Journal of Development Economics 155 (2022) 102780

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Regular Article

Social protection amidst social upheaval: Examining the impact of a multi-faceted program for ultra-poor households in Yemen☆

Lasse Brune [a], Dean Karlan [a,*], Sikandra Kurdi [b], Christopher Udry [a]

[a] Northwestern University, USA
[b] IFPRI, USA

ARTICLE INFO

JEL classification:
O12
I30
J24
F51
C93

Keywords:
Conflict
Social protection
Poverty
Transfer programs

ABSTRACT

Social protection programs are needed more than ever during periods of social upheaval, but are also likely to be even harder to implement successfully. Furthermore, social upheaval makes measuring the impact of such policies all the more difficult. We study the impact of a multi-faceted social protection program, often referred to as a "Graduation" model program, in Yemen during a period of civil unrest. We are unable to measure outcomes for four years, thus much remains unknown about what transpired in the intermediary time. After four years we find positive impacts on savings behavior and asset accumulation, albeit substantially less than the amount the household originally received.

1. Introduction

The poorest members of society are often chronically food insecure and lack stable income-generating activities. Conflict settings undoubtedly exacerbate such issues for most if not all, and conflict also may alter the effectiveness of social protection programs, whether simple cash transfer or more complex multi-faceted programs. Such efforts may be more effective if the program helps to mitigate negative consequences of the conflict; or may be less effective if implementation fidelity weakens or if the conflict creates constraints that render the program less effective. We examine the impact of a multi-faceted grant-based program, often referred to as a "Graduation" program, in Yemen during a period of civil conflict.

The core program combines short-term relief with a productive asset transfer, training, and ongoing support, and the design is predicated on a theory that sources of poverty are multi-faceted and intertwined, and thus solutions that aim to tackle multiple constraints are likely more effective. This theory is supported by prior randomized evaluations in Bangladesh, Ethiopia, Ghana, India, Honduras, Pakistan, and Peru, in which the program generated positive impacts that persisted after two and three years (Banerjee et al., 2015), as well as after four and seven years (Bandiera et al., 2017).

While this approach is adapted to each setting, all sites share several key elements. Each tested program begins by identifying the poorest households within a community. Selected households receive a productive asset to be used for generating income (such as livestock, inventory for petty trade, or sewing equipment), with concurrent training about how to profitably manage that asset. Households also receive consumption support, either in the form of cash transfers or food aid. Individuals are encouraged (and in some sites, required) to save in order to improve their resiliency to shocks. Finally, households receive regular coaching and mentoring throughout the implementation period.

☆ We thank Nate Barker, Caton Brewster, Callan Corcoran, Sami Horn, Hideto Koizumi, and Lalchand Luhana for excellent research assistance, and Nathanael Goldberg and Rachel Strohm for research management. We thank in particular Matt Lowes for coordinating the field work and project management in Yemen. We also thank the Social Fund for Development and the Social Welfare Fund for their partnership (in particular Lamis Al-Iryani, Arafat Alsalhy, Osama Al Shami, and Doaa Bahubaish); Essam Al-Fadhli and Husam Al-Sharjabi from Apex Consulting; Aude de Montesquiou, Syed Hashemi, and Mohammed Khaled at CGAP for their collaboration on data collection on implementation, as well as the Ford Foundation (in particular, Frank deGiovanni) and the UK Department for International Development for funding support. IRB approval from Yale University #1006006972. All errors and opinions are our own.

* Corresponding author.
E-mail addresses: lasse.brune@northwestern.edu (L. Brune), karlan@northwestern.edu (D. Karlan), s.kurdi@cgiar.org (S. Kurdi), christopher.udry@northwestern.edu (C. Udry).

https://doi.org/10.1016/j.jdeveco.2021.102780
Received 16 July 2020; Received in revised form 27 July 2021; Accepted 30 October 2021
Available online 24 November 2021

But the above-mentioned sites were all mostly stable settings, i.e., neither civil unrest nor conflict was present. Here we present results from a test of the same approach, but in a setting of civil war. Conflict could in theory alter the effectiveness of the program, in either direction. Such programs may generate even stronger welfare gains if they help households build more diversified income and extra assets that support their ability to manage risk or if they mitigate the need for labor and credit markets, thus providing households with a path to grow self-employment activities in a setting where labor and credit markets are not functioning well due to the civil conflict. Two recent studies are consistent with this theory. Chowdhury et al. (2017) examines the impact of a similar program in South Sudan. The study area was affected by a conflict partway through the program, and participants in the program were less likely than a control group to say that they were unable to invest in business because of the conflict and had 16% higher consumption after six months. This impact on consumption did not persist two years later, though participants retained higher levels of livestock assets and (weakly statistically significantly) higher livestock income. In Afghanistan, a setting with increasing levels of sporadic violence, a recent evaluation found strikingly large impacts of 30% on consumption one year after the end of the yearlong intervention. The Afghanistan intervention was notable for especially large levels of asset transfer (focused on cows rather than sheep and goats) and included on-going training and veterinary services, as well as replacement of sick or deceased animals during the period of the intervention (Bedoya et al., 2019).

Naturally, the impact of these multi-faceted, Graduation programs could also be worse in a civil conflict setting. The program is not designed to teach households how to flourish as autarkic subsistence households, but rather promotes market-level activities to generate cash income. If markets collapse, such engagement may not be viable, rendering those aspects of the program ineffective. Furthermore, participants may be even more vulnerable to shocks if the program has encouraged them to invest in businesses that are affected by conflict-related shocks at the expense of businesses that may have been less affected. Lastly, the impact may be smaller in a civil conflict setting as the fidelity of implementation may suffer, or the program may not be implemented at all if, for example, employees are not able to visit households regularly.

A further possibility has mixed welfare implications: the civil unrest may lead households to divest of the productive asset sooner than they would have without the civil unrest, which undermines the long-term aspiration to build a stable income source, but does provide the household with an effective tool to absorb the immediate shock from the civil unrest.

We use a randomized evaluation to examine the four-year impact of a Graduation program in Yemen, which was implemented by the Social Fund for Development (SFD) and the Social Welfare Fund (SWF) in the governorates of Aden, Lahj, and Taiz. The program targeted beneficiaries of the national cash transfer program (run by SWF) and used public lotteries to randomly choose a subset to participate in the Graduation program. Targeted households could choose from several types of assets, including sheep and goats, stock for kiosks, a sewing machine and materials for tailoring, or other goods. All households continued to receive consumption support in the form of cash transfers from SWF, so this does not distinguish the project beneficiaries from the comparison group. Beneficiary households received an initial training in how to profitably manage the enterprise they chose. They then received regular visits from SFD staff, which were meant to provide training, ensure that individuals did not simply liquidate their assets, and provide households with the encouragement needed to persist in the program.

Due to the political instability in Yemen, which started shortly after the baseline survey, our data collection efforts were cut short. We only have one follow-up survey, which was conducted in 2014, four years after the program began. Thus, while we are able to measure the living standards of our sample households after four years, we do not have data from the intervening period of unrest. This limits our ability to describe the path of impact over the four years. Further limiting our ability is a statistical power challenge, induced by imperfect compliance at the beginning of the project. The implementer, post-randomization, did a "validation" that removed 39% of households from treatment status. No control households were included in this process, unfortunately. We attempted to reconstruct this post-randomization selection process with an independent process in which control groups households were also reassessed. Unfortunately, the proportion identified as ineligible in the control was different enough that we deemed this reassessment unusable to reconstruct the sample for the study. Thus, we instead focus on intent-to-treat estimates throughout our analysis. This provides us with the average treatment effect on all households initially identified as eligible for the program, compared to the control group (regardless of whether they were re-verified and participated in the program). Aside from being lower powered for detecting average treatment effects, this also makes queries regarding heterogeneous treatment effects all the more difficult to answer.

We find modest positive results four years after the start of the program. Households selected into the program have a higher level of assets and savings, though this increase in wealth is substantially less than the value of the transfers received by the household four years earlier. We do not have precise estimates on per-capita consumption or household income, and thus can draw no conclusions for these outcomes. We find evidence to suggest increased participation in livestock rearing and slaughtering. We do not have precise enough evidence to suggest that borrowing or food security increased as a result of the program. We also observe a potentially important result: mortality is higher in the treatment group; we discuss both positive and negative reasons, as well as survey attrition, that may explain this mortality difference.

The long-run nature of the measurement and the intermediary crisis could be masking positive benefits from the program on resilience. A reasonable and more positive interpretation of the results would focus on the asset increase as evidence that, despite the political instability, the program was able to make important and long-lasting impacts on households, albeit at a high cost.

