focusing on the ones that are most directly related to the program. We use the 21 variables reported in the balance table and do not interact them with the treatment variables. Controls available for selection by PDSLASSO are consistent across all regressions..

Before selecting control variables, we will first impute median values for all missing observations of all candidate control variables in the OG and original PIF samples. We will then impute median values for all missing observations of time invariant observations in the expanded PIF sample.

Due to the nature of missing data in our PIF sample we will select control variables using PDSLASSO in two steps. After imputing medians for all missing observations, we will first select among all possible time-invariant variables using the full PIF dataset. In the second step, using only the original PIF sub-sample, all time-invariant variables from the first step will be included in the second stage amelioration set, and then we will select additional time-variant control variables. For analysis, we will control for selected variables and include a dummy variable adjustment for each selected control variable for which there is at least one missing observation as well as a dummy variable for being in the expanded PIF sample (and therefore missing all time-variant control variables).

We do not do this because we do not use the data from the expanded PIF sample.

5.2.2 Multiple hypothesis testing

We account for multiple hypothesis testing in two ways. First, the summary index for each welfare dimension consolidates several individual tests into a single test. Second, because we still have multiple outcome dimensions, we control for the false discovery rate (FDR). Specifically, we calculate both naive p-values and q-values for multiple hypothesis tests for our main results across summary indices, using the Benjamini and Hochberg (1995) step-up method as described in Anderson (2008). We do not adjust p-values across treatments, time, or subpopulations. We also do not adjust p-values for the exploratory analysis of mechanisms or quantile regression.

We control for FDR for welfare indicators as planned. We also control for FDR for the key proximate outcomes (a subset of what we call here the "exploratory analysis of mechanisms"). These include a goat good practices index, goat herd size, net goat revenue, women's decision making power over goats index, respondent savings, and household debt. We chose these variables because outside of the welfare variables, these are most indicative of whether the program benefits households, both targeted and non-targeted.

5.3 Differential impacts and pooling data

The experimental design and sample structure allows us to estimate results from different treatments for different sub-populations at different periods of time. Looking for differences along these dimensions is an important aspect of this study. However, we will likely be underpowered to detect differences between treatments, between beneficiary type, or between time periods unless one impact is very small and the other large. We will therefore focus on statistical comparisons of outcomes between a given treatment for a given population at a given time and zero. To allow the reader to make statistical comparisons we will display standard errors in tables and confidence intervals in figures.

If impacts are very similar across treatments or time periods we can gain statistical power through pooling, allowing us to detect statistically significant differences (from zero) we otherwise may not have. For instance, if the provision of goats through the program do not matter we would expect T1 and T2 to have similar impacts (and our analysis of midline data shows that this is usually the case). It is possible that we see similar point estimates for an outcome under T1 and T2 where neither is significantly different from zero on its own, but the impact of receiving either treatment is statistically different from zero (and the weighted average of the point estimates for T1 and T2 independently). It could also be that point estimates of impacts do not change much between endline 1 and endline 2, but at neither point are the impacts significantly different than zero. In such cases we can increase

precision by pooling data from the two endline surveys and estimating an average treatment effect across the two years (McKenzie, 2012).

We will pool data or treatments if impacts are similar (across treatments and/or time periods) and the analysis is underpowered for detecting significant effects. For treatment type, we will consider pooling any combination of T1, T2, and T3 to allow for the possibility that the goats, values-based training, or neither add to the impact of the basic intervention. For time period, we will consider pooling endline 1 and endline 2. For direct and PIF beneficiaries we can pool observations, applying sample weights so that we can interpret results as average impacts for a randomly selected household in the entire ward.

We do not do this because there are already so many dimensions of analysis in the paper.

5.4 Heterogeneous treatment effects

If we can identify observable characteristics that predict which households are likely to benefit from the HI intervention, or which households are most likely to require a particular program component (like goats) in order to benefit, we could leverage such results to improve program design and/or targeting.

With this in mind, we will use machine learning to analyze heterogeneous treatment effects. Following Athey and Imbens (2016) and Wager and Athey (2018), we will use random forests to estimate treatment effect heterogeneity, to test hypotheses about the differences between the effects in different subpopulations, and to identify characteristics associated with greater program impacts. Using this approach, we will analyze heterogeneity of impacts across the 28 covariates defined above as possible control variables. We will not include the expanded PIF sample in this heterogeneity analysis. We will pool data or treatments for heterogeneity analysis if impacts are similar (across treatments and/or time periods) and the analysis is underpowered for detecting significant effects.

In addition, we will employ quantile regression for each of the seven welfare indices. This provides the flexibility to identify different estimates at different parts of the distribution of each welfare index. We will include the expanded PIF sample in the quantile regression analysis. We will pool for quantile regression if impacts are similar (across treatments and/or time periods) and the analysis is underpowered for detecting significant effects. We may also do quantile analysis of certain subindicators in an exploratory analysis of mechanisms.

