Mathematica
Progress Together

Jordan Refugee Livelihoods Development Impact Bond Evaluation Framework

Report

February 4, 2022
Evan Borkum, Paolo Abarcar, Laura Meyer, Matt Spitzer

Jordan Refugee Livelihoods Development Impact Bond Evaluation Framework

February 4, 2022

Evan Borkum, Paolo Abarcar, Laura Meyer, Matt Spitzer

Submitted to:

Stichting IKEA Foundation
Dellaertweg 9G
2316 WZ Leiden, the Netherlands
Attention: Laura Costica

Submitted by:

Mathematica
1100 First Street, NE, 12th Floor
Washington, DC 20002-4221
(202) 484-9220
Fax: (202) 863-1763

Contents

Tables

Figures

ACRONYMS

CBO	community-based organization
DFC	United States Development Finance Corporation
DHS	Demographic and Health Survey
DIB	Development Impact Bond
FGD	focus group discussion
HIES	Household Income and Expenditures Survey
IGA	income generating activity
IRB	institutional review board
LSMS	Living Standards and Measurement Study
NEF	Near East Foundation
RCT	randomized controlled trial
TVET	technical and vocational education and training

I. Introduction

The ongoing Syrian civil war, which began in 2011, generated the world's largest refugee crisis since the Second World War. Almost 6.6 million people—about one quarter of Syria's pre-war population—are estimated to have fled the country, with most seeking refuge in neighboring Turkey, Lebanon, and Jordan. As of late 2020, Jordan hosted about 660,000 registered Syrian refugees, together with another 90,000 registered refugees from other crisis-hit countries such as Iraq, Yemen, and Sudan (United Nations High Commissioner for Refugees 2020).

Among Syrian refugees in Jordan, about 8 in 10 lived below the national poverty line in 2019, even before the COVID-19 pandemic worsened economic conditions in the country (United Nations High Commissioner for Refugees 2019). Female-headed refugee households are particularly vulnerable, with cultural expectations and home care duties acting as additional constraints to securing employment and livelihoods. An estimated 40 percent of refugee households are headed by women and, as of 2016, only 3 percent of refugee women worked (Krafft et al. 2019; United Nations 2018). Although Jordan has several refugee camps in which governmental and non-governmental organizations provide free services, limited livelihood opportunities and a lack of privacy in the camps lead most refugees to live in cities outside the camps, where they face high housing costs (Aziz et al. 2019; Wall et al. 2017; United Nations High Commissioner for Refugees 2019). A lack of livelihood opportunities has forced many refugees to deplete their assets and savings and accumulate large debts (Culbertson et al. 2016, ReliefWeb 2017). Many refugees have also resorted to other negative coping strategies such as accepting socially degrading, exploitative, high risk, or illegal jobs or reducing expenditures on essentials (United Nations High Commissioner for Refugees 2019).

The large influx of Syrian refugees has occurred in a context in which there is a large population of vulnerable Jordanians—often living in the same communities as refugees—facing livelihoods-related challenges. For example, 19 percent of Jordanian adults were unemployed and about 16 percent of Jordanians lived below the poverty line even before the pandemic (United Nations Children's Fund 2020). The influx of refugees has affected Jordanian citizens in complex ways, both real and perceived. For example, increased competition for housing may have worsened average housing quality for poor Jordanians and increased rental prices (Al-Hawarin et al. 2018). In contrast, recent evidence suggests that the labor market outcomes of Jordanian workers were little affected, likely because refugees and Jordanians were largely not competing for the same jobs and because increased public sector investment in response to the refugee crisis improved job opportunities for Jordanians (Fallah et al. 2018; Malaeb and Wahba 2018). Nevertheless, resentment over differential access to employment and perceptions of Syrians competing with Jordanians for jobs might still have increased tensions and adversely affected social cohesion in host communities (REACH 2014).

More recently, the pandemic has exacerbated the economic challenges facing both refugees and vulnerable Jordanians. Due in large part to the economic slowdown resulting from pandemic-related shutdowns and mobility restrictions, the national unemployment rate reached 25 percent in the fourth quarter of 2020, with youth unemployment at 50 percent (World Bank 2021). Among those who were employed before the pandemic, workers in informal employment—which is the most common form of employment among refugees and vulnerable Jordanians—were especially hard-hit due to the absence of written contracts to secure their employment and because they were ineligible for social security payments that were introduced to mitigate the effects of the pandemic (Kebede et al. 2020; Kebede et al 2021). Many of these workers permanently lost their jobs, while others experienced a substantial decline in household income.

In the first years of the Syrian refugee crisis, the response by the government of Jordan, foreign donors, and international organizations was primally humanitarian in nature, focused on meeting refugees' short-term needs for shelter, food, and cash. However, as the protracted nature of the displacement from Syria became apparent, these stakeholders sought a longer-term, more development-oriented approach to build self-reliance and resilience among Syrian refugees. Such an approach, which focuses on helping refugees secure sustainable livelihoods, could position refugees to become assets and major economic contributors to their host countries in the long run (Clemens et al. 2018; Legrain and Burridge 2019).

An important step in this new approach was the 2016 Jordan Compact, signed by the Jordanian government and the European Union, which outlines the objectives of facilitating refugees' labor market access while mitigating adverse impacts to local citizens through financial assistance and trade concessions (Lenner and Turner 2019). Under the Compact, the Jordanian government agreed to allow as many as 200,000 Syrian refugees to obtain work permits in certain sectors, simplified the fees and administrative procedures for these permits, and allowed Syrians to operate certain types of home-based businesses. This initiative is believed to have contributed to a modest increase in labor force participation among adult Syrian refugees and a large decrease in the unemployment rate among adult Syrian refugees in the labor force between 2014 and 2018 (Tilnes et al. 2019). Nevertheless, in 2018, about 40 percent of adult Syrian adult refugees remained out of the labor force and a further 15 percent were in the labor force but unemployed. Further, among those employed, the majority were employed in the informal sector, where their jobs lack legal protection. This suggests that, despite improvements in recent years, many refugees have still been unable to attain secure livelihoods.

Building on the paradigm shift toward a development-oriented approach to the Syrian refugee crisis, a group of international partners are partnering on an innovative Refugee Livelihoods Development Impact Bond (DIB) in Jordan. The DIB, which is coordinated by KOIS, will finance a four-year microenterprise training and grants program for refugees and vulnerable Jordanians in host communities. This program will be implemented by the Near East Foundation UK (NEF) in collaboration with local community-based organizations (CBOs). The DIB investors (the United States International Development Finance Corporation [DFC] and Ferd) are providing NEF with the upfront financing for the program. Under the DIB mechanism, the funders (IKEA Foundation, Novo Nordisk Foundation, and Norad) will pay the investors at the end of the program, with the final payment amount depending on the results achieved. The main objectives of this DIB are: (1) to sustainably improve program participants' abilities to meet basic needs; their economic well-being, self-reliance, and resilience; and women's confidence, bargaining power, and agency; and (2) to encourage international development actors to devote more resources to long-term livelihoods programs in refugee contexts and demonstrate the potential of innovative funding mechanisms to achieve this.