2. Setting, experimental design and data

2.1. Partners and site selection

The Yemen Graduation program was a new program created as a partnership between Yemen's Social Fund for Development (SFD) and Social Welfare Fund (SWF). SFD operates as a non-governmental organization, receiving regular funding and loans from the Government of Yemen but without direct oversight and control by the Government. Its main goal is to help alleviate poverty and reduce unemployment in Yemen through the implementation of targeted development projects. The SWF is a government agency tasked with delivering social protection and operates the national, unconditional cash transfer program.

About a quarter of households in Yemen included a SWF beneficiary as of 2012 (IIPC-IG and Unicef-Yemen 2014). While SWF beneficiary households are more likely to be poor on average compared to non-beneficiaries, the targeting performance of the SWF system is low due to the use of social categories for targeting (i.e. elderly, handicapped) and a lack of updating of listings. In 2012, only 44% of SWF beneficiaries were found to be in the poorest two quintiles of the population as measured with a wealth index based on household characteristics and asset ownership. Additional targeting, described below, was carried out to determine eligibility for the Graduation program.

The project took place in three governorates in southern Yemen—Aden, Lahj, and Taiz. The implementers chose these three governorates because they are home to a relatively large number of ultra-poor households, have an accessible terrain, and are in close proximity to one another.

2.2. Program eligibility and randomization

SFD identified eligible individuals for the program from the list of households in SWF's unconditional cash transfer program in project areas. As a result, both the treatment and the control households received consumption support, as all sample households were recipients of SWF cash transfers. SFD then surveyed 7300 individuals across 50 villages using the IPA's Poverty Probability Index (PPI),[1] with the intention of identifying approximately 1000 households for the program (i.e. about 14% of those surveyed were designated as eligible). Individuals were deemed ineligible if they received a PPI score above 40 or if they did not meet a number of criteria related to participation in existing programs and ease of logistics. The criteria were as follows: the household was not benefiting from any program run by a non-governmental organization, government agency, or microfinance institution; the household had a potential program participant who was willing and able to work and was between the ages of 18 and 60; the head of the household was not employed by the government; and the household was not nomadic.

In addition, SFD had the general aim of identifying the poorest members of society. Community leaders verified the selection and all eligible households received a visit from the program's management to check there was no erroneous targeting. A final list of 1002 eligible households was initially confirmed at the time of the baseline survey. On average at baseline, households had total monthly consumption of PPP $755 and 7.8 members (4.6 adults and 3.3 children). A female was the head of 31% of households. About 40% of the sample was below Yemen's National Poverty Line of PPP$2.7 per day in 2010 (YER 179; (Chen and Schreiner 2009). In our sample, 70% of households report having an adult who skipped a meal at some point in the last 12 months due to a lack of food, and 37% report having an adult who had gone a full day without food.

After compiling the final eligibility list, households were randomly chosen for the program using a public lottery in each village. In total, 505 households were assigned to the treatment group for participation in the Graduation program (i.e. about 7% of 7300 surveyed as part of the eligibility assessment), and 497 households were assigned to the control group. The control group did not receive any assets or training through the project but, as pre-existing beneficiaries of SWF, they still received their regular cash transfers. After the baseline survey and randomized assignment but (for most part) before the asset-transfer –a key component of the program– the implementing agencies conducted an additional verification exercise which resulted in the exclusion of nearly 40% of treatment households. We discuss the reverification in detail below.

Among treatment households one member was designated as the participant of the various program components. Unfortunately, we do not have reliable information on the identity of the target participant. Given the household eligibility criteria, almost all participants should be between 18 and 60 years old. In addition, the implementer reported that for 62% of households, the individual who attended an initial classroom-based training (see next section below) was female. As is the case in other Graduation programs, while there is an individual identified for training, the household as a unit is often considered the recipient, and often many share responsibility for the new livelihood activity.

2.3. Graduation program

The Yemen Graduation program consisted of four main components:[2]

- Enterprise development training
- Productive asset transfer
- Encouragement to save
- Education in social awareness, health care, and financial management

Once households were chosen for the program, SFD provided each household with an asset to help jump-start economic activity. The household had the option to choose, based on preference and past experience, from a list of livelihood options. The livelihood options included both agricultural (sheep and goat rearing) and off-farm activities (petty trade, tailoring, barber shops, ice cream vending, etc.). The implementation team ensured that the chosen activity had the potential to be economically viable, was easily manageable, and was socially acceptable within the communities served. The average cost per participant of each project was 70,000 Yemeni riyals (US$327, or US $963 in Purchasing Power Parity (PPP) terms). Livestock rearing was the most commonly chosen activity, with 75% of the beneficiary households choosing this option. The remaining 25% selected petty trade or other business activities.

Prior to the asset transfer, SFD provided participants with initial classroom-based training specific to their chosen means of income generation. The aim of the training was to help them acquire the skills needed to manage their assets or small businesses, and to teach them best practices. Furthermore, they were given information on where to obtain assistance and services from SFD should they need further help. The length of the training varied according to the enterprise, but typically took around three days. For more specialized enterprises, such as hairdressing and tailoring, training took up to one month.[3]

While the asset transfer leads to a substantial increase in the households' asset wealth in the immediate-term, the underlying theory of the program is that people will increasingly engage in the encouraged income-generating activity (i.e., livestock or micro-enterprise). Key outcomes are therefore (a) whether the asset base is maintained over time; and (b) whether revenues and incomes increase due to these assets. Ultimately, the program aims to increase living standards in the form of per capita consumption.

As pre-existing SWF beneficiaries, participants also received 6000–12,000 Yemeni riyals (PPP$82–165) per quarter in consumption support from SWF.[4] This safety net provides a predictable source of income to participants and helps them stabilize their livelihood.

Participants were encouraged to participate in saving schemes, either formally (e.g., through postal savings), or informally (e.g., through "hakba" savings clubs or in a secure saving box). Households were encouraged to save 100–150 riyals (PPP$1.37–2.06) a month to foster financial discipline and build-up their asset base to better cope with shocks and emergencies. Due to the 2011 political crisis (described further in the next section), this component was not fully executed, with only 104 participants reporting having saved. Furthermore, the bulk of savings we observe is on an informal and intermittent basis. We consider savings another key outcome of the program, given that they improve the ability of households to finance expenses and mitigate shocks.

[1] https://www.povertyindex.org/country/yemen; at the time of the data collection the index was called "Progress out of Poverty Index" and was managed by the Grameen Foundation.

[2] There were also plans to provide a link with microfinance after the end of the program, but this did not happen.

[3] The livestock training did not require any basic literacy skills because the training was focused on teaching participants how to tend to their livestock (in terms of feeding, vaccination, etc.). For trade activities, however, participants who lacked basic literacy skills were advised to bring another family member who was able to read, write, and understand basic math in order to teach families some basic principles of running a business. We do not have evidence suggesting that the brining of family members happened at a high enough rate to alter the interpretations of our results; but if it did happen systematically, evidence from, for example, Field et al. (2016) suggests this could be important.

[4] The levels of support were YER 6,000, 7,200, 8,400, 10,800 and 12,000 (USD PPP 82, 98, 115, 148, and 165, respectively). The amount each household received mainly depended on their number of dependents.

SFD field workers were hired for the program, with the intent of providing regular monitoring, coaching, and skills training for participants throughout the project. In addition, the implementation team was tasked with organizing weekly visits to the households to raise awareness on a number of social concerns, such as early marriage, and to educate participants on health issues, such as vaccination, qat use, water, sanitation, and hygiene. While the program was designed with the aim that SFD staff would meet with beneficiaries on a regular schedule, it is worth noting that we are unable to directly report the extent to which this coaching was successfully implemented, especially in light of the political crisis (see next section).

It is important to note that both treatment and control households received quarterly consumption support from the SWF fund (participation in SWF's unconditional cash transfer program was the first criterion establishing eligibility for the program tested here). Thus, the comparison between the two groups only measures the total impact of the *additional* components of the Graduation program: asset transfer, livelihoods, coaching, heath and community-building trainings, and savings. Most other evaluations within the seven country evaluation of Banerjee et al. (2015), with the exception of Ethiopia and to some extent Peru, evaluate the impact of the full program (including consumption support). The Yemeni case provides evidence on the impact of the remaining components, given the presence of preexisting consumption support for everyone in the study. This becomes particularly important when considering the program's cost effectiveness.