We do not test for heterogenous treatment effects or conduct quantile regression because there are already so many dimensions of analysis, and because the focus of the paper is on the functionality of the pay-it-forward mechanism as opposed to the welfare indices.

5.5 Mechanisms

Although our main analysis relies on summary indices, we will also consider impacts using all subindicators as the outcome variable. This supplements the welfare analysis, and it also serves as part of our exploratory analysis of mechanisms. We will not adjust p-values for the impact on subindicators.

Furthermore, the data provides additional information that can be useful for understanding mechanisms. We thus propose to analyze the impact of the program on a number of behavioral and other outcomes directly related to the livestock-based intervention but are not included in any welfare summary indices. Our approach to this analysis of mechanisms will follow the empirical strategy outlined above. In cases where summary indices seem natural we will construct summary indices; otherwise we will not. We will not adjust p-values for multiple hypothesis testing. For analysis of mechanisms we will not pool across time or treatments. Proposed outcome variables related to mechanisms are defined following the definitions of welfare indicators and subindicators in section 6.8 below.

6 Defining outcomes

We consider seven primary welfare outcomes of interest: women's empowerment, financial inclusion, psychological well-being, assets, income, expenditures, and food security & nutrition. Each dimension consists of multiple subindicators described and summarized in the subsections below. These subindicators are then aggregated into a summary index for each dimension of welfare. Summary indices for empowerment, financial inclusion, psychological well-being, assets, and food security & nutrition will be calculated as standardized inverse covariance weighted (ICW) averages of subindicators following Anderson (2008). Income is total household annual income and expenditures is total household annual food and non-food expenditures. Summary indices will allow us to draw tractable conclusions about the program reaching broad objectives. Their use can also increase power by aggregating a number of impacts that are not statistically significant but move in the same direction, resulting in a statistically significant impact on index, and reduce dimensionality to mitigate problems arising from multiple hypothesis testing.

6.1 Empowerment

We employ subindicators (modified to the local context) from the Five Domains of Empowerment (5DE) of the Abbreviated Women's Empowerment in Agriculture Index (A-WEAI)

to calculate an empowerment score for all women in the sample (Alkire et al., 2013; Malapit et al., 2015). The A-WEAI was developed based on pilot surveys conducted in three countries through extensive collaboration between the United States Agency for International Development, the International Food Policy Research Institute and Oxford Poverty and Human Development Initiative. The A-WEAI aggregates an empowerment score across decisions about production, ownership of productive assets, access to and decisions on credit, control over income, group membership & leadership, and workload. Each binary subindicator equals one if the respondent achieves "adequate" empowerment in that area, and zero otherwise. Although we employ subindicators based on the A-WEAI, in a deviation from the A-WEAI, we will calculate the ICW average following Anderson (2008) rather than using standard weights defined by the A-WEAI. We do this for consistency across indices and to better leverage components of the index where there is greater variation among respondents.

Definitions of adequacy are based on the A-WEAI, but adjusted to the local context. Specifically, we use the following definitions of adequacy: A respondent is adequately empowered in production decisions if she has at least some input into at least one production decision. Adequate ownership means the household owns at least one productive asset, livestock or land, and the respondent (individually) has at least some ownership of an asset, livestock or land. A respondent is adequately empowered in access to and control over credit if the household has at least some credit and the respondent participated to any extent in the decision to borrow. Adequacy in control over income means the respondent participates in at least one decision regarding non-food expenditures. A respondent is adequately empowered in group membership if she is a member of any group. A respondent is adequately empowered in leadership if she holds a leadership position in any group. A respondent is adequately empowered for time use if she worked 10.5 hours or less in the previous 24 hours.

The baseline empowerment index is a weighted standardized average of the same subindicators, and we use the same definitions of adequacy. However, some of the survey questions differ between baseline and endline, such that the baseline empowerment index is calculated slightly differently.

We do not use the A-WEAI as proposed because by endline empowerment levels using this metric were extremely high in both the treatment and control group; there simply was no room for improvement. Instead we created indices for empowerment over spending, empowerment over assets, and empowerment over production, all of which are important components of the A-WEAI.

6.2 Financial inclusion

Financial inclusion subindicators reflect household savings and credit behaviors. Savings indicators include amount saved in the last month, a dummy variable equal to one if the household saved in the last month, the total amount of current savings, a dummy variable equal to one if the household currently has any savings, and a dummy variable equal to one if the household belongs to a savings group. Indicators related to credit disaggregate informal and informal loans, and also consider loans for investment. For each of the three types of loans (which can be overlapping), two subindicators are used: the amount owed formal/informal/investment, and a dummy variable equal to one if the household currently has a informal/formal/investment loan.

The baseline financial inclusion index does not include the total amount of current savings or a dummy variable equal to one if the household currently has any savings due to differences in the survey.