The Refugee Livelihoods DIB is one of 15 DIBs in lower- and middle-income countries launched since 2018 (Brookings 2021). [1] DIBs are become an increasingly popular mechanism for funding development programs and offer two main benefits compared to traditional grant-based financing. First, they appeal to a wider range of potential investors, including those in the private sector, who have access to large pools of capital. This could potentially lead to larger volumes of capital and thus increased scale. Second, they intensify all partners' focus on measurement, learning, and results, given the payments at stake. This could potentially lead to improved quality and thus effectiveness of the program. Whether, and to what

[1] Social Impact Bonds (SIBs), in which the outcome payer is the domestic government instead of a foreign government or private foundation as is the case with a DIB, are much more prevalent, with over 200 launched globally as of December 2021.

extent, a DIB-funded program in fact achieves results that differ from a traditional-grant funded one has not been rigorously measured, and likely depends on the context. Previous DIBs have tackled challenges in health (Cameroon, India, Nigeria, Democratic Republic of Congo, Mali), employment and training (Palestine), education (India), poverty reduction (Kenya), and agriculture and the environment (Peru) (Oxford Government Outcomes Lab 2021). To the best of our knowledge, the Refugee Livelihoods DIB is the first to focus on supporting livelihoods in a refugee context.

The IKEA Foundation has contracted with Mathematica to conduct an independent evaluation of the DIB. Mathematica's evaluation seeks to both measure the metrics to determine payments to investors and generate broader learning about the program's impacts to support future adaptation and scale-up. This evaluation framework report describes Mathematica's proposed design for the evaluation. We begin by describing the program that the DIB is funding in Jordan (Chapter II) and providing context for the evaluation by reviewing the relevant literature (Chapter III). Next, we describe the design of the evaluation and our approach to data collection and analysis, including for the DIB payment metrics and the broader evaluation (Chapter IV). We also address the limitations and risks of the evaluation, as well as plans to mitigate them (Chapter V). We conclude by presenting the evaluation timeline and reporting schedule (Chapter VI).

II. The Refugee Livelihoods DIB program

In this chapter we describe the refugee livelihoods program—including the planned rollout and program activities—and discuss how the program activities are expected to lead to the desired outcomes. We also provide more context for the DIB mechanism by describing in more detail the roles of the various parties.

1. Program description

The livelihoods program that NEF will implement under the DIB focuses on supporting participants to create sustainable, mostly home-based businesses. It builds on similar work conducted since 2013 through several iterations of the Enhancing Economic Resilience project in Jordan and Lebanon. Over the course of the program, NEF expects to serve 5,040 refugees and vulnerable Jordanians, spread equally across three cohorts.

For each cohort, NEF will partner with local CBOs, selected on a competitive basis, to identify participants and deliver the program across five locations in Jordan.[2] NEF and its partner CBOs will deliver the program through NEF's existing *Siraj* centers in the selected locations. These are physical hubs managed by the CBOs at which vulnerable individuals can access training, financial resources, and advisory services to support their livelihoods, or find referrals to or information about other services related to their physical, economic, and mental wellbeing. By training and supporting local CBOs to implement the program, NEF hopes to build their capacity to identify and support refugees and vulnerable Jordanians in the future.

NEF and its partner CBOs will identify potential participants for each cohort during a three-month inception period through door-to-door canvassing (prioritizing neighborhoods that have not yet been reached by the program or other livelihoods programs), coordinated referrals (for example, from municipalities, other CBOs, non-governmental organizations, United Nations agencies, or government anti-poverty programs), or *Siraj* center walk-ins. Potential participants will be screened using a vulnerability assessment tool to identify those who need livelihoods support to meet their basic needs and are willing to commit to the program. At least 30 percent of participants will be refugees, at least 75 percent will be women, and at least 35 percent will be youth (ages between 18 and 25); a maximum of 10 percent will be existing business owners.

Under current implementation plans (which might be adjusted over time), the program will include several activities aimed at building participants' skills and supporting the development and growth of new businesses:

- A core approximately 11 day blended in-person and remote **sequenced training in life and business skills** for all participants, culminating in the preparation of a business development plan;
- Brief **technical and vocational skills training** for about 35 percent of participants in topics relevant to their business development plan, conducted by third party experts or training providers;
- **Grants of about $850** for about 68 percent of participants to finance their business development plans, subject to an application from participants and approval of their plans by a grants committee;[3]

[2] These locations are Zarqa, Amman, Aidon, Kuforsum, and Russifieh. Each CBO will cover one location and NEF expects to engage the same CBOs for all three cohorts.

[3] Jordanian applicants must commit to spending the grants on business-related needs; refugees will not be required to make commitments on how the grant will be used.

- An **advanced tailored business training and additional financial support** for a small number of grantees (about 3 percent) whose businesses demonstrate particularly strong potential for growth; and
- Additional advanced and specialized support for grantees as part of the "*Siraj* accelerator" initiative, with topics driven by beneficiary demand and informed by a monthly business support tool. This support includes **small group coaching** on relevant topics, creation of **peer business networks** to provide peer-to-peer mentorship through a digital platform, and **one-on-one business support** in the form of structured mentorship sessions.[4]

For each cohort, NEF and its CBO partners will provide the core business and life skills training, the technical and vocational skills training, and grants to participants over a six-month implementation period. Participants will be divided into small groups that will receive training at different times within the implementation period; it is expected that grants will be awarded in the second half of the implementation period, based on the assessment of a business plan developed during the core training. NEF and its partners will provide the advanced business training, as well as other ongoing support, over a subsequent five-month period. Overall, implementation for the first cohort will run from February 2022 to October 2022, for the second cohort from January 2023 to September 2023, and for the third cohort from April 2024 to February 2025.