2.4. 2011 political crisis

The Yemen Graduation program was launched in 2010, just prior to the Arab Spring and the accompanying political crisis in Yemen. Since the late twentieth century, Yemen has been one of the poorest countries in the Middle East, with a weak central government and high levels of instability. In January 2011, demonstrations erupted in the major cities of Sana'a, Aden, and Taiz. The situation deteriorated into a period of active armed conflict as a major tribal alliance began supporting the opposition. Former President Ali Abdullah Saleh fled to Saudi Arabia and was eventually convinced to resign in November 2011, at which point his former Vice President, Abdrabbuh Mansour Hadi, became Acting President pending the drafting of a new constitution by an internationally supported National Dialogue Conference. Significantly for the Graduation program participants, the delivery of the SWF transfers for most beneficiaries was suspended for five quarters from April 2011 until September 2012. While beneficiaries received a large lump-sum that included the missing payments when the transfers were resumed, the lack of regular delivery during the crisis period was a challenge for poor households (Unicef Yemen and IPC-IG 2014; Moqueet 2013), the period from 2012 to 2014 (the year of our follow-up survey) was relatively peaceful and allowed for economic recovery after the disruptions of the 2011 revolution. There was, however, continued conflict with al-Qaeda during this period, taking the form of frequent bombings of government installations in major cities and oil pipelines, as well as clashes and drone strikes in some of the more isolated rural areas where al-Qaeda is based, including areas near the study's intervention sites according to the UCDP Georeferenced Conflict Event Dataset (Sundberg et al. 2012). Additionally, the large share of Yemeni households reliant on remittances from migrant labor experienced a significant economic shock as hundreds of thousands of Yemenis working in Saudi Arabia were expelled (Economist 2013). The National Dialogue Conference failed to reach a final agreement among the various parties and in September 2014 the Houthi seized control of the government in Sana'a. A coalition led by Saudi Arabia intervened in 2015 to attempt to re-install the Hadi government, precipitating the ongoing civil war.

A baseline survey was completed in July 2010. Enterprise selection started immediately after the baseline survey and continued until January 2011. Fig. 1 summarizes the timing of the project components and the political instability in Yemen at the time.

2.5. Impact of the crisis on program implementation

Despite the crisis, the implementing teams maintained that they were able to provide services. This was confirmed with the monitoring data collected at endline: 229 out of 287 households in the program that were interviewed at endline said they received regular visits from SWF or SFD, at least initially, while three households said they did not receive regular visits (note however that this question was unanswered by 55 households). Out of those who reported regular visits initially, only 25% reported that visits stopped at any point due to political instability. Among beneficiaries, 231 households received training from the implementing partners, while two reported not receiving this training (and for 56 households we do not have responses for this question). The interruptions of the regular household visits may have reduced the effectiveness of the intervention. A study of a similar program in Uganda tested the importance of household visits and found that a lack of regular supervisory visits were associated with significantly lower business survival (Blattman et al., 2016). Internal communications with SFD suggest that during the crisis visits were sometimes canceled due to road closures and unrest. There were also delays in distributing consumption support. Despite these challenges, field staff were able to travel to visit program beneficiaries with some regularity and internal monitoring suggested that few beneficiary households sold off their assets or lost their capital.

Even if the program had been implemented perfectly despite the crisis, the crisis may have created new constraints that households do not normally face. For example, the most direct effect of the crisis was the increase in transport costs due to the lack of security and the proliferation of checkpoints that demanded toll payments and caused long delays. Attacks on the pipelines also led to shortages and high black market fuel prices, further increasing transport costs as well as the cost of pumping ground water for agriculture (Reuters Staff 2013). Other evaluations of household welfare covering this time period found that employment in the private non-agricultural sector declined by half during 2011 (Christian et al. 2015) and proximity to conflict events during 2012 and 2013 was associated with significant decreases in child anthropometric status (Ecker et al. 2021). For program participants, the high transport costs meant that prices for inputs could be higher and prices for outputs could be lower (if middlemen needed to transport them elsewhere to sell), and prices could be more variable over time and space. Increases in food insecurity also gave households an incentive to consume the assets provided by the program.

2.6. Reverification of participants

Initial screening of households for program eligibility was done using the Progress out of Poverty Index (PPI), a country-specific proxy means test based on 10 survey questions that are used to identify the likelihood that a household is below the national poverty line. As described further above, households were generally quite poor both in terms of consumption and in terms of food insecurity at the time of our baseline survey. However, the baseline data showed that despite the high incidence of poverty, a substantial share of individuals chosen from SWF's initial list and SFD's further screening had income levels and asset holdings that suggested that not all of them were among the poorest of the poor.

Several factors caused the program implementers to doubt the effectiveness of the targeting process. First, the baseline report showed that the averages for water, power, and gas usage for beneficiaries in two of the three governorates of this project (Lahj and Aden) were higher than expected. Second, some beneficiaries were found to have received recent loans from the Social Welfare Fund or Al-Amal Microfinance bank. Third, a few beneficiaries in some villages seemed to be selling their assets immediately after receiving them. The fact that the selected

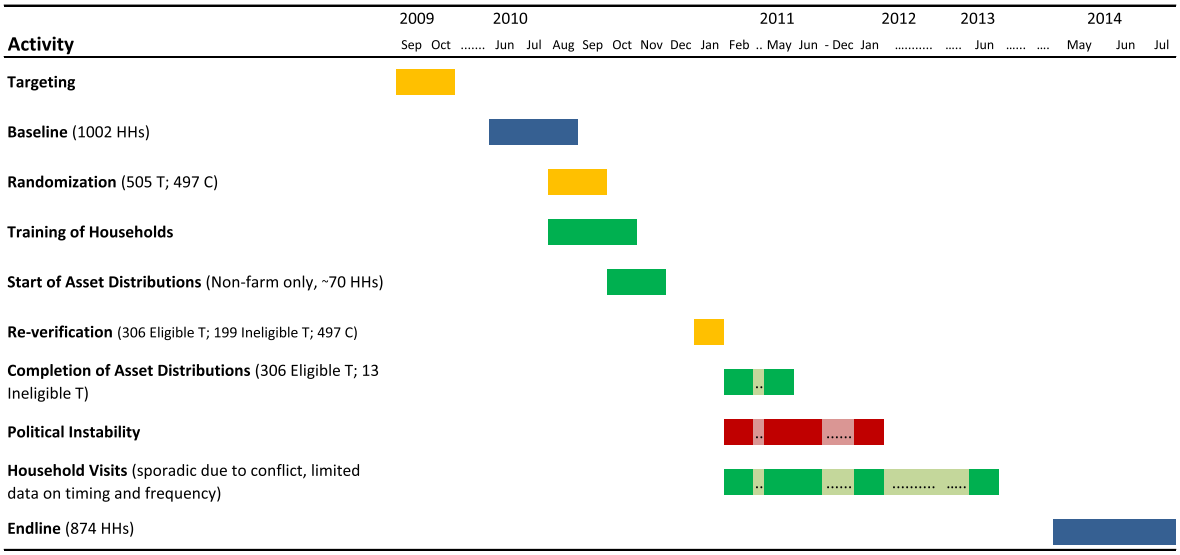

Fig. 1. Timeline of program actvities and surveying.

participants were relatively well off was likely a result of known issues with targeting within the SWF system from which the eligible households were selected, so that the selected households were among the poorest in the SWF system, but not necessarily the poorest in their communities.[5] The implementers decided to re-screen households and 199 people from the treatment group were determined to be *ineligible* for the program after the re-evaluation process was completed.[6] The rescreening was conducted after the assignment to treatment and control groups. Thus, some households who were originally identified as eligible for the program and were randomly selected into the treatment group did not actually receive the treatment. Unfortunately, an identical re-evaluation process was not carried out for the control group. Instead, several months later, a separate process was put in place to attempt to recreate the rescreening process. However, the proportion identified as ineligible among the control group was considerably different and, as a result, we do not have a reliable way to model selection into receipt of actual treatment.

Households that were excluded during reverification have, on average, more children, larger houses, are more likely to own land and to have a loan from a formal source and have more durable assets. The differences are not necessarily large economically, but they are all statistically significant (see Appendix Table 1). The households do not differ statistically significantly in total household size or in likelihood of owning any livestock at baseline.

Since we present intent-to-treat (ITT) estimates, the fact that many treatment households were removed from the program reduces our power to detect the impacts of the program. For example, the ITT minimal detectable effect (MDE) for per-capita consumption is 14.5% of the mean of the control group.[7] This number is equal to the minimum effect of the program on those that participated that we could have expected to detect had 100% of treatment households been eligible. Given that 39% of treatment households were ineligible after rescreening, the MDE of the program *on those who participated* is 23.8%, calculated from the ITT MDE scaled by 164% (=1/0.61), the inverse of the share of those who were eligible after rescreening.

2.7. Data collection

Initially, 1002 households were considered eligible to participate in the project. We conducted a baseline survey with all eligible households prior to the start of the program, which included information about assets, health status, land use or access, livestock, and business activity, among other things.

An endline was originally planned for 2012, two years after the start of the program, but was delayed due to the violence and insecurity associated with the Arab Spring. Instead, households were re-surveyed for an endline in May and June 2014, four years after the baseline survey was completed and one year after the intervention ended.,[8][9] A total of 874 households were found and surveyed at endline.