We did not use an index for financial inclusion because we realized we could not clearly distinguish formal from informal loans, nor could we easily disaggregate loans by purpose. Thus it was unclear which direction (qualitatively) different kinds of loans should move the index. Are loans good or bad? Instead, we focus on simpler credit and savings metrics when looking at proximate program outcomes.

6.3 Psychological well-being

The index is an ICW average of several aspects of psychological well-being. Depression (inverse) is based on nine questions from an abbreviated version of the CES-D scale Radloff (1977) with a high value indicating *low* levels of depression. Worry (inverse) employs nine questions from the Penn State worries questionnaire, and a high value indicates *less* worried. Self-esteem is based on eight questions from Rosenberg (CITE). Life satisfaction is based on one question from the World Values Survey (CITE). Optimism is based on six questions from Sheier (CITE). Locus of control is an abbreviated Rotter (1966) scale based on 19 questions where a high value indicates a stronger internal locus of control.

The baseline psychological well-being index is a weighted standardized average of the same subindicators, but many of the subindicators are calculated differently due to survey changes between baseline and endline. At baseline, depression (inverse) was calculated using four questions from an abbreviated version of the CES-D scale Radloff (1977) with a high value indicating *low* levels of depression. Worry (inverse) was calculated using four questions from the Penn State worries questionnaire, and a high value indicates *less* wor-

ried. Self-esteem was calculated using six questions from Rosenberg (CITE). Optimism was calculated using four questions from Sheier (CITE). Locus of control was calculated using an abbreviated Rotter (1966) scale based on six questions where a high value indicates a stronger internal locus of control. Life satisfaction was calculated in the same way at baseline as at endline 1 and endline 2.

6.4 Assets

The asset index is the ICW average the value of productive assets, the value of non-productive assets, land, and livestock. Productive assets include draft animals (excluding water buffalo), plows, tractors, motor cars or other vehicles, computers, printers, grinders, threshers, looms, sewing inventories, mechanical tools, hand tools, bicycles, motorcycles/scooters, solar panels, batteries, inverters, and improved livestock pens. Non-productive assets include phones, radios, cassette recorders, DVD players, televisions, satellite dishes, cameras, camcorders, electric fans, heaters, refrigerator/freezers, gas stoves, cupboards, jewelry, watches, tables, chairs, sofas, and mattresses. Livestock represents owned goats, cattle, water buffalo, swine and chickens aggregated into tropical livestock units (TLUs)[9]. Land is measured by total hectares owned.

The baseline survey does not accommodate calculating the value of productive or no-productive assets. The baseline survey also varies in the types of assets. Instead of using value, the non-productive asset index is a principle components index of dummy variables indicating ownership of non-productive assets (including radios, televisions, mobile phones, heater/pressure lamps, electric fans, camera/camcorders, furniture, irons, jewelry and watches). For productive assets, principle components analysis led to inexplicably signed weights. Instead we sum dummy variables indicating ownership of productive assets (including plows, grinders, threshers, sewing inventories, mechanical tools, tractors, motorbikes, bicycles, cars, computers and improved livestock pens).

Rather than combine monetary and non-monetary values in a single index, we separately looked at asset value (for assets we had a value for), land (measured in hectares), and livestock (measured in tropical livestock units). The asset value data is incredibly noisy, whereas the land and especially livestock data is less so.

[9]We follow the FAO's guide (FAO, 2005) to calculating TLUs in Nepal: cattle = 0.5, buffalo = 0.5, sheep & goats = 0.1, pigs = 0.2, chicken = 0.01.

6.5 Income

The income index is total household income, which is the sum of all subindicator categories: livestock income, value of livestock produced and consumed at home[10], crop income, value of crops produced and consumed at home[11], small business or entrepreneurial income (income from non-crop, non-livestock enterprises), income earned as a day laborer, permanent salaried income, remittances, and other miscellaneous income including cash transfers, gifts, pension, rental income, social security allowance or other income.

If a respondent claims zero income across all categories we will treat all income data as missing. If income data is missing in at least one but not all subcategories we will assume income is zero for missing subcategories.

The endline survey did not capture income data in the same way as the baseline survey. With three exceptions, these differences do not affect our ability to disaggregate income across the subindicators of interest. At baseline we did not collect data on the value of livestock and crops produced and consumed at home, so these baseline control variables will be missing. We are also unable to disaggregate remittances from other miscellaneous income, so the baseline control variable for both of those subindicators will be the sum of remittances and other miscellaneous income.