2. Program logic

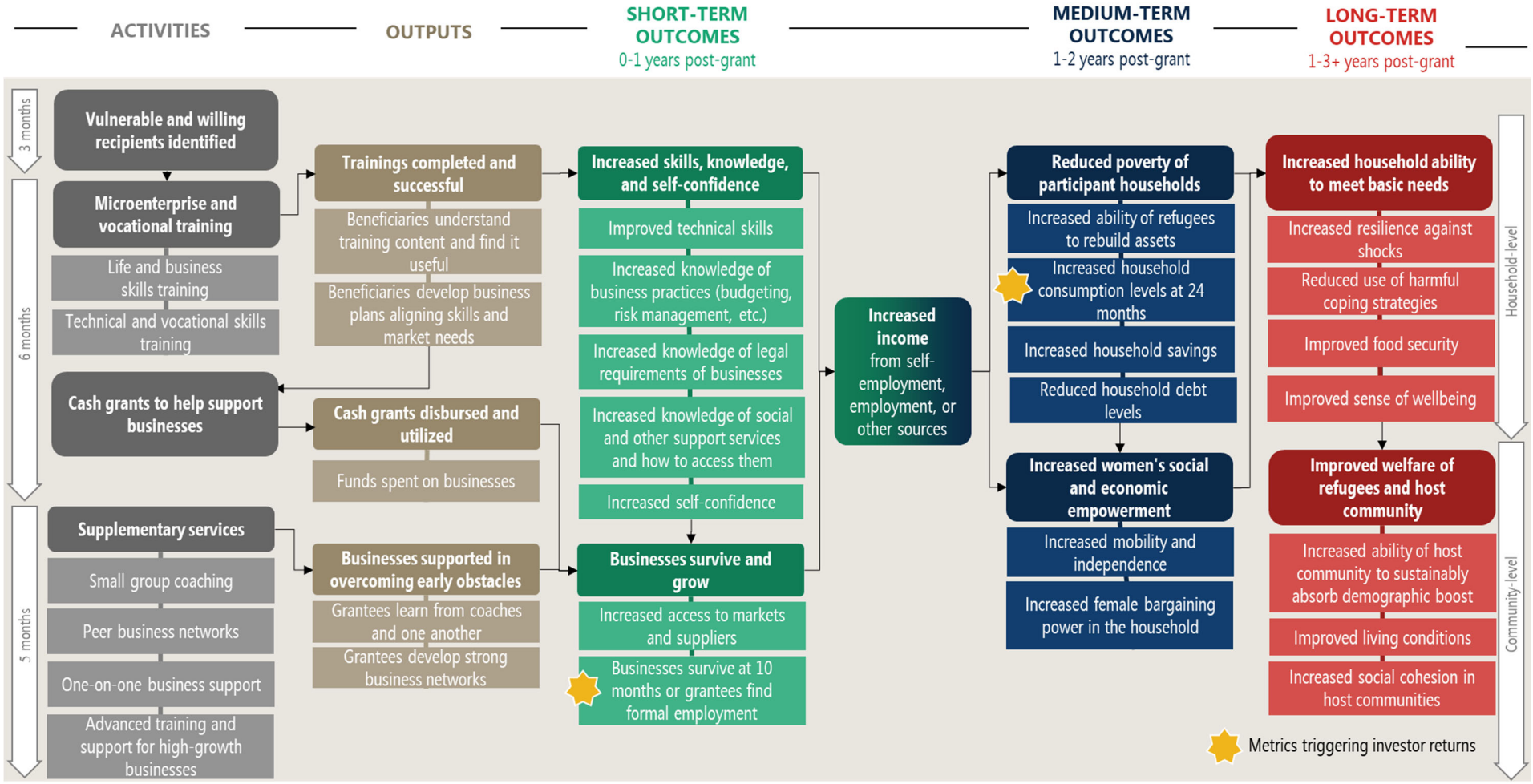
The program logic (**Figure II.1**) illustrates how the program's activities are expected to lead to outputs and, ultimately, short-, medium-, and long-term outcomes. The training activities conducted during the implementation period are expected to culminate in participants developing viable business plans. Cash grants will provide selected participants with the seed capital required to establish these businesses as income generating activities (IGAs), and a set of supplementary services will support them to overcome early obstacles. In the short-term, program participants are expected to build their knowledge of business practices and legal requirements, technical skills, and self-confidence. The increase in knowledge and skills, together with the direct support to overcome early obstacles, is expected to help these newly established businesses survive and grow. Some participants might use their new skills, self-confidence, and increased knowledge of other support services to find formal wage-earning employment as an alternative and secure IGA.

In the medium term, these short-term outcomes are expected to translate to a sustained increase in participants' income from self-employment or wage employment. This increased income is expected to reduce poverty as refugees rebuild their assets, increase their savings, reduce their debt levels, and increase consumption. Women participants' increased contribution to household income is expected to increase their social and economic empowerment within the household.

In the longer term, these changes are expected to contribute to transformations both at the household and at the community level. Households will be better able to meet their basic needs—even in the face of adverse shocks—without having to resort to harmful coping strategies. As a result, they will experience a greater sense of wellbeing. In the aggregate, host communities will be better able to integrate refugees as both refugees and vulnerable Jordanians are able to meet their basic needs, resulting in improved living conditions and social cohesion.

[4] NEF anticipates that 75 percent of grantees will access at least some *Siraj* accelerator activities. NEF expects that each of these grantees will receive three small group coaching sessions (two in-person and one digital) and seven mentorship sessions (five in-person and two digital), on average.