2.8. Survey attrition

The endline survey response rate was 87%. Given the gap of four

[5] The SWF system was originally designed based on inclusion of social categories such as the elderly and disabled, and a shift towards targeting the poorest households via a Proxy Means Test was never fully implemented due to the challenge of removing existing non-poor beneficiaries from the rolls (Unicef Yemen and IPC-IG 2014).

[6] Seven households sold their assets immediately upon receipt, triggering the re-evaluation of beneficiaries, most of whom had yet to receive their assets. The process was informal and based on an interview and observation by the implementers. The program implementers removed households from treatment prior to the transfer based on four criteria: selling asset, taking a loan, being insufficiently poor, or qualitatively appearing unwilling to participate in the program. Poverty status was based on observables, including observable assets and housing construction (floor material, roof material, etc.), size of SWF stipend, and whether the household owned or rented their house.

[7] For 80% power, with a test size of 5%, a control group of 497 and a treatment group of 505, the MDE is 0.177 standard deviations. The mean and standard deviation of per capita monthly consumption in the control group is PPP$91.94 and PPP$75.15, respectively.

[8] For the endline survey, as with the baseline, Innovations for Poverty Action (IPA) contracted Apex, a Yemeni firm with local staff, to carry out the data collection. IPA provided training to Apex staff members on the endline survey instrument and conducted quality checks on the incoming data. The initial training on the instrument was held in Turkey given the ongoing security concerns. Although on-site supervision was not feasible for this survey round, the use of electronic data collection allowed IPA to routinely monitor incoming data to check for inconsistencies and other errors.

[9] The endline survey had only limited overlap with Ramadan: the survey endline took place between May 20th to July 4th; Ramadan in 2014 was from June 28th to July 29th. As a result, only about 3% of observations were collected during Ramadan and the vast majority were collected *before* Ramadan.

Table 1
Endling survey response rate, baseline summary statistics and joint orthogonality tests.

Sample: Surveyed in Baseline (Panel A); Surveyed in Endline (Panels B & C)				
	(1)	(2)	(3)	(4)
	Control mean (S.D.)	Treatment mean (S.D.)	Obs.	p value of H0: (1)=(2)
Panel A: Survey response rate				
Endline survey response rate	0.85	0.89	1002	0.03
Panel B: Household demographics and health				
Num. of adults (≥18 years)	4.43 (2.42)	4.78 (2.50)	874	0.07
Num. of children (<18 years)	3.37 (2.46)	3.22 (2.36)	874	0.29
Household Size	7.78 (3.32)	7.79 (3.11)	874	0.84
Average age of adult household members	24.30 (8.94)	25.09 (8.42)	874	0.22
Avg yrs of schooling of adult hh mem.	4.10 (2.83)	4.38 (2.60)	874	0.27
Household head:				
Age	50.17 (12.64)	51.75 (13.86)	831	0.18
Age>60	0.17 (0.38)	0.23 (0.42)	831	0.06
Female	0.32 (0.47)	0.32 (0.47)	868	0.80
Years of schooling	2.47 (3.92)	2.27 (3.84)	868	0.30
Work impeded by illness or disability	0.43 (0.50)	0.41 (0.49)	868	0.53
Panel C: p-values of joint orthogonality tests				
Baseline variables from Panel B above = 0				0.32
Baseline variables from the primary outcomes in Table 2 = 0				0.59
Baseline variables from the secondary outcomes in Table 3 = 0				0.69
Baseline variables from Panel B above, Table 2, and Table 3 = 0				0.69

Notes: Randomization was stratified by village. p values are based on regressions that include a full set of village indicators.

years between baseline to endline, and the political upheaval in Yemen, we consider this a high resurvey rate. Importantly, however, the response rate is 4 percentage points higher in the treatment group than in the control group and the difference is statistically significant (Table 1, Panel A, $p < 0.05$).[10] However, we do not find evidence that baseline variables predict attrition differentially across treatment and control groups (Appendix Table 2, p-value of joint test = 0.39). Nevertheless, we conduct robustness analysis following the presentation of our main results to explore the potential bias stemming for selective survey attrition (see Section 4).

2.9. Randomization balance

We present orthogonality analysis in Table 1 for household demographics and baseline health (Column 4) and in Tables 2 and 3 for all outcome variables (Columns 10 and 6, respectively; in addition, summary statistics for the baseline values of outcome variables by treatment status are shown in Appendix Table 3). In univariate analysis, we reject equality of means for two of the ten baseline demographic variables (Table 1) and two of the 24 primary and secondary outcomes (Tables 2 and 3). Table 1 Panel C reports the joint tests for orthogonality, and we fail to reject orthogonality for all tests: all demographic and baseline health variables (p = 0.19), primary outcomes from Table 2 (p = 0.59), secondary outcomes from Table 3 (p = 0.69), and all of the above (p = 0.69).

3. Results

To measure the intent-to-treat (ITT) impact of assignment to the Graduation program, we estimate OLS regressions of the outcomes on the treatment group indicator and the baseline value of the outcome variable if available:

$$y_{i,t=1} = \alpha + \beta_0 + \beta_1 T_i + \beta_2 y_{i,t=0} + v_j + \varepsilon_i$$

where $y_{i,t}$ is the outcome variable of interest for household i at baseline *(t=0)* or endline *(t=1)* and v_j are stratification cell (i.e., village) fixed effects. The coefficient β_1 captures the average effect of being assigned to participate in the program. Under the assumption of no spillover effects from treatment households to control households, we can interpret the coefficient as the causal effect of being assigned to the treatment group. We discuss the implications of potential bias due to spillovers for the interpretation of our results in a separate section following the description of our results. Under the assumption of no spillovers, we can also scale the ITT estimates from the treatment effect tables –both the coefficient estimates and the standard error estimates– by 164% (=1/0.61) to compute the estimate of the Treatment on the Treated (TOT) that accounts for the fact that 39% of the treatment group did not receive the program.

To adjust for multiple hypotheses testing, following Banerjee et al. (2015), we calculate a *q* value: the minimum false discovery rate (i.e., the expected proportion of rejected null hypotheses that are actually true) at which the null hypothesis would be rejected for that test, given the other tests run on other outcomes in the same family (Anderson, 2008; Benjamini and Hochberg, 1995). We compute *q* values for the primary outcomes in Table 2 but not for secondary outcomes, because analyses on the latter are exploratory.

3.1. Primary outcomes

Table 2 presents results for the key welfare outcomes that the program aimed to improve.

We find a large and statistically significant increase in total assets (PPP$290, se = PPP$67, with a control group mean of PPP$744) as well as in an index of savings outcomes (0.44 sd increase, se = 0.12), but do not find an accompanying positive treatment effect on either consumption or income. These results are robust to adjustments for multiple hypothesis testing (*q* values < 0.01 for both total assets and the savings index). Importantly, we can rule out effects of consumption and income that are similar in magnitude to that on assets; the confidence interval for change in monthly consumption per household ranges from negative

[10] Attrition took place for the following reasons: 29 treatment and 35 control households declined to be interviewed; seven treatment and five control households relocated; six treatment and six control households dissolved; two treatment households were too old to participate; one treatment household was not found due to a member's death; and one control household was travelling at the time of survey. For another 11 treatment and 24 control households, we do not have a recorded reason.

Table 2
Treatment effects on key welfare outcomes.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)		(9)		(8)
	Coef.	(S.E.)	Control Mean	Control S.D.	Obs.	p value	q value	Appendix table # with details		P-values: baseline balance test		q value (FWER)
Total asset value (PPP$)	290.15***	(66.97)	743.56	922.05	874	<0.01	<0.01	4		0.97	†	<0.01
Monthly consumption per capita (PPP$)	−2.22	(4.80)	91.94	75.14	874	0.64	0.93	5A		0.88		1.00
Monthly consumption per HH (PPP$)	9.32	(32.07)	616.36	483.78	874	0.77	0.93	5B		0.61		1.00
Total income, past 12m (PPP$)	−29.08	(342.90)	1406.68	4532.56	847	0.93	0.93	6 & 7		0.57		1.00
Livestock income, past 12m (PPP$)	−20.09	(34.53)	−70.73	484.30	874	0.56	0.93	6		0.07	†	1.00
Non-livestock income, past 12m (PPP$)	42.79	(338.43)	1411.33	4439.80	847	0.90	0.93	7		0.81	†	1.00
Food security index (z-score)	0.04	(0.07)	−0.00	1.00	874	0.55	0.93	8		0.58		1.00
Savings index (z-score)	0.44***	(0.12)	−0.00	1.00	874	<0.01	<0.01	9		0.17	†	<0.01
Perceived economic status (1–10)	−0.04	(0.14)	3.68	2.23	864	0.76	0.93	–		0.45		1.00
Housing index (z-score)	0.11*	(0.06)	−0.00	0.98	874	0.09	0.32	10		0.25		0.78
Debt index (z-score)	−0.05	(0.05)	−0.02	0.86	874	0.39	0.93	11		0.14		1.00

Notes: Regressions with the components of each outcome variable in this table are shown in Appendix Tables 3 through 10. Regressions include as controls a full set of village indicators (= level of stratification of randomization) and the baseline value of dependent variable or the closest proxies available. Missing values of dependent variables at baseline were replaced by zero and indicators for missing observations are added to the regression as controls. Variables are winsorized at the 99th percentile. Column 9 shows p values of test of equality of means of baseline values of the outcome variables. † denotes outcomes for which the balance tests was done with the closest baseline proxy available; specifically, for "Total asset value" we use an asset index, for "Livestock Income" we use an indicator variable for any livestock ownership; for "Non-Livestock Income" we use an indicator non-farm income over the prior 12 months; for "Savings Index" we use an indicator of whether someone in the household has a savings account. * denotes statistical significance at the 10-percent level; ** at the 5-percent level; and *** at the 1-percent level.