6.6 Expenditures

The expenditures income includes the sum of annual food and non-food expenditures. Subindicator categories are described as follows:

1. medical: formal and traditional medicine
2. apparel: men's, women's, and children's (including school uniforms) and
3. jewelry: non-ceremonial jewelry and watches
4. education: tuition and related expenses
5. home: telecommunications, utilities, fuel, personal items, household items and services, home improvement and maintenance, household durable goods, electronics, migration of household member, rent or mortgage
6. temptation goods: gambling, alcohol, tobacco
7. celebrations: weddings, funerals, festivals, dowries, ceremonial items, recreation and entertainment
8. charitable giving: gifts and donations

[10] from the livestock enterprises module of the survey
[11] from the crop enterprises module of the survey

9. services: insurance, banking, legal or other administrative services
10. travel: transportation and travel costs, private vehicle purchase, and any related vehicle expenses
11. food: value of food purchased
12. food: value of food consumed that was produced at home (while not technically an expenditure, we include it here because it is an important aspect of consumption that could be affected by the intervention)[12]

For expenditure categories not reported annually, we multiply the weekly, monthly or quarterly figures by the appropriate factor to achieve an annualized amount. If a respondent claims zero expenditures across all categories we will treat all expenditures data as missing. If expenditures data is missing in at least one but not all subcategories we will assume expenditures is zero for missing subcategories.

The endline survey did not capture expenditure data in the same way as the baseline survey. With the exception of services, travel and food, these differences do not affect our ability to disaggregate income across the subindicators of interest. In the subindicator analysis of service, travel and food (both purchased and food consumed/produced at home) expenditures, we will use no baseline control variable.

Given the small increases in livestock income and the extremely noisy nature of the expenditure data we decided not to use this outcome.

6.7 Food security & nutrition

The food security & nutrition index is the ICW average of a dummy variable equal to one if the respondent indicates all household members get enough to eat every day, a dummy variable equal to one if the household cut back on meals following a shock, the number of days the household consumed meat, fish or eggs, and a food consumption score (FCS).

We do not have baseline data for the number of days the household consumed meat, fish or eggs. Instead, we use the number of meals in the last three days the adults in the household consumed meat, fish or eggs.[13] Baseline does not accommodate calculating FCS, so we will use no baseline control variable in the subindicator analysis, and the baseline index will include only the three remaining subindicators.

Households getting enough to eat (almost always happens) and households cutting back on meals (almost never happens) did not have adequate variation to include. The FCS data was

[12]This data is from the food expenditures module, not constrained to be consistent with income value of livestock and crops produced and consumed at home.

[13]If the adults consumed eggs and fish in the same meal, this will count for two meals.

extremely noisy so instead we focus on meat and vegetable consumption, the two outcomes we felt were most likely to be impacted by the program given its focus.

6.8 Outcomes for analyzing mechanisms

In addition to the welfare analysis using the indices and subindicators described above, we will assess impacts on the following outcomes to analyze mechanisms. Since we are studying a livestock-based livelihoods intervention, most of the outcomes considered for the analysis of mechanisms are directly related to livestock, and more specifically goats. The indicators identified below consider goat herd dynamics, profit from goat production, and goat production practices. The intervention targets women, so we also consider decision-making over livestock production, which may lead to empowerment, which may in turn affect other welfare measures like food security and nutrition. Since the intervention includes training in entrepreneurship and encourages savings for investment, we assess investment in all entrepreneurial activities. Finally, the intervention seeks to alter the way individuals think about and plan for the future. To this end, we will assess impacts on future outlook, including aspirations. Although aspirations could be considered a welfare outcome, we consider them in our mechanisms analysis because they are most important if they lead to other behavioral changes, such as investment, which results in higher income.

1. Goat herd dynamics (no summary index)
   (a) herd size
   (b) number male (no baseline control)
   (c) number purchased
   (d) number born
   (e) number received as gifts/dowries
   (f) number received from Heifer
   (g) number given away
   (h) number died
   (i) number sold
   (j) number slaughtered for consumption (no baseline control)
   (k) growth rate (intake - offtake)

   We do not present findings for number of male goats born. Our reason to do this was to investigate heterogeneity in some outcomes based on the sex of goat offspring. The idea was that beneficiary households who had male goat kids would be able to sell the male kid for income, while households with female goat kids had an obligation to pay it forward. However, even absent the programs households tend to sell male kids

and keep female kids in the herd so we did not find this line of analysis particularly informative about program functionality. Since households had pre-existing herds, the outcome was quite noisy. We do not present findings for growth rate (intake-offtake) because it is just a sum of other outcomes and should be reflected in differences in herd size across treatment groups.

2. Goat profit: Summary index is revenue minus investment
   (a) Goat revenue
      i. value of goat sales
      ii. value of goats consumed at home in levels (no baseline control)[14].
      iii. value of goat products in levels (use value all livestock products at baseline)[15].

      We did not consider value of goats or goat products consumed. Households rarely consume goats and never consume goat milk.

   (b) Goat investment[16]
      i. cost of fodder
      ii. cost of veterinary care
      iii. cost of breeding fees
      iv. cost of improved shelter
      v. cost of marketing (no baseline)
      vi. total investment in goat production

      We present results for total expenditures only to reduce the number of outcomes examined. Total expenditures is the sum of i-v above.