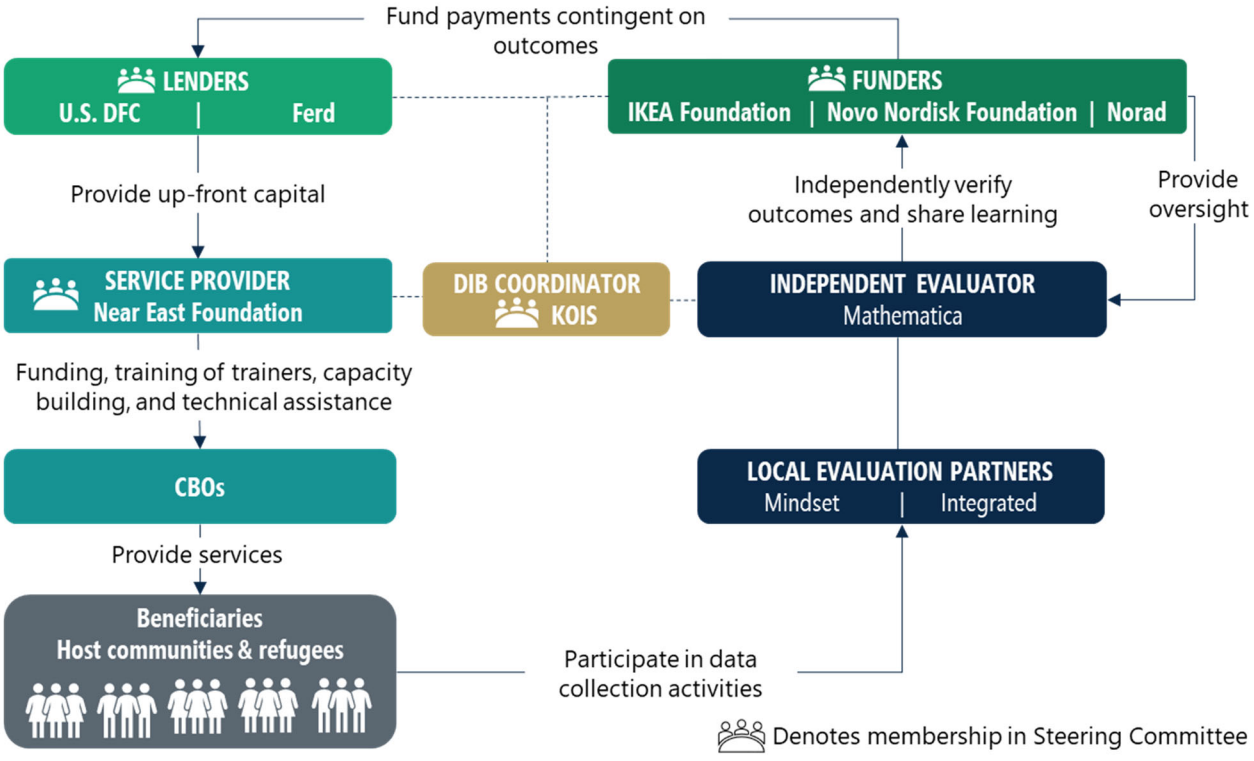
Figure II.1. Program logic



3. Structure of the DIB

The four-year DIB includes several parties, each with a distinct role (**Figure II.2**). NEF (the service provider) will receive up-front capital from DFC and Ferd (the lenders) to implement the program for each cohort, working through the local CBOs. Mathematica (the independent evaluator)—in close collaboration with its local evaluation partners, Mindset and Integrated—will conduct an evaluation of the DIB. As part of the evaluation, Mathematica will measure the two DIB payment metrics: (1) the percent of grantees across all three cohorts actively engaged in IGAs about 10 months after grants are disbursed (the "business metric"), and (2) impacts on household consumption for the first cohort about 24 months after grants are disbursed (the "household consumption metric"). At the conclusion of the DIB, the IKEA Foundation, Novo Nordisk Foundation, and Norad (the funders) will repay the lenders, potentially with interest, with the total amount of the payment dependent on the value of the two payment metrics reported by Mathematica. Under this DIB model, the lenders take on the risk of program performance, the service provider benefits from the stability of multiple years of funding (subject to satisfactory performance at intermediate milestones), and the funders repay lenders based on concrete achievements in targeted social outcomes that are independently verified by the independent evaluator. KOIS (the DIB coordinator) has worked to financially structure the DIB, coordinate the agreements between the other parties, and engage service providers, and will continue to facilitate coordination between parties during implementation.

Figure II.2. Structure of the DIB



III. Literature review

In this chapter we contextualize the refugee livelihoods program by reviewing the literature on similar programs that seek to improve business- and household-level outcomes for entrepreneurs in low- and middle-income countries. Overall, we find a solid body of rigorous evidence suggesting that programs that provide human capital (training or mentorship) and financial capital (grants, loans, or productive assets) can effectively support business outcomes such as sales and profits, and in some cases can improve household outcomes such as consumption. However, there are several important caveats regarding the relevance of these findings for the refugee livelihoods program. First, little is known about the impacts of these programs in a refugee context. As a result, we focused this review more broadly on interventions with similar characteristics to those of the refugee livelihoods program. Second, many of these studies focus on existing businesses, not newly created ones, which are the main target of the refugee livelihoods program. Third, most studies focus on business outcomes and not household outcomes; we included studies that focus on business outcomes in our review because there is a logical link between the success of an entrepreneur's business and their household outcomes.

We begin by presenting the evidence on the effects of human capital augmentation, which can include traditional classroom-based training programs and more tailored mentoring interventions. We then turn to the effects of financial capital, which can be provided through business grants, microfinance, or unconditional cash transfers. Next, because most beneficiaries of the refugee livelihoods program are expected to be women, we examine the effects of entrepreneurship programs on women specifically. We conclude by discussing the contributions that the DIB evaluation will make to the literature and potential implications of the evaluation for future investments and government policy in Jordan and the region.

1. Human capital

In this section we review the evidence on programs that seek to improve the business-related knowledge and know-how of existing and prospective entrepreneurs. These include training on business skills and practices, technical and vocational education and training, and mentorship initiatives—all of which are components of the refugee livelihoods program.

a. Business skills and practices

The refugee livelihoods program will support entrepreneurs in developing ideas for businesses, creating business plans, and running home-based businesses. Several studies assess the impact of classroom-based business training for small-scale entrepreneurs who are either looking to start a business or currently own one. For aspiring entrepreneurs, common topics covered in the training include the development of business ideas and plans, and the logistics of permits, pricing, and budgeting. For current owners, training often covers record-keeping, accounting, marketing, people management, and operations.

In a meta-analysis of 19 studies in low- and middle-income countries that use Randomized Controlled Trials (RCTs) to assess the impact of programs that provide training on business skills and practices, McKenzie and Woodruff (2021) find that traditional classroom training typically results in a 5.6 percent increase in business sales and a 12.1 percent increase in business profits. In an RCT of business training for aspiring female entrepreneurs who plan to start subsistence enterprises in Sri Lanka, de Mel et al. (2014) find that those who participated in training create businesses sooner than their untrained counterparts and that their businesses have higher profits. McKenzie and Woodruff (2017) use panel data on micro and small firm productivity in Bangladesh, Chile, Ghana, Kenya, Mexico, Nigeria, and Sri

Lanka to identify specific business practices that are associated with higher business sales and profits. Their findings suggest that effective trainings might need to emphasize—and support participants in adopting—a key set of business practices (in marketing, stockkeeping, record-keeping, and financial planning.

In addition to business skills and practices, coaching on soft skills will also be part of the refugee livelihoods program. In an RCT for existing businesses in Togo, Campos et al. (2017) find that a training focused on personal initiative resulted in increased profits more than two years after the training ends. Similarly, Ubfal et al. (2019) conduct an RCT with existing businesses in Jamaica to show that a stand-alone soft-skills training that focused on personal initiative resulted in positive, significant effects on business sales and profits, whereas a training that emphasized topics more common in traditional business training—such as strategic management, financial management, and business plan formation—did not achieve positive effects.

b. Technical and vocational education and training

For a subset of entrepreneurs who plan to specialize in a particular trade, NEF will provide technical and vocational education and training (TVET) that teaches skills specific to that trade. Studies of TVET programs in low- and middle-income countries typically show modest effects on employment and earnings, although there is evidence of promising effects for women-led home-based businesses.

In a review of 12 RCTs of TVET programs, McKenzie (2017) finds that fewer than a third report a statistically significant effect on employment or earnings. One study that does find positive effects is Maitra and Mani (2012), which finds that a training program focused on stitching and tailoring design for women in low-income households in India resulted in an increased likelihood of employment, hours worked, and earnings. Chinen et al. (2017) review 35 experimental and quasi-experimental studies globally focused on vocational and business training programs for women and find that training increased likelihood of employment, although these effects waned over time.