PPP$54 to positive PPP$72 (with a control group mean of PPP$616), which means that the upper bound equals a 12% increase in consumption in the treatment group relative to the control group, compared to a 39% increase in assets. We discuss this in more detail below.

Examining the components of total asset value (Appendix Table 4), we find the increase in assets comes from productive assets, particularly livestock and, to a lesser extent, agricultural tools and structures. The large share of the livestock impact as a proportion of the impact on total asset value lines up with the choices of participants for their program asset transfers, where 75% chose livestock. The increase in agricultural tools and structures, our best measure of assets not given out as part of the program, is in line with the idea that the program might have helped households start diversifying their income sources by investing income generating activities beyond the set of activities directly supported by the program.

Note that the average value of the asset transfer of the program was PPP$963 (US$326) per participant. Since only 61% of the treatment group actually received the program after reverification, the cost figure that is comparable to the ITT—i.e. the cost per treatment group member—is PPP$587 and so the measured effect after four years on total asset value is about 49% of the value of the initial transfer. This could reflect the selling or consumption of some assets over the four years between the initial transfer and the endline; the positive impact on the housing index may be evidence of spending of this type. In addition, measurement error and potential changes in prices over time complicate the comparison of our survey measure based on self-reports with the known cost of the initial transfer. Sidestepping the issue of prices, we can examine the treatment effects on the *number* livestock owned, the asset type chosen by the majority of participants (Appendix Table 15). Adding up the treatment effect for sheep and goats, treatment households had 1.7 additional animals on average at endline, which is almost exactly the same as the 1.8 animals initially transferred on average per household in the treatment group (4 animals per treated household x 61% recertification rate x 75% choosing livestock). This leaves open the possibility that lower value of assets as measured at endline could either be driven by non-livestock assets or by issues of measurement regarding the valuation of assets.

Treatment households also increased savings activities as captured by a savings index that combines balance and deposit behavior. Ultra-poor households are often characterized by extreme financial vulnerability; they tend to have low and irregular income and are unlikely to have formal savings to draw from in case of emergencies or for investment. In Yemen, one of the core components of the Graduation program was encouragement to save and the monitoring of records of savings by field officers. At the time of the baseline survey, only 1.1% of the entire sample reported having any kind of saving so it is reasonable to expect that the inclusion of a savings component would have visible effects on the savings habits of treatment households. This can be clearly seen in the endline results, indicating that the savings component did in fact change savings habits; treatment households have a higher monthly average savings balance and report higher levels of deposits in the three months prior to the survey (Appendix Table 9). In addition to the results on savings, the housing index, another primary outcome, increases by 0.11 standard deviations (se = 0.06), although this result is not robust to adjusting for multiple hypothesis testing (q value = 0.32).

However, treatment households did not experience improvements in other key, primary welfare outcomes shown in Table 2, namely income, food security, and subjective economic status. Households also did not change their overall debt activity. But we do observe changes in income generating activity that is consistent with the higher levels of productive assets. Specifically, livestock activity increases: we find positive effects on livestock revenue and expenditures, but not net income (Appendix Table 6). These results are important since increases in income and diversification of income sources were two goals of the program. In other income components we see no statistically significant positive effects (Appendix Table 7). However, for some of them we cannot rule

Table 3
Treatment effects on shocks, household composition, mortality and travel.

	(1)	(2)	(3)	(4)	(5)	(6)
	Coef.	(S.E.)	Control Mean	Control S.D.	Obs.	P-values: baseline balance test
Panel A: Shocks in past 12m						
Had any shock	0.03	(0.03)	0.55	0.50	874	0.41
Had livestock shock	0.07***	(0.02)	0.03	0.18	874	0.07
Had shock and used […] to finance coping						
Savings	0.04*	(0.02)	0.13	0.34	874	0.11
Loan	0.00	(0.03)	0.25	0.43	874	0.90
sale of livestock or crops	0.00	(0.01)	0.02	0.14	874	0.89
Panel B: Household composition						
Household size	0.19	(0.16)	7.48	3.50	874	0.84
Average household education (adults)	0.08	(0.11)	4.37	2.88	874	0.27
Average household age	−0.70	(0.48)	28.58	10.27	874	0.22
Hh has new members since baseline	0.01	(0.03)	0.44	0.50	874	–
Number new members since baseline	0.14	(0.11)	0.91	1.48	874	–
New household head since baseline	−0.01	(0.01)	0.02	0.13	874	–
Panel C: Mortality						
Hh head died since baseline	0.06***	(0.02)	0.05	0.22	874	0.50 †
Number of members since baseline …						
who died	0.06*	(0.03)	0.17	0.40	874	0.95 †
present at baseline who died	0.06**	(0.03)	0.16	0.40	874	–
who died of an illness	0.01	(0.02)	0.06	0.23	874	0.96 †
who died from an accident	−0.01	(0.01)	0.02	0.13	874	0.86 †
Average age of members who died	1.58	(5.67)	49.35	27.58	158	0.71 †
Panel D: Travel						
Number of members who travelled	0.15**	(0.07)	0.48	0.89	874	–
Proportion of travelling members	0.02**	(0.01)	0.08	0.15	874	–
Number of members travelling for work	0.07**	(0.03)	0.21	0.50	874	–
Proportion of members travelling for work	0.02***	(0.01)	0.03	0.08	874	–
Panel E: Transfers						
Crops given to other HH last season ($PPP)	−0.11	(0.13)	0.28	2.30	874	–
Remittances received in past 12m ($PPP)	0.18	(0.16)	0.31	1.99	874	–

Notes: "Hh" = household. Variables are top winsorized at the 99th percentile. * denotes statistical significance at the 10-percent level; ** at the 5-percent level; and *** at the 1-percent level. For additional notes, see Table 2. † balance tests were done with the equivalent variables over the past 5 years prior to the baseline interview.

out substantively important changes. For example, we see large point estimates relative to the control mean for business profit (+26%) and expenses (+21%) – but less so for revenue (+8%) and confidence intervals are wide for all three business outcomes.

3.2. Secondary outcomes

Table 3 presents results on a second set of outcomes, encompassing experiencing and coping with shocks; household composition; mortality; travel; and transfers.

The Graduation program aims to improve the resiliency of program participants—to enable households to respond to shocks in ways that do not harm long-term investment and income generating activities. While treatment households do not have a statistically significantly higher incidence of experiencing a shock of any type in the past 12 months (Table 3, Panel A), they do report a 7 percentage points higher likelihood of livestock shocks (se = 2pp). The latter is likely a mechanical by-product of having more livestock (Appendix Table 4) and thus should not be interpreted as a negative consequence per se. We examine households' ability to cope with shocks and find a 4-percentage-point increase (se = 2pp) in the likelihood of having a shock in the past 12 months *and* using savings to cope with it. On net, it is ambiguous whether this a good or bad outcome. Taking on more risks, when coping strategies are viable and risk is positively correlated with expected returns, can lead to higher and more sustainable long-term income. Of course, higher risk without compensating higher returns is bad. The effects here are not large enough, nor are the data granular enough, nor is there is enough variation in relevant observable states of the world, to be able to ascertain whether the net effect here on shocks and coping is evidence of positive or negative impacts.

Importantly, we find an increase in mortality in the household (Table 3, Panel C). Treatment households are 6 percentage points (se = 2pp) more likely than the control group to have seen their household head die since the baseline four years earlier.[11] The control group mean is only 5 percent; thus a 6-percentage-point increase represents a doubling of the mortality rate of the household head.