3. Goat production practices (summary index is the I)CW average):
   (a) have improved animal shelter (dummy variable)
   (b) goat manure removal once per week (dummy variable)
   (c) goat manure removal at least once per month (dummy variable)
   (d) goat manure used for fertilizer (dummy variable)
   (e) used any medicine (dummy variable)
   (f) had goats vaccinatted against anything (dummy variable)
   (g) harvested home grown fodder (dummy variable)
   (h) used mineral block (dummy variable)
   (i) received any animal health worker visit at home (dummy variable), (no baseline control)

[14]We anticipate this being practically non-existent given households rarely slaughter their own goats

[15]This could include milk and hides. We anticipate this being practically non-existent given that goat milk production is rare and households rarely slaughter their own goats

[16]Baseline is missing goat-specific expenditures. Use livestock investment as baseline control variables.

(j) have access to CAHW in the community (dummy variable), (no baseline control)

4. Decision making over goats: summary index is a weighted standardized average
   (a) ownership (dummy variable)
   (b) decisions over care/maintenance (dummy variable)
   (c) decisions over sales/renting (dummy variable)
   (d) decisions over livestock income (dummy variable)
5. Investment in entrepreneurial activities: summary index is the sum of total investment
   (a) total livestock investment (goat and other livestock)
   (b) total crop investment(no baseline)
   (c) entrepreneurial investment (no baseline)

   We do not present these less proximate outcomes to limit the scope of analysis.
6. Future outlook: summary index is a weighted standardized average
   (a) income aspirations[17]
   (b) social status aspirations[18]
   (c) ordered categorical "patience" variable[19]
   (d) ordered categorical "planning horizon" variable[20]

   We do not present these outcomes to limit the scope of analysis. Four years after the program, we can expect any changes in future outlook would have resulted in behavioural changes impacting other dimensions of welfare (e.g., savings, investment) that are more easily quantifiable.

7 Cost-benefit analysis

7.1 Program costs

Program costs associated with the intervention fall into the following broad categories:

1. basic intervention costs

[17] The income and status aspirations-related subindicators are based on the subindicators used in Bernard and Taffesse (2014).

[18] We will use a count of the number of women within the same ward who might one day seek advice from the respondent, specifically, "In the future, how many women in this ward do you think might actually seek your advice?"

[19] Based on questions designed to elicit a discount rate following Ashraf, Karlan, and Yin (2006).

[20] Based on how many days individuals plan ahead following Laajaj (2017).

(a) technical trainings on nutrition, improved animal management, home gardening, fodder and forage production

(b) cash support for home gardens, fodder, and forage

(c) cash support for goat shed improvements

(d) provision of access to a community animal health worker

(e) administrative expenses, equipment, supplies (including those expenses associated with SHG mobilization)

2. livestock transfer costs

   (a) two doe goats

   (b) a buck of improved breeding stock to be shared among SHG members

3. values-based training costs

   (a) self-help group management training

   (b) gender justice training

   (c) HI cornerstones training (including encouragement to "pay-it-forward")

Cost per beneficiary in our sample varies by treatment arm. Some program costs are common or shared across all treatments. All three treatment types receive the basic intervention. Other costs are not incurred at all in certain treatment arms: the NG treatment arm incurs no costs for livestock, while NVT incurs no costs for values-based trainings.

Per beneficiary costs also vary by beneficiary type (direct or PIF). Some PIF costs are incurred by the NGO, but many costs associated with PIF are actually incurred by direct beneficiaries. (Recall that direct beneficiaries take on much of the responsibility for paying benefits forward —sharing livestock and knowledge to second generation beneficiaries and providing financial support for improved goat pen construction.) We consider only program costs incurred directly by the NGO. For our calculations, we assume there are no PIF beneficiaries in the NVT arm and therefore zero PIF costs because encouragement to pay it forward is not a program component.

Some costs are reported at the beneficiary level whereas others are reported at the SHG or ward level. Allocating costs per-beneficiary therefore requires us to clearly define the aggregate cost (numerator) and the number of beneficiaries among whom an aggregate cost is shared (denominator).

1. **Individual beneficiary costs** include the actual unit cost of each goat (if any) transferred to a direct beneficiary, as well as the value of cash support provided for home garden, fodder and forage. These unit costs map 1:1 to an individual cost per direct beneficiary.

2. **Costs shared directly across all beneficiaries (both direct and PIF beneficiaries)** include the costs of mobilizing and providing each SHG with access to a community animal health worker, and all administrative costs.

    - The numerator for community animal health worker costs per beneficiary is the reported sum of all costs to mobilize and provide community animal health workers. The denominator for per-beneficiary calculations is the total number of direct (all treatments) and PIF (FT and NG treatments only) beneficiaries.
    - The numerator for administrative costs per beneficiary is the reported sum of all administrative costs. The denominator for per-beneficiary calculations is the total number of direct (all treatments) and PIF (FT and NG treatments only) beneficiaries.