Impacts of TVET programs on self-employment are especially relevant to the refugee livelihoods program. Borkum et al. (2017) use an RCT to assess the impact of a TVET scholarship program for youth in Namibia and find no impacts on employment—including self-employment—and earnings. In an RCT in Mongolia, Alzua et al. (2019) show that a TVET program for disadvantaged urban youth resulted in increased self-employment and earnings, although it did not have a statistically significant impact on overall employment. Similarly, Chakravarty et al. (2019) use a regression discontinuity design to estimate impacts of a TVET program for disadvantaged youth in Nepal and find positive impacts on non-farm employment and earnings, but no effect on overall employment. The gains in non-farm employment and earnings appear to be driven by women starting businesses at home, a relevant finding for the refugee livelihoods program which aims to encourage women to start their own enterprises.

c. Business mentorship

NEF plans to offer peer business mentorship networks and one-on-one structured business mentorship to grantees. A range of studies look at the role that mentorship plays in supporting businesses. Mentorship programs have the potential to provide more specific, tailored knowledge than that taught in typical business trainings.

McKenzie and Woodruff's (2021) review of the literature concludes that mentoring by peers can lead to higher business profits in some cases, especially in the period when mentoring is the most active, but the

effects can attenuate over time. For example, Brooks et al. (2018) conduct an RCT with existing businesses owned by inexperienced women in Kenya and find that, compared to a formal business education intervention which did not achieve any effect on business profits, the mentorship program improved business profitability in the short term. However, the effect of mentorship on profits faded over time as the frequency of mentor-mentee meetings fell. In an experiment with existing business owners in Uganda, mentorship was shown to have large, durable effects on sales and profits when advanced firms expand to new markets where their knowledge is limited (Anderson et al. 2020 and Anderson et al. 2021). In addition, Anderson et al. (2021) confirms the intuitive notion that higher-quality mentors lead to larger improvements in business profits and sales. This might help explain why not all mentorship programs have been successful.

2. Financial capital

In this section we summarize evidence on the role of financial capital in supporting entrepreneurs to create and grow their businesses. We focus primarily on the effects of business grants but also discuss the related literature on microfinance programs and unconditional cash transfers.

a. Business grants

The business grant (of approximately $850, on average) is a cornerstone of the refugee livelihoods program, providing initial start-up capital with the aim of facilitating business creation and growth. However, a grant (or cash infusion of any kind) will be effective only if capital is in fact a limiting constraint. The evidence on this question is mixed. On the one hand, Giné and Mansuri (2014) use a field experiment with entrepreneurs in rural Pakistan to show that the provision of a $1,700 business loan had little effect on household welfare, implying that there may be constraints to business growth other than financial capital. Further, in an RCT with both new and existing women-led businesses in Sri Lanka, de Mel et al. (2014) find that providing a $130 grant conditional on completing a business training raised business profits initially, but that gains in profits faded after two years.

On the other hand, an evidence review by Blattman and Ralston (2015) that synthesizes the results of multiple impact evaluations of business skills, soft skills, and TVET programs in fragile or conflict-affected states finds that programs that combine grants with skills training tend to be more effective than skills training alone. The grant sizes of the programs included in Blattman and Ralston (2015) range from $150 to $1,200 per person. The refugee livelihoods program fits the description of the programs that Blattman and Ralston (2015) find most effective, since it combines cash with other forms of assistance to address multiple constraints to starting a business.

Further, several studies have found that entrepreneurship programs that include grants have positive impacts on household consumption, the secondary payment metric for the DIB. Blattman et al. (2016) conducted a RCT of a program for war-affected women in Uganda that provided a $150 grant, business training, guidance developing a business plan, and supervision to implement the plan, with the goal of supporting these women in establishing new businesses. The study found that program participants experienced a 0.33 standard deviation increase in durable assets and a 39 percent increase in non-durable consumption relative to the control group, 16 months after grant disbursement. Also in conflict-affected areas of Uganda, Blattman et al. (2014, 2020) conducted an experimental evaluation of a program that provided grants to groups of poor and unemployed young adults to invest in training or business start-up, with the aim of setting up their members as craftspeople (the grants had a value of about $400 per person, on average). The authors find that grant recipients were more likely to work and had higher levels of

earnings and consumption compared to the control group four years after disbursement. Specifically, their durable assets increased by 0.16 standard deviations and their nondurable consumption increased by a similar amount (equivalent to an 11 percent increase), although the impacts on nondurable consumption had largely faded nine years after disbursement. Bedoya et al. (2019) conduct an RCT in conflict-affected areas of Afghanistan to look at the impact of a comprehensive program that combines skills training and business coaching for aspiring women entrepreneurs with livestock assets and a cash consumption stipend. They find positive impacts on women's time spent working on self-employment businesses and a 30 percent increase in household consumption, driven mainly by food consumption.

There is some evidence that the effects of financing differ for aspiring entrepreneurs as opposed to existing business owners. In a systematic review of 113 rigorous studies on active labor market programs for youth in 31 countries, Kluve et al. (2017) find that grants—either combined with training or administered alone—result in larger growth in profits for new business owners than existing ones. The authors posit that grants might mitigate capital constraints for new entrepreneurs, whereas for existing entrepreneurs, capital might not be the binding constraint to growing the business. The size of the grants in the studies in low- or middle-income countries included in Kluve et al. (2017) range from $150 to $400.

Finally, in addition to the basic grants that are available to most program participants, NEF also plans to offer a form of acceleration which pairs advanced business training with an additional financial support for a select group of businesses with a high promise for growth. Stand-alone studies on acceleration and incubation programs—such as those that offer seed capital, coworking space, mentoring, and networking events—are limited compared to those on more traditional business training practices, because acceleration programs are often an element of a more comprehensive business training package (McKenzie and Woodruff, 2021), as is the case with the refugee livelihoods program. However, Gonzalez-Uribe and Leatherbee (2017) use a regression discontinuity design in Chile and find that a business accelerator program that provides funding and coworking space did not impact sales, number of employees, or business survival rates. Businesses that performed higher on these measures after the accelerator program were those that showed more promise initially.

b. Microfinance programs

Microfinance programs are a common alternative source of capital for microentrepreneurs. Banerjee et al. (2015) conducted a seminal review of six RCTs on microfinance programs in Bosnia and Herzegovina, Ethiopia, India, Mexico, Mongolia, and Morocco targeted toward microentrepreneurs. They find that microfinance programs have positive average impacts on business outcomes, such as entrepreneurial investment, business size, and profits in some contexts, but that there is no average impact on household income or consumption. However, Morduch (2020) has pointed out the limitations of RCTs as a method to assess the average impact of microfinance since most studies only measure the impact of microfinance on borrowers to whom the program is expanded as part of the study and not on the core group of borrowers. In addition, microfinance RCTs tend to be statistically underpowered because of low take-up rates, making it difficult to detect meaningful impacts on outcomes like household income or consumption (Dahal and Fiala 2020). The average impacts of microfinance programs might also mask important variation by subgroup. Specifically, there is strong evidence that microfinance programs have larger impacts for existing businesses than new businesses (Meager 2019; Banerjee et al. 2019), likely because entrepreneurs who managed to establish their business when credit was tight are those with the highest potential returns to investment. (This contrasts with Kluve et al.'s [2017] finding, discussed earlier, that business grants have larger impacts for new business owners than existing ones.)

c. Unconditional cash transfers

Like business grants, unconditional cash transfers might alleviate capital constraints to starting a business. (As mentioned in Chapter II, in the refugee livelihoods program, the grants for refugees are unconditional.) There is some evidence that grants earmarked for businesses appear to be more effective than unconditional cash transfers at improving business-specific outcomes. In an RCT in Somalia, Hassan et al. (2018) find that business training combined with grants (ranging from $100 to $1,000) increase the likelihood of establishing a business, income, savings, and food security, and that higher grants translate into larger impacts on household income. They compare business grants of various sizes to an unconditional cash transfer and show that business grants achieve larger effects on business ownership, business income, savings, and food security.