We posit a number of possibilities for why we measure an increase in mortality for the household head, some of which imply a negative impact on welfare while others are ambiguous. First, the mortality effect could be due to increased economic activity that led treatment households to leave the house and village more often to trade and do business. To examine this, we look at data on travel incidence (Panel D) and indeed find higher travel rates for treatment households. We do not have travel data broken down by household member, however, which would be useful to pinpoint whether travel could be responsible for the increased mortality of the household head in particular. Increased economic activity can only be the driver of household head mortality to the extent that the household head was more economically active, and in many cases the household head was not the primary program participant. In addition, we do not find an increase in the number of household

[11] This outcome variable is equal to 1 if since baseline a household member died who was head of household at the time of death, and 0 otherwise. We do also know who was head of the household at baseline and if they died,. Since there is only one death of a household head who was not the head at baseline, the alternative outcome variable definition "Baseline household head has died" yield essentially identical results.

members dying in an accident, which provides evidence against the travel mechanism. Many causes of death, however, were not recorded in the survey in the first place since the cause of death was only asked for deaths that occurred for members younger than 50 years old.

We do not expect that mortality would be directly related to the conflict, as during this period, the character of the conflict was not such that civilians were often direct causalities. The study period was a period of instability and unclear governmental authority; however, the conflict was much lower intensity compared to the period of the civil war beginning in 2015. Civilian casualties reported in the governorates of the intervention during the period from 2010 to 2014 were primarily in the urban centers of Aden and Taiz during insurgent attacks on government installations and government crackdowns on protestors, while the increase in mortality seen in our sample is driven by deaths of household heads in peri-urban villages in Lahj governorate. According to the UCDP Georeferenced Conflict Event Dataset (Sundberg et al. 2012), conflict events in Lahj were related to clashes with Al Qaeda in the Arabian Peninsula in more remote areas. Proximity to these conflict events does not correlate with greater impacts on mortality within Lahj.

As an alternative to the causal explanations above, the mortality results could also be an artefact of differential survey attrition. We examine attrition in more depth below but discuss the elements here that are relevant for the mortality results. One univariate comparison in particular may be important: relative to the control group, older household heads are more likely to respond in the treatment group (Appendix Table 2). Although this is a weak result statistically, it still may be the underlying explanation. There are two mechanisms to consider that would lead to an upwardly-biased estimated mortality effect. First, if treatment effects led to positive income-generating changes, a widow may have been more likely to support herself and her remaining family with the new income-generating activity and, thus, more likely to have stayed in her home. Control household widows on the other hand may have had to move, and thus may also have been harder to find for surveying. Second, treatment households maintained further contact with the implementers of the program, whereas control households did not. And, thus, if a household did relocate after the death of the household head, this might make treatment households relatively more easily findable by the survey team. We repeat our analysis using inverse probability weights as an attempt to address potential bias from differential attrition but find our results from Table 3 do not substantively change (see Appendix Table 14; for details on the weighting used, see the section 4 further below). Pushing further against the story that differential attrition is driving the mortality results, we observe that both attrition levels and differences between treatment and control are highest in the urban area (Aden, one of the three governorates, with an average attrition of 28% and a difference between treatment and control of 8 percentage points), however mortality is not higher in Aden than the other two governorates.

Another potential explanation for the mortality results is differential underreporting of deaths of household heads that are SWF beneficiaries. In some cases, households were discovered to have failed to disclose deaths of the household's SWF beneficiary due to concerns about losing benefits. A study of the SWF in 2013 found that across Yemen, in approximately 1% of households reporting receiving SWF benefits, the primary beneficiary had passed away (Unicef Yemen and IPC-IG 2014). Relative to control households, treatment households might have been less concerned about exposing the death of the primary beneficiary after the experience of having regular visits of SFD staff and may have reported previously underreported deaths during the household member listing at endline at a higher rate. This explanation is consistent with the geographic pattern of both mortality levels and treatment effect estimates by governorate: in our sample, Lahj has the highest levels of death and the largest differences between treatment and control – and Lahj was also the governorate with the highest rate of underreported deaths in the SWF report (in 4.5% of SWF beneficiary households in Lahj the targeted individual had passed away).

Finally, we do not see any impacts on two outcomes capturing inter-household transfers that we have available at endline (Panel E). Neither the value of crops given to other households in the last season nor the value of remittances in the past 12 months are statistically significantly different between treatment and control households, and the point estimates do not imply economically important differences. A lack of effects on transfers to other households is in line with the idea that the lack of impacts on consumption is not driven by within-village spillovers via transfers from treatment to control households.

3.3. *Why would assets increase but not consumption?*

Given the importance of these divergent results, we put forward four possible explanations.

First, the difference could be due to imprecision. However, as discussed above, the upper bound on monthly household consumption after four years is a 12% increase over the control group, whereas the asset increase is 39%. We can therefore safely rule out that the increase in consumption at endline was proportionally the same size as the increase in assets.

Second, the difference may relate to a higher proportion of travelers amongst treatment households. We do observe more travelling by household members, and the consumption survey did not include their consumption. This could bias downward the estimated treatment effect on consumption. When we adjust for this in the household consumption per capita by computing consumption per non-travelling member, the point estimate for change in consumption is no longer negative (Appendix Table 5, Panel A). The adjustment is not large enough, however, to explain the difference between the consumption and asset results.

Third, unobserved consumption prior to the period captured by the endline survey may be responsible for the difference. Consumption increases that may have occurred over the four years since the program start are not captured by the endline. The most generous estimate of this would be to use the result on monthly consumption per household, a PPP $9.32 monthly increase per household at endline (Table 2), and multiply this by four years, yielding a PPP$447 increase in consumption over four years. This is likely to be the upper bound, in that the PPP$9.32 per household point estimate has a standard error of PPP$32, thus encompassing both zero and noticeably negative values, and the monthly consumption per capita (rather than per household) point estimate is actually negative (because of the increase in household size in the treatment group relative to the control group).

Fourth, unobserved categories of consumption in the endline survey may explain some of the gap. The results on housing may be evidence of this since we do see improved housing stock in treatment relative to control. We do not have a way of valuing this differential, but improved housing stock is evidence that treatment households were investing more in durable assets (which we find direct evidence of as well).

Lastly, to reconcile these results, we should consider why, if consumption had indeed not improved, households did not divest of more of the livestock. The shocks from the civil conflict could have made them value their assets more, preserving them as a buffer stock for even worse times. Times were already harsh, however; for example, 20% of adults skip entire days of eating (Appendix Table 8). While this may make the buffer stock mechanism somewhat less convincing, the interpretation is both in line with results from other settings (Kazianga and Udry 2006) and with the implementer's assessment of the buffer stock role of livestock in a setting where conflict is salient in household's decision-making. Program officers may also have reinforced buffer-stock behavior through messages delivered during their weekly visits. As reported in a qualitative study of the intervention, program officers strongly discouraged participants from selling off their productive assets (Moqueet 2013).

4. Robustness

4.1. Selective attrition

Given the differences in survey attrition rates at endline between treatment and control groups, we explore the robustness our results with bounding and reweighting exercises. Overall, the additional specifications do not suggest that the interpretation of results is qualitatively affected by bias from differential survey attrition.

In Appendix Table 12 we compare our unadjusted treatment effect estimates (column 4) on the key welfare outcomes from Table 2 with lower and upper bounds from three approaches. For the first, we compute standard Lee (2009) bounds, trimming the empirical distributions at the extremes (from the top for the lower bound, from the bottom for the upper bound) until the number of observations in the group with higher response rate –the treatment group it our setting– is such that the survey response rate is equal to that with the lower response rate. For the second and third approach we follow Kling and Liebman (2004) to estimate lower bounds, we impute values for non-responders in the treatment group with the mean *minus* 0.1 and 0.25 standard deviations of the observed treatment distribution and impute values for non-responders in the control group with the mean *plus* 0.1 and 0.15 standard deviations of the observed control distribution (and vice vera to estimate the upper bound). Overall, we see that the main interpretation of our original results is robust to the bounding exercise. The lower bound effects on assets are positive (though not statistically significant in case of the Lee bound) and the upper bound is substantially less than the value of the initial transfer. For consumption, all three upper bounds rule out substantial positive effects, in line with our interpretation of the unadjusted estimates.

Lastly, we show impacts on our primary outcomes using inverse probability weighting with two reweighting schemes in Appendix Table 13. The first scheme uses only the rates of program recertification in the treatment group. Households who were initially deemed eligible but were deemed ineligible during the reverification of eligibility by the implementer have a high rate of attrition (19%) – a rate that is higher than that of the control group (11%) and much higher than the rate than rate for treatment households who continued to be eligible (6%). We reweight observations such as to proportionally increase the importance of observations of those households who were deemed ineligible during recertification and restore representativeness of endline observations for the entire treatment group (irrespective of recertification status). The second reweighting scheme adds to the information used in the first scheme by including the full set of baseline characteristics used as controls in the main regressions in Tables 2 and 3 as well as the baseline variables of Table 1. Neither approach substantively changes the treatment effect estimates on key welfare outcomes presented in Table 2. Results from an analogous analysis on the outcomes in Table 3 are shown in Appendix Table 14 and reveal no substantive differences either.