3. **Costs shared among all direct beneficiaries (but not PIF beneficiaries)** include the cost of direct technical trainings, and equipment and supply costs.

    - The numerator for technical trainings is the sum of all costs reported for technical trainings (recall that direct beneficiaries later incur the costs of technical training for PIF beneficiaries). The denominator for per-beneficiary calculations is the total number of direct beneficiaries (all treatments).
    - The numerator for equipment and supplies is the sum of all costs reported for equipment and supplies. The denominator for per-beneficiary calculations is the total number of direct beneficiaries (all treatments).

4. **Costs shared among a subset of direct beneficiaries** include the cost of a shared buck (applies only to FT and NVT), and the cost of values-based training (applies only to FT and NG).

    - The numerator for the cost of a shared buck is the sum of the cost of all bucks purchased for direct SHGs; The denominator for per-beneficiary calculations is the total number of direct beneficiaries in FT or NVT.
    - The numerator for the cost of values-based training is the sum of the cost of all values-based training allocated to direct beneficiaries. The denominator for per-beneficiary calculations is the total number of direct beneficiaries in FT or NG.

5. **Costs shared among all PIF beneficiaries (but not direct beneficiaries)** includes the cost of values-based training provided to all PIF beneficiaries (applies only to FT and NG).

    - The numerator for the cost of values-based training is the sum of the cost of all values-based training allocated to PIF beneficiaries. The denominator is the total number of PIF beneficiaries in FT and NG.

6. **Costs shared among a subset of PIF beneficiaries** includes the cost of a shared buck provided to PIF beneficiaries (applies only to FT).

    - The numerator is the sum of the cost of all bucks purchased for PIF SHGs. The denominator is the total number of PIF beneficiaries in FT.

In the end, all costs were reported at the VDC level, making it difficult to disaggregate costs by targeted and non-targeted households. In addition, some targeted individuals joined PIF groups and many non-targeted individuals joined original groups, further complicating cost disaggregation. For these reasons we did not disaggregate costs by targeted and non-targeted. This also makes it easier to deal with fixed costs for providing benefits to a village. All costs are per household, irrespective of whether that household was targeted or not, or whether they joined an original group, a PIF group, or no group.

Finally, we must define what qualifies as a beneficiary. Recall that we are estimating ITT impacts. To make the costs comparable to our estimated benefits, we will consider ITT costs, i.e. the cost per targeted beneficiary. For example, if an individual is targeted as a direct beneficiary but does not join a SHG, they are still a targeted beneficiary (even though the costs of the program are zero for them). Since ITT impacts are an average impact on those whom actually join and those who do not, costs will be calculated in the same way. The number of targeted beneficiaries will be obtained from the original sample frame (disaggregated by direct and PIF). The average number of direct beneficiaries in a ward is 22. The average number of potential PIF households is 104.

7.2 Program benefits

The proposed impact analysis includes non-monetary and monetary welfare impacts that vary by both treatment and beneficiary type. The cost-benefit analysis considers only monetary benefits. Some monetary impacts (expenditures and income) are estimated annually. Previous work revealed no statistically significant short-term impacts on expenditures or

income (Janzen et al., 2018), so we only consider impacts at endlines 1 and 2, noting that this is conservative.[21]

To better align with program costs, we calculate program benefits using a weighted regression to account for the fact that targeted beneficiaries are sampled with greater frequency than non-targeted beneficiaries.

The data provides several options for calculations of monetary benefits. We will use the following:

1. **Income**: Sum of total annual household income at endlines 1 and 2. We will assume impacts at endline 2 persist in perpetuity.

   Income is our preferred measure of cash flows to be used for cost-benefit analysis. It can be used on anything: consumption of durable and non-durable goods, cash savings, home improvement, and investment in livestock and other productive assets. However, it may be noisily measured because the survey asks for income over the course of a year.

   We do not use total household income given how noisy it is. Instead, we use goat profit which is precisely measured.

2. **Expenditures on non-durable goods**: Sum of total annual expenditure on non-durable goods, including food, at endlines 1 and 2. We will assume impacts at endline 2 persist in perpetuity.

   We do use total household expenditures for the same reason.

3. **Assets**: Sum of the value of durable assets, livestock, and current savings at endline 2. We note this is a conservative estimate of asset values because it omits land.[22].

   We only use goat value, as this is the only asset impacted by the program, and the data on the value of other assets is extremely noisy.

4. **Expenditures on non-durable goods plus assets**: Sum of 2 and 3 above. This approach is similar to that used by Banerjee et al. (2015) except for that we also include productive assets. Our rationale is that while the household does not enjoy the benefit of consuming an asset such as a sewing machine or a goat, the asset does have

[21] Income and expenditures were measured differently at midline relative to endlines 1 and 2. Expenditures at midline excluded food expenditures, and thus underestimate total expenditure impacts. The income module was also substantially revised with a goal of reducing noise. Sensitivity analysis could consider alternative assumptions about impacts between midline and endline 1.