Nevertheless, unconditional cash transfers can still be effective in improving business and household outcomes. An RCT by Haushofer and Shapiro (2016) in rural Kenya finds that transfers ranging from about $400 to $1,500 increased business revenues (though not profits), increased the average number of income-generating activities reported by households, and increased household consumption by 23 percent. In another RCT on the impact of unconditional cash transfers in rural Kenya, Egger et al. (2019) find that these transfers had positive spillovers to the broader local economy, increasing revenues of non-recipient firms and consumption of non-recipient households in the areas in which they were disbursed.

In the refugee context, most studies have focused on cash transfers as part of a humanitarian approach. In a systematic review of the effects of cash transfers in humanitarian settings, Doocy and Tappis (2017) find that these transfers improve household-level outcomes for refugees. The total size of the unconditional cash transfers in the included studies range from $130 to $575, with a median of slightly over $200. In an RCT evaluating the impact of unconditional cash transfers sent to Syrian refugees in Lebanon, Lehman and Masterson (2014) find positive impacts on school enrollment, child labor supply, and negative coping strategies. These results are promising for NEF's program, which will also target Syrian refugees. However, NEF's approach differs by tying the grant to a comprehensive program to support new businesses, even if grants are technically unconditional for refugees.

3. Effects of entrepreneurship programs on women

Because most participants in the refugee livelihoods program are expected to be female, we examined the literature on how entrepreneurship programs affect women specifically. Women are often participants in programs that implement business training, grants, or both, although not all programs are tailored specifically toward them. Chinen et al. (2017) conduct a systematic review of 35 rigorous studies of vocational and business training programs for women in low- and middle-income countries. The authors find that programs that are tailored to women's needs—both in their provision (for example, providing childcare while women participate) and in their content (for example, discussions of gender norms that inhibit women's ability to develop a busines)—tend to have the strongest impacts on employment and income. Consistent with this, McKenzie and Puerto (2021) find that a business training for female entrepreneurs that had a gender component (such as encouraging collaboration with other women business owners and emphasizing the benefits of dividing household and business tasks) raised business profits and sales, an effect which was sustained three years after the training ended. One explanation for these findings is that female entrepreneurs in low- and middle-income countries often face social constraints—such as expectations that they are responsible for household tasks such as cooking and childcare—that men often do not. Therefore, easing these constraints through gender-focused program components can lead to positive business outcomes. Other examples of gender-focused adjustments to

business support programs suggested by Buvinić and Furst-Nichols (2014) are providing capital in-kind or through a private mobile phone and setting up savings accounts that are earmarked for the business—both of which limit interference from a woman's spouse and family.

In contrast, programs not tailored to women often find effects that are bigger for men than for women, or in some cases find no effect for women at all. In an RCT with young business owners in Uganda, Fiala (2015) finds that a grant of $200 improves the outcomes of men's businesses, but not those of women. In fact, the businesses of married women in the treatment group with extended family nearby had worse outcomes than the control group. Fiala (2015) proposes that these negative effects could be a result of pressure from the women's husbands or extended family to spend the grant on family needs. This is consistent with Buvinić and Furst-Nichols (2014)'s hypothesis that gender-specific social dynamics can play a significant role in modulating outcomes for women. In an RCT in rural Pakistan, Giné and Mansuri (2014), find that a business training and loan lead to better business knowledge and higher business survival rates, but mainly for men. The effect on women was relatively limited: they improved their business knowledge but did not exhibit changes in other outcomes. Like Fiala (2015), Giné and Mansuri (2014) hypothesize that social norms limited the labor supply of women, preventing them from taking full advantage of the business loans.

Finally, some studies have examined the impacts of entrepreneurship programs on indicators related to women's social and economic empowerment. For example, in the RCT of a female entrepreneurship program in Afghanistan discussed earlier, Bedoya et al. (2019) find a large impact on an index of women's empowerment that includes, among other components, women's influence over fertility and mobility, ownership of mobile phones, saving or borrowing, and participation in income-generating activities or being the owner or manager of a self-employment enterprise. In Vietnam, an RCT of a widely implemented program for female entrepreneurs that includes both entrepreneurship and empowerment components found positive impacts on several aspects of empowerment, including women's influence on intrahousehold decision-making (Huis et al. 2019). However, these effects were not immediate and only emerged about one year after the end of training.

4. Contribution and policy relevance of the evaluation

Although the varied content, intensity, and context of other programs makes it challenging to compare them to one another and to the refugee livelihoods program, the available evidence suggests that the refugee livelihoods program has the potential to achieve impacts on business outcomes and related household outcomes. Many of the programs studied in the literature include multiple components, but the refugee livelihoods program is among the more comprehensive because it includes business skills training, TVET, mentoring, financial grants, and business acceleration. Additionally, the grant of $850 included in the program is among the largest grants provided in comparable programs, even when adjusting for local purchasing power.[5] The main contribution of the evaluation will be to assess the extent to which this promising program can be effective in the highly challenging refugee context.

The findings from the evaluation will be key in demonstrating a proof-of-concept—of both the program itself and the DIB financing mechanism—to attract resources for NEF or other organizations to

[5] The $850 grant in NEF's program amounts to about $2,000 once adjusted for purchasing power parity (World Bank 2020). This is considerably larger than the grant sizes in de Mel et al. (2014), Blattman et al. (2016), Blattman et al. (2014, 2020), and Bedoya et al. (2019) which range from approximately $400 to $800 once adjusted for purchasing power parity. In Giné and Mansuri (2014), businesses received a higher sum of about $4,000 once adjusted for purchasing power parity, but this was a loan rather than a grant.

implement similar programs for refugees and vulnerable populations elsewhere. Further, this proof of concept might encourage the governments in Jordan or the region to consider adopting a version of the program for its own vulnerable citizens.[6]

[6] The cost of the program will be a critical consideration for potential scale-up and replication even if it shows promising results in terms of outcomes and impacts. Therefore, we will explore with IKEA Foundation the possibility of adding a cost and cost effectiveness analysis to the evaluation. The cost analysis would use the up-front and operating costs of the program to estimate a program cost per beneficiary, which could be used to assess the costs for implementation at different scales. The cost-effectiveness analysis would combine estimates of costs and outcomes to estimate, for example, the cost per active IGA or per dollar increase in household consumption.