4.2. Spillovers

Spillovers, through one of many mechanisms (e.g., risk-sharing or general equilibrium price effects on inputs or outputs), could be present and could thus lead to misinterpretation of the results. Randomization took place at the household level and study households in the same area interact directly and through markets. While the experimental design does not allow us to quantitatively assess the degree of spillovers, we do have some (scant) data on sharing across households. Table 3 Panel E presents these results: value of crops given to other households last season ($PPP -0.11, se = 0.13) and value of remittances received in past 12 months ($PPP 0.18, se = 0.16). Neither result indicates a noticeable shift in sharing. The within-village randomization does not allow us to examine general equilibrium effects on wages or prices.

To further consider spillover effects, we look to the literature on similar programs. Bandiera et al. (2017) do not find evidence of sizeable spillovers of a Graduation program in Bangladesh on ineligible households in treatment villages with a treatment density of 6% of households in the village, about twice the treatment density in the Yemen Graduation project (based on the 12 out of 50 villages for which we have population information for our study). More broadly the limited treatment density suggests that unless the spillover effect sizes from any given treatment household in the community are large, total spillover effects will not be so large as to substantially bias our estimates.

Banerjee et al. (2015) studies the Graduation program in six countries and test for spillovers from treatment households to control households within the same village in three of the sites. In Ghana, Honduras, and Peru randomization took place first at the village-level, followed by household-level randomization within villages. On average across all three sites the point estimate for mean differences between control households in treatment villages (subject to spillovers from treatment households in the same village) and households in control villages (not subject to spillovers) three years after the start of the intervention is 0.003 standard deviations (ibid., Table S6b, column 7, row 1).

Thus, in other contexts, similar interventions[12] do not appear to generate sizeable spillovers to other households. An exception is Raza et al. (2018) which documents sizeable spillovers on childhood malnutrition from the same program studied by Bandiera et al. (2017), but childhood malnutrition is not an outcome we study and spillovers could have operated through a knowledge mechanism about dietary diversity and feeding practices, a type of channel is less likely to be operative for assets or consumption. Another caveat to this interpretation is that the spillover effects in those other sites studied in Banerjee et al. (2015) are imprecisely estimated. The 90%-confidence interval ranges from negative 0.05 to positive 0.06 standard deviations and so potentially meaningful spillover effects (albeit smaller than the estimated primary treatment effects) cannot be ruled out.

5. Cost-benefit analysis

To determine the effectiveness of the program, we must consider the cost-benefit ratio (particularly since this is a fairly expensive program). However, it is impossible to be comprehensive without data on what transpired over much of the four years, and it is reasonable to conjecture that we are underestimating the benefits relative to the costs; the costs are fully measured, and any consumption benefits realized but not measured over the four years are not included in our calculation.

With that caveat in mind, we first examine the change in wealth as a percentage of costs. The size of the transfer to each participating household was PPP$963 on average. Under the assumption of no spillovers, we can scale the ITT estimates from the treatment effect tables by 164% (=1/0.61) to compute the estimate of the Treatment on the Treated (TOT) that accounts for the fact that 39% of the treatment group did not receive the program (and therefore did not benefit from the program but also had no associated costs). Assets increased by an average of PPP$475 per participant, while savings rose by PPP$12 (based on the scaled estimates in Table 2 and Appendix Table 9, respectively). Therefore, the increase in total wealth four years after the asset transfer corresponds to approximately 51% of the size of the transfer.

Purely in terms of wealth, thus, the *measured* benefits after four years to households were substantially less than the size of transfers received.

[12] The number of households per village assigned to treatment was somewhat lower in Ghana compared to our study but similar in Honduras and Peru, suggesting that treatment intensity per village was broadly comparable. In Ghana, Honduras and Peru the ratios were 4.3 (666 treatment households in 154 villages), 10.0 (800/80), and 9.1 (785/86), respectively. In comparison, our ratio is 10.1 (505/50).

But as described above when discussing the results, the fact that the average treatment effect on the number of livestock owned matches the number animals transferred highlights the potential for issues of measurement to be an important factor when comparing our survey measures with the known cost of the transfer. In addition, much remains unmeasured in a single follow-up survey after four years, thus rendering the cost-benefit analysis incomplete. For example, households may have liquidated some of the assets over the four years, perhaps due to the civil unrest, in order to absorb the shock and smooth consumption; or households may have earned and consumed more over the first three of the 4 years as a by-product of the program.

Next, we examine consumption, a perhaps stronger long-term measure, since eventually increases in wealth ought to lead to higher consumption. Estimating the total benefit to individuals requires some assumptions about how consumption levels differed between control and treatment households over time. Using the most generous approach from the above section "Why would assets increase but not consumption?" would yield a PPP$733 increase in consumption (PPP$447 inflated by 1/0.61) over four years. This takes the point estimate for consumption as-is and ignores the fact that this estimate is not statistically significant. When combined with the increase in assets, the total benefit per participant add up to PPP$1120. This figure surpasses the cost of the asset transfer (PPP$963) and approaches the total cost of the program (PPP$1175 after adding costs of supervision of PPP$170 and costs of operation PPP$42).

Lastly, two other potentially important outcomes are not incorporated into the cost-benefit analysis. The housing improvements amongst the treatment group may be indicative of more stable households, able to invest in long term durables. Second, the shift in mortality could be important, and could be indicative of a positive change (more economically active households) or negative (households that are taking on more risks in travel and exposure to civil conflict).

6. Conclusion

As the events of the Arab Spring brought political and economic crisis to Yemen, the Yemen Graduation program became an important test of the Graduation approach in difficult contexts. Could the livelihood, asset-base, and coaching provided by the project give households what they needed to weather sudden macroeconomic shocks? Or would economic conditions prevent beneficiaries from developing livelihoods and moving out of extreme poverty? Answering this question is hampered by two factors. A decision by the implementing agencies to drop a large portion of the treatment group from the program reduced the ability to detect statistically significant results, and the security situation in Yemen delayed the collection of follow-up data until four years after the baseline and the subsequent training and transfer of assets.

Despite these challenges, the evaluation of the Yemen Graduation program shows positive gains in some, though not all, areas. Compared to the control group, treatment households have considerably higher values of productive assets. Treatment households also report saving more and keeping more money in their savings accounts. No statistically significant differences were found for income or total household consumption, and negative impacts were observed on per capita consumption (because household size increased). We observe a potentially important increase in mortality rate that could be the result of positive impacts of the program. The mortality increase could be a consequence of an unintended negative impact from the positive economic impacts (increased economic activity led to more travel and exposure to the civil unrest); or a result of differential survey attrition (which itself may be a by-product of positive economic impacts on households that experience a death for unrelated reasons).

The results differ considerably from the other six sites evaluated as part of this larger worldwide effort. Banerjee et al. (2015) which studied the same program in Ethiopia, Ghana, Honduras, India, Pakistan, and Peru, finds strong increases in income, consumption, assets, food security, mental health, and female empowerment. Furthermore, the program was cost effective in five out of the six sites. We highlight several differences in the program, context and evaluation, to help explain the difference between the results in Yemen and the other sites.

Firstly, the program differed with respect to consumption support. The underlying theory behind the provision of consumption support is that households need a certain degree of "breathing room" in order to profitably manage the asset they receive. If households are chronically food insecure, and on the margin of eating or not eating, they are likely to be tempted to liquidate (or slaughter) any assets received for the sake of improving their consumption in the short-term. In Yemen, both treatment and control households received consumption support as part of a pre-existing government social protection program and this support was irregularly delivered during the political crisis, whereas in most of the other sites only the treatment group received the consumption support.

Secondly, the civil unrest in Yemen could be a critical underlying contextual variable that lowers the long-run returns to the program. It could make it more difficult for participants to access markets, or (more positively) could lead to higher incentives to sell or slaughter assets in order to absorb the shock from the civil unrest and address immediate consumption needs.

Thirdly, the long gap in surveying (four years, from 2010 to 2014, as a result of the political instability) limits our ability to observe what happened in the early stages of the project and is also likely to increase outcome variance. It could be that the program had a large short-run positive impact, specifically in helping households absorb the shock from the civil unrest. We do find some evidence, albeit limited, to support this: households in the treatment group were more likely to report using savings to absorb shocks. Note, however, that evaluations have found persistent long-run impacts of similar programs in India and Bangladesh (Banerjee et al., 2016; Bandiera et al., 2017). The gap in surveying therefore limits our ability to measure mechanisms but is not itself an explanation for the lack of positive treatment effect on many outcomes, in particular consumption.