[22] At the time of writing this plan, we do not have a precise way of assigning value to land owned. If we can confidently assign value to land owned, we will include value of land owned

value and will contribute to future income. It could also be liquidated in the event of a shock.

We do not use this outcome for the reasons mentioned above.

Of the above, we believe 1 or 4 will best reflect the entirety of monetary benefits of the program. If beneficiaries have begun selling more or better quality livestock as part of the program, benefits should be reflected in income. If households are increasing their herd size but have not yet begun selling more or better quality livestock, assets will better reflect monetary benefits, and this is captured in 4 but not 1 above.

For the above, non-durable goods include: medical expenses, apparel, education, telecommunications, fuel, personal items, household items and services, home improvement and maintenance (we acknowledge major renovations are durable goods but cannot separate them from small ones and maintenance), migration support, rent, mortgage, gambling, alcohol, tobacco, weddings, funerals, festivals, dowries, ceremonial items, recreation and entertainment, gifts and donations, insurance, bank fees, legal fees, other administrative fees, transportation and travel costs, vehicle maintenance and other expenses, food purchased, and value of food produced at home. This data will be taken from the expenditures and food consumption modules of the survey.

Durable assets include radios, cassette recorders, DVD players, televisions, satellite dishes, computers and printers, mobile phones, cameras and camcorders, bicycles, motorcycles or scooters, solar panels, batteries or inverters, fans, heaters, refrigerators or freezers, gas stoves, cupboards, jewelry and watches, tables, chairs, sofas, mattresses, plows, tractors, cars and trucks, grinders, threshers, looms, sewing inventories, mechanical tools, hand tools, and other productive assets. This data will be taken from the assets module of the survey, but we will use expenditure data to fill in missing or unreasonable values as needed when possible.

We only consider the value of the goat herd and value of goat sales in years 2, 3, and 4 plus a stream of expected future goat sales (we assume the household will continue to sell as many goats per year as they do in year 4).

Because monetary benefits will be estimated with error, and may not be significantly different than zero, we will report confidence intervals and include them in our discussion. To do this we will need to calculate 1, 2, 3, and 4 as described above for each individual in the sample and estimate treatment effects on those outcomes (e.g. we will estimate the treatment effect on calculated lifetime income). We will do this using different discount rates (described below) to provide several ranges of benefits.

We report standard errors for our benefits regressions in Table D1.

We will consider benefits accruing to both direct and PIF households in all three treatment groups. Whereas there are no program costs associated with PIF beneficiaries under the NVT treatment, it is possible that benefits accrue to individuals not directly targeted. Benefits to these households could arise either by joining the group (despite not being targeting) or otherwise integrating themselves into the program, or through spillover effects.

7.3 Benefit/cost ratios

The cost-benefit analysis obviously relies on assumptions regarding the discount rate. We will consider annual discount rates of 5, 10 and 20 percent. Costs and benefits will be considered from the time of endline 2, thus program costs and benefits from endline 1 will be inflated and future benefits will be deflated accordingly. We will also calculate the internal rate of return to assess at what social discount rate costs equal benefits.

We will calculate benefit/cost ratios for the VDC as a whole, which is a weighted average of direct and PIF benefit/cost ratios. We will do this separately for all treatment arms, unless pooling is necessary to achieve adequate power. We will also calculate the benefit/cost ratio for direct beneficiaries only in an effort to understand what the ratio would be for an asset transfer and training program without a pay-it forward mechanism or any non-programatic spillovers. We believe this is a better approach than using benefit/cost ratios from the NVT treatment, which does not include the pay-it-forward mechanism but also does not include other aspects of the values-based training.

As stated above, we use weighted regressions to calculate average benefits per household (targeted or non-targeted).

References

2005. *Livestock sector brief: Nepal*. Livestock information, sector analysis and policy branch: Food and Agriculture Organization of the United Nations, July.

Ahrens, A., C.B. Hansen, and M.E. Schaffer. 2018. "PDSLASSO: Stata module for post-selection and post-regularization OLS or IV estimation and inference." Working paper, Boston College Department of Economics.

Alkire, S., R. Meinzen-Dick, A. Peterman, A. Quisumbing, and G. Seymour. 2013. "The Women's Empowerment in Agriculture Index." *World Development* 52:71–91.

Anderson, M.L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103:1481–1495.

Ashraf, N., D. Karlan, and W. Yin. 2006. "Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines." *The Quarterly Journal of Economics*, pp. 635–672.

Athey, S., and G. Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113:7353–7360.

Bandiera, O., R. Burgess, N. Das, S. Gulesci, I. Rasul, and M. Sulaiman. 2017. "Labor markets and poverty in village economies." *The Quarterly Journal of Economics* 132:811–870.