IV. Evaluation design, data collection, and analysis approach

In this chapter we present the key research questions that the evaluation will seek to address and describe the evaluation design that we propose to answer them. This evaluation design comprises three components: (1) IGA validation; (2) an impact evaluation; and (3) a process evaluation. For each of these components, we define the outcomes (or qualitative themes) we will measure, including how these outcomes will translate into DIB payments. We also detail our approach to data collection, sampling, and analysis. We conclude by describing our approach to data quality assurance and data privacy protection.

1. Research questions and overview of evaluation approach

Table IV.1 presents the proposed research questions for the evaluation. These questions cover the two payment metrics (shown in bold) but also seek to assess the broader effects of the program on participants' social and economic wellbeing, and the mechanisms and context underlying these effects. To answer these questions, we propose a rigorous mixed-methods evaluation comprising three components:

- **IGA validation**, which will seek to estimate the total percent of grant recipients who are actively engaged in IGAs about 10 months after grants are disbursed, based on a short survey with a representative sample of grant recipients from each of the three cohorts. This metric, known as the "business metric," is the primary metric that will be used to determine DIB payments.[7]
- An **impact evaluation**, which will assess the impacts of the program on household consumption and other outcomes related to social and economic wellbeing for participants in the first cohort about 24 months after grants are disbursed. To estimate these impacts, we will implement a matched comparison group design with a comparison group of participants from the third cohort, who will just be starting the program at that point in time. The impact evaluation will draw on a household survey conducted with the treatment and comparison groups. The estimated impacts on household consumption will be the secondary DIB payment metric, known as the "household consumption metric," used to determine additional DIB payments.[8]
- A **process evaluation**, which will seek to summarize the programmatic context, explore participants' experiences with and perceptions of the program, and identify facilitators and barriers to achieving the outcomes envisaged in the program logic. The process evaluation will draw primarily on the analysis of qualitative data from focus group discussions (FGDs) with program participants in the first two cohorts, as well as semi-structured interviews with program implementers. We will complement this with a descriptive analysis of quantitative program monitoring data collected by NEF and its partner CBOs.

[7] DIB payments to investors will comprise a base lump sum plus an additional payment for each tenth of a percentage point by which the business metric exceeds 44 percent, up to a cap of 75 percent. The payment per tenth of a percentage point decreases once the business metric exceeds 55 percent, the minimum target for the DIB.
[8] DIB payments for the household consumption metric will only be made if the business metric exceeds the minimum target of 55 percent. In that case, payments to investors will begin for impacts on household consumption of 0.22 standard deviations and will increase for each additional impact of 0.01 standard deviations up to a cap of 0.38 standard deviations. If the cap is reached, no additional payments will be made to investors beyond the cap, but NEF becomes eligible for an outcomes payment based on the business metric. Specifically, NEF will receive a payment for each tenth of a percentage point by which the business metric exceeds 67.5 percent, up to the original cap of 75 percent.

We envisage strong synergies and integration between the three components to generate a comprehensive body of evidence that can be used to guide possible scale-up and the design of similar investments in the future. In the following sections, we describe each of the evaluation components in further detail.

Table IV.1. Research questions

Research question	Evaluation component
1. What percent of grant recipients were actively engaged in IGAs 10 months after grant disbursement?	IGA validation
2. What were the impacts of program participation on social and economic wellbeing 24 months after grant disbursement?	Impact evaluation
a. What were the impacts on household consumption?	
b. What were the impacts on household savings?	
c. What were the impacts on household durable asset stocks?	
d. What were the impacts on participants' self-confidence?	
e. What were the impacts on women's social and economic empowerment?	
f. What were the impacts on other outcomes related to social and economic wellbeing (including coping strategies and food security, sense of safety and wellbeing, school enrollment and attendance, and receipt of social assistance and social protection)?	
g. How did these impacts vary by subgroup (for example, by gender, refugee status, and age)?	
3. What were the key elements of the program that led to achieving the desired program outcomes?	Process evaluation
a. Did participants understand and find value in the training content? Which components of training were the most valuable?	
b. What was the role of the cash grants in triggering the successful launch of individual businesses?	
c. Do participants view the program's supplementary support services and activities as valuable? How have they engaged with these support services?	
4. What is the community and business environment in which participants live and work?	Process evaluation
a. What are the barriers and facilitators to business growth and sustainability at the individual, household, and community levels?	
b. How has participants' level of community integration changed because of the program?	
c. How has participants' awareness and use of other social protection schemes changed because of the program? How has this influenced achievement of the desired program outcomes?	

Note: Payment metrics are highlighted in **bold**.

IGA = Income generating activity.

2. IGA validation

The purpose of the IGA validation component is to estimate the business metric that determines the bulk of the payments to DIB investors. This business metric is defined as the percent of grant recipients across all three cohorts who have an active IGA about 10 months after grant disbursement, either through

businesses supported by grants or through formal employment.[9] We will measure this metric through short in-person surveys of a random sample of grant recipients in each of the three program cohorts.

1. Outcome definition

The business metric is intended to measure active IGAs among grant recipients, who are expected to comprise the majority of program participants. The DIB parties have agreed to define an active IGA as (1) an active business, which is one that "has conducted at least one business transaction in connection with the grant received from NEF in a short reference period before the data audit," or (2) formal employment that meets minimum conditions in terms of average hours per week and wages. Below, we clarify how we will operationalize these definitions for the evaluation; in the next subsection we describe how we will measure active IGAs following these definitions.

a. Active businesses

To operationalize the definition of active businesses, we plan to define the key terms as follows:

- **Business transaction.** We will define a business transaction as the exchange of a good or service for cash or kind (that is, a sales transaction) or the acquisition of goods or services related to business for cash or kind (that is, a purchase transaction). Further, because it is likely that some businesses supported by the grants have less frequent but higher value transactions, we will also classify a business as active if the grantee was engaged in production-related activities related to the business in the reference period, even if they did not have a business transaction over this period. These activities could include manufacturing products from raw materials for sale at an upcoming market fair or providing a large order of goods and services that has not yet been paid for. They would indicate that the business was active and likely to make a business transaction in the near future.
- **Connection with the grant received.** The connection with the grant received will be based on grantees' self-reports. Specifically, we will ask the grantee what business they established using the grant and then ask about transactions associated with that business. This is aligned with NEF's measurement approach for previous iterations of the program.
- **Short reference period.** We will define the reference period for a business transaction as two months (60 days) prior to the IGA validation survey. Based on discussions with NEF, this reference period

[9] The logic of including formal employment in the definition is that increased skills, knowledge, and self-confidence, as well as stronger networks obtained through DIB-funded training, might lead to -previously inaccessible- formal employment opportunities for the grant recipients.

should appropriately capture business activity for almost all types of businesses that the program is expected to fund under its implementation model.[10,11]

b. Formal employment

The business metric definition also allows for grantees to count as having an active IGA if they are not operating their own businesses but instead have found formal employment. To qualify, their employment must meet certain conditions: (1) they have an employment contract or contribute to the Jordanian social security system through their employer (formal employment); (2) they worked an average of at least 20 hours weekly over the previous month; and (3) their monthly wages during the previous month are equal to or greater than the minimum wage for their nationality and sector of employment (at present 260 Jordanian dinars per month for Jordanians, 230 dinars for non-Jordanians, and 220 dinars per month for those in the garment and textile industries and who load or unload goods in the transportation sector [International Labour Organization 2021]).