The lack of consumption support and presence of civil unrest potentially interacted with each other, as consumption support would be particularly relevant in a setting where participants were exposed to more adverse shocks, especially over a longer time horizon. In two of the six stable sites (Ethiopia and Peru) reported on in Banerjee et al. (2015), consumption support did not differ between treatment and control groups and yet evaluations found impacts from the asset transfer and training alone. On the other hand, in both of the other evaluations so far conducted in conflict-affected settings (Afghanistan and South Sudan), consumption support has differed between treatment and control groups. In the Afghanistan Graduation program, where impacts were particularly strong, participants received not only ongoing consumption support, but also veterinary support and replacement of sick or deceased animals during the operation of the program (Bedoya et al., 2019; Chowdhury et al., 2017).

There is some external evidence for the potential for the Graduation program asset transfers to increase rather than reduce vulnerability to shocks in unstable contexts such as Yemen. The Yemen Social Fund for Development's Rainfed Agriculture and Livestock Program is similar to the Yemen Graduation program in that it provided livestock assets and

training, but it relied on self-selection into cooperative groups and required co-investment rather than targeting the ultra-poor. However, an evaluation of this program during the same time period (2009–2011) found that even for somewhat better-off households, increases in livestock ownership were not accompanied by increases in consumption measures and participation in the program caused an increase in the probability that households had experienced a money shortage during the past 12 months (Christian et al. 2015). Raising sheep and goats is costly for farmers in Yemen, even if the animals are originally transferred for free. In addition to the time spent grazing animals, owners invest in constructing pens, providing water, purchasing feed supplements, and paying for medications and veterinary services. The baseline for this evaluation showed that most households in rural Yemen do not have access to veterinary services for their livestock, and conversations with program staff and participants revealed that this lack of access was compounded by the conflict and participants tended to sell off livestock during times of crisis if they did not expect that they would receive sufficient support from the program (Egel and Yeslam 2010). Faced with shocks exacerbated by the conflict—such as animal disease, high cost or difficulty in access to watering sources, or high costs of transportation to markets and low prices—households may find that their expected returns to investing in raising livestock are negative.

The fact that some initially randomized beneficiaries preferred to sell off the livestock assets rather than engage in an income generating project, as well as the significant impacts found by the evaluation on durable good ownership compared to somewhat low though positive impacts on livestock asset ownership, also suggest that for risk averse households, the expected returns from the livelihood projects may have been too low to accept the risk of losses in an already highly unstable environment.

Finally, as discussed above, the re-qualification process amongst the treatment group led to lower statistical power for assessing the impact of the program. However, the India site from Banerjee et al. (2015) had a similar issue (albeit through participant choice rather than disqualification), and the results in India are quite strong across almost all outcome measures, both two and three years after the asset transfers.

Overall, and even during exceptionally difficult circumstances, the Graduation approach in Yemen was able to have important positive impacts on asset accumulation and savings behavior four years after the asset transfer, albeit substantially less –as measured– than the amount the household had originally received.

Still, the question remains whether the lack of impact on other dimensions is due primarily to missing short-run benefits from the program, to the crisis, to the low precision from the requalification process, or to some other factor. While the Graduation approach has been largely successful elsewhere, there has been a wide range of impacts among the nine sites previously studied (Bangladesh, reported in Bandiera et al. (2017); Ethiopia, Ghana, Honduras, India, Pakistan and Peru, reported in Banerjee et al. (2015); South Sudan reported in Chowdhury et al. (2017); and, Afghanistan reported in Bedoya et al. (2019). Further research is needed to better understand how such programs can help households to survive system-wide crises. Such research would likely benefit from more frequent data collection during the crisis, though this poses obvious logistical challenges.

For any social protection program, ongoing conflict could diminish implementation fidelity or inhibit households from embracing and flourishing with opportunities, but could on the other hand be critical for helping households mitigate the shock and its ensuing economic effects. In the context studied, we find a multi-faceted, modified Graduation program yielded some important benefits to recipients but was not as successful as has been observed in non-conflict settings.

Data availability

Analysis data and code on the Harvard Dataverse (https://doi.org/10.7910/DVN/5NM88Z).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jdeveco.2021.102780.

References

Anderson, Michael, 2008. Multiple inference and gender differences in the effects of early intervention : a reevaluation of the abecedarian , perry preschool , and early training projects. J. Am. Stat. Assoc. 103 (484), 1481–1495. https://doi.org/10.1198/016214508000000841.

Bandiera, Oriana, , et al.Burgess, Robin, Das, Narayan, Gulesci, Selim, Rasul, Imran, Rasul, Imran, Sulaiman, Munshi, 2017. Labor markets and poverty in village economies. Q. J. Econ. November https://doi.org/10.1093/q, 811–70.

Banerjee, Abhijit, Duflo, Esther, Chattopadhyay, Raghabendra, Shapiro, Jeremy, 2016. The Long Term Impacts of a 'Graduation' Program: Evidence from West Bengal.

Banerjee, Abhijit, Duflo, Esther, Goldberg, Nathanael, Dean, Karlan, Osei, Robert, Parienté, William, Shapiro, Jeremy, Thuysbaert, Bram, Udry, Christopher, 2015. A multifaceted program causes lasting progress for the very poor: evidence from six countries. Science 348 (6236). https://doi.org/10.1126/science.1260799.

Bedoya, Gaudalupe, Coville, Atdan, Haushofer, Johnannes, Isaqzadeh, Mohammad, Shapiro, Jeremy, 2019. No household left behind : Afghanistan targeting the ultra poor program impact evaluation. World Bank Policy Research Working Paper 8877.

Benjamini, Yoav, Hochberg, Yosef, 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 57 (1), 289–300.

Blattman, Christopher, Green, Eric, Annan, Jeannie, Jamison, Julian, 2016. The returns to microcredit support among the ultra-poor. A Field Experiment in Post-War Uganda" 8 (2), 35–64.

Chen, Shiyuan, Schreiner, Mark, 2009. Yemen Progress Out of Poverty Index Design Documentation.

Chowdhury, Reajul, Collins, Elliott, Ligon, Ethan, Sulaiman, Munshi, 2017. Valuing Assets provided to Low-Income Households in South Sudan, pp. 1–26.

Christian, Sikandra, de Janvry, Alain, Egel, Daniel, 2015. Quantitative Evaluation of the Social Fund for Development Labor Intensive Works Program (LIWP). *University of California at Berkeley*, pp. 1–76. CUDARE Working Papers.

Ecker, Oliver, Maystadt, Jean-Francois, Guo, Zhe, 2021. Can unconditional cash transfers mitigate the impact of civil conflict on acute child malnutrition in Yemen?, April, 1–33.

Economist, 2013. Down and Out. *Economist*. December 2013.

Egel, Daniel, Yeslam, Tareq Al-Bass, 2010. Conditions in Rural Yemen : Findings from the RALP Baseline Survey. Social Fund for Development Working Paper. http://sfd.sfd-yemen.org/uploads/issues/Conditions in Rural Yemen Findings from the 2010 RALP Baseline Survey-20120924-113944.pdf.

Field, Erica, Jayachandran, Seema, Pande, Rohini, Rigol, Natalia, 2016. Friendship at work: can peer effects catalyze female entrepreneurship? Am. Econ. J. Econ. Pol. 8 (2), 125–153. https://doi.org/10.1257/pol.20140215.

Kazianga, Harounan, Udry, Christopher, 2006. Consumption smoothing? Livestock, insurance and drought in rural Burkina Faso. J. Dev. Econ. 79 (2), 413–446. https://doi.org/10.1016/j.jdeveco.2006.01.011.

Kling, Jeffrey R., Liebman, Jeffrey B., 2004. Experimental Analysis of Neighorhood Effects on Youth. https://users.nber.org/~kling/483.pdf.

Lee, David S., 2009. Training, wages, and sample selection: estimating treatment effects. Rev. Econ. Stud. 76 (3), 1071–1102. https://doi.org/10.1111/j.1467-937X.2009.00536.x.

Moqueet, Nazia, 2013. Graduation Out of Poverty in Yemen : Impact of the Political Crisis on the Extreme Poor CGAP-Ford Foundation Graduation Pilot.

Raza, Wameq A., Van de Poel, Ellen, Van Ourti, Tom, 2018. Impact and spill-over effects of an asset transfer program on child undernutrition: evidence from a randomized control trial in Bangladesh. J. Health Econ. 62, 105–120. https://doi.org/10.1016/j.jhealeco.2018.09.011.

Reuters Staff, 2013. Yemen's Main Oil Pipeline Attacked, Pumping Halted. *Reuters*. January 26, 2013.

Sundberg, Ralph, Eck, Kristine, Kreutz, Joakim, 2012. Introducing the UCDP non-state conflict Dataset. J. Peace Res. 49 (2), 351–362. https://doi.org/10.1177/0022343311431598.

Unicef Yemen, and IPC-IG, 2014. Yemen National Social Protection Monitoring Survey (NSPMS): 2012-2013 Final Report. http://nspms-yemen.ipc-undp.org/.