Banerjee, A., E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parienté, J. Shapiro, B. Thuysbaert, and C. Udry. 2015. "A multifaceted program causes lasting progress for the very poor: Evidence from six countries." *Science* 348:1260799.

Banerjee, A., D. Karlan, R.D. Osei, H. Trachtman, and C. Udry. 2018. "Unpacking a Multi-Faceted Program to Build Sustainable Income for the Very Poor." Working paper, National Bureau of Economic Research.

Belloni, A., V. Chernozhukov, and C. Hansen. 2014. "Inference on treatment effects after selection among high-dimensional controls." *Review of Economic Studies* 81:608–650.

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300.

Bernard, T., and A.S. Taffesse. 2014. "Aspirations: An approach to measurement with validation using Ethiopian data." *Journal of African Economies* 23:189–224.

Bloniarz, A., H. Liu, C.H. Zhang, and J.S. Sekhon. 2015. "Lasso adjustments of treatment effect estimates in randomized experiments." *Proceedings of the National Academy of Sciences* 113.

Darrouzet-Nardi, A.F., L.C. Miller, N. Joshi, S. Mahato, M. Lohani, and B.L. Rogers. 2016. "Child dietary quality in rural Nepal: Effectiveness of a community-level development intervention." *Food Policy* 61:185–197.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár. 2022. "Contamination Bias in Linear Regressions." Working paper, National Bureau of Economic Research.

Janzen, S.A., N.P. Magnan, S. Sharma, and W.M. Thompson. 2018. "Short-term impacts of a pay-it-forward livestock transfer and training program in Nepal." *AEA Papers and Proceedings* 108:422–425.

Jodlowski, M., A. Winter-Nelson, K. Baylis, and P.D. Goldsmith. 2016. "Milk in the data: food security impacts from a livestock field experiment in Zambia." *World Development* 77:99–114.

Kafle, K., A. Winter-Nelson, and P. Goldsmith. 2016. "Does 25 cents more per day make a difference? The impact of livestock transfer and development in rural Zambia." *Food Policy* 63:62–72.

Laajaj, R. 2017. "Endogenous time horizon and behavioral poverty trap: Theory and evidence from Mozambique." *Journal of Development Economics* 127:187–208.

Lee, D.S. 2009. "Training, wages, and sample selection: Estimating sharp bounds on treatment effects." *The Review of Economic Studies* 76:1071–1102.

Lin, W. 2013. "Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique." *The Annals of Applied Statistics* 7:295–318.

Malapit, H., C. Kovarik, K. Sproule, R. Meinzen-Dick, and A. Quisumbing. 2015. "Instructional Guide on the Abbreviated Women's Empowerment in Agriculture Index (A-WEAI)." Unpublished.

McKenzie, D. 2012. "Beyond baseline and follow-up: The case for more T in experiments." *Journal of Development Economics* 99:210–221.

Miller, L.C., N. Joshi, M. Lohani, B. Rogers, M. Loraditch, R. Houser, P. Singh, and S. Mahato. 2014. "Community development and livestock promotion in rural Nepal: Effects on child growth and health." *Food and Nutrition Bulletin* 35:312–326.

Phadera, L., H. Michaelson, A. Winter-Nelson, and P. Goldsmith. 2017. "Do Asset Transfers Build Household Resilience?" Unpublished.

Radloff, L.S. 1977. "The CES-D scale a self-report depression scale for research in the general population." *Applied Psychological Measurement* 1:385–401.

Rawlins, R., S. Pimkina, C.B. Barrett, S. Pedersen, and B. Wydick. 2014. "Got milk? The impact of Heifer International's livestock donation programs in Rwanda on nutritional outcomes." *Food Policy* 44:202–213.

Rotter, J.B. 1966. "Generalized expectancies for internal versus external control of reinforcement." *Psychological Monographs: General and Applied* 80:1.

Wager, S., and S. Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113:1228–1242.

Table 1: Description of program components by treatment arm

Description of Program Components	**T1 (FT)**	**T2 (NG)**	**T3 (NVT)**
Basic intervention	x	x	x
SHG formation			
SHG savings encouragement			
training on nutrition			
training on improved animal management			
training and cash support ($5) for home gardening			
training and cash support ($10) for fodder & forage production			
cash support ($40) for goat shed improvement			
access to community animal health worker			
Productive asset transfer	x		x
2 doe goats			
1 shared buck of improved breeding stock (per SHG)			
Values-based trainings	x	x	
encouragement to "pay-it-forward"			
training on SHG management			
training on gender and justice			
*training on remaining HI Cornerstones**			

*The remaining HI Cornerstones not noted elsewhere in this table include: accountability; sharing and caring; sustainability and self-reliance; improving the environment; income; full participation; training, education, and communication; and spirituality.

Figure 1: Study timeline