Because wages are typically received per calendar month, we will define the previous month for hours and wages as the last complete month before the survey. For example, if we conduct the survey sometime in February, we will ask about the respondent's current formal employment status but will capture hours and wages for the month of January.[12] If the respondent began their current employment after the start of January, we would capture average hours and wages for January and prorate them into monthly measures. If the respondent began their current employment in February, we would capture expected hours and income for the month of February and prorate them into monthly measures.

2. Data collection approach

We will use administrative data from NEF to obtain the list of grant recipients, their socio-demographic characteristics, and their contact details for each of the three cohorts. Information about the types of businesses funded by the grants will also provide valuable context as we develop the IGA validation survey. We will conduct the IGA validation for each cohort about 10 months after the expected midpoint of grant disbursement through short, in-person surveys; in-person surveys will be critical to building rapport with respondents and conducting the observations that we have proposed (for example, reviewing

[10] The entrepreneurship and microenterprise literature offers limited guidance on the appropriate reference period because existing studies typically measure business survival based on visiting the business during its operating hours or using simple self-reports of whether the business is active (for example, whether they are working in "the same line of business" [de Mel et al. 2014] or whether the business "still sells any goods" [Calderon 2013]). In its monitoring and evaluation efforts for previous iterations of the program, NEF has also primarily used a self-reported measure of whether the business is currently active. A reference period approach is more common when gathering information about income, expenditure, revenue or profits following business training interventions; for those purposes, a one-month reference period is common (Berge et al. 2012; de Mel et al. 2014; Anderson et al. 2018; Brooks et al. 2018). However, based on discussions with NEF, a one-month (30 day) period might not be appropriate for many of the businesses supported by the refugee livelihoods program, which might have less frequent but higher value transactions.

[11] IGA validation for the first two cohorts will take place in the winter. If extreme weather events such as major snowfall or flooding affect any of the project locations, we will extend the reference period for the affected location(s) so that it covers 60 days of non-extreme weather conditions. We will also extend the reference period if other extreme events (for example, strikes or civil unrest) affect any of the project locations.

[12] If the respondent was employed in a formal job in January but was no longer employed when we conducted the survey in February, we would not count them as formally employed.

records). Specifically, we will conduct the validation in mid-2023 for the first cohort, early 2024 for the second cohort, and mid-2025 for the third cohort.

Because, as members of vulnerable populations, the respondents may be reluctant to participate in "audits" of their businesses, we will emphasize that the objective of the survey is to understand the respondents and their business practices, with the goal to assess the quality of the program and improve it for future participants. This framing is consistent with the approach to audits of other business outcomes (typically business practices) conducted as part of studies of entrepreneurship programs in other low- and middle-income countries. For example, Anderson and McKenzie (2020) framed their survey in Nigeria as an effort to better understand businesses in priority sectors, while in Sri Lanka the research team paid business mentors to collect information about businesses while at the same time providing suggestions for areas of improvement (de Mel et al. 2014). As part of the data collection training we will also emphasize the importance of approaching the grantees with sensitivity, building rapport with them, and emphasizing our relationship with NEF and the *Siraj* centers.

a. Active businesses

The approach to measuring active business involves a tradeoff between the rigor of the evidence and the likelihood that such evidence is available. We therefore propose a tiered approach that prioritizes the strongest possible evidence but also accepts weaker—but still plausible—evidence in other cases. Specifically, we propose to count a business as active if it meets one of the following criteria, consistent with the definitions described earlier (all of these criteria pertain to the business established with the grant):

- The business is observed to be conducting a business transaction (sales or purchase) at the time of the survey.
- The respondent reports that the business is active and can show the enumerator a copy of a receipt, ledger, electronic record, or other documentation of a sales transaction that clearly shows it was conducted in the past 60 days. In keeping with the framing of the overall survey, we will frame the request to see records as an effort to understand grantees' business practices that were taught during training.
- The grantee reports that the business is active, that the most recent sales transaction was conducted within the past 60 days, and provides the following details of the transaction: (1) the date of the transaction; (2) the good(s) or service(s) provided to the customer; (3) the mode of payment and value of the transaction; and (4) a description of the customer, such as their gender, approximate age, whether they are a new or returning customer, and whether they had a preexisting relationship with the grantee before becoming a customer. Although it will not be possible for us to verify these details, being able to provide them makes it more plausible that the respondent is reporting a valid sales transaction.
- The grantee reports that the business is active and can show the enumerator documentation of a purchase transaction that clearly shows it was conducted in the past 60 days.
- The grantee reports that the business is active, that the most recent purchase transaction was conducted within the past 60 days, and provides the following details of the transaction: (1) the date of the transaction; (2) the good(s) or service(s) purchased; and (3) the mode of payment and value of the transaction. Purchases could include raw materials or supplies, inventory, equipment, payments to employees, and so on.

- The grantee reports that the business is active and in the past 60 days engaged in the production-related activities described earlier. Like the self-reports of sales and purchase transactions, we will require the grantee to provide additional details of these activities—for example, showing the enumerator a sample of the products produced and describing the event at which they are expected to be sold, or reporting a down payment and the total expected payment for the large ongoing order. To count for the business metric, the grantee will have to provide sufficient detail to suggest that they have a clear expectation to sell the goods or services produced.

We will pilot the survey instrument to test and refine this approach. If we find during piloting that respondents struggle with or are reluctant to share certain information, we will revise the survey and our criteria for the business metric accordingly while following the core approach described above. For example, piloting will enable us to assess the appropriate level of detail to require regarding self-reported transactions in the previous month, as well as to refine the list of purchases and production-related activities we will count towards the business metric.

To provide a more complete picture of grantees' business activities as part of the process evaluation described later, our IGA validation survey will also include a small number of additional questions about grantees' businesses, to the extent feasible given the relatively short survey that we have planned. This will complement information about these businesses from NEF's monitoring and evaluation efforts; we will work closely with NEF to maximize synergies and avoid duplication in the information collected. For example, for active businesses this information could include the number of sales transactions in the past 60 days, total revenues and costs (in the past 60 days, or more broadly for a good, average, or bad month), and the adoption of practices that have been shown to be correlated with the success of microenterprises in low- and middle-income countries (McKenzie and Woodruff 2017). For non-active businesses this could include when the last sales transaction occurred and whether the business expects to make sales transactions in the future.

b. Formal employment

We will ask the grantee to self-report the details of their employment—including whether they have an employment contract and/or contribute to the social security system through their employer, their average weekly hours, and their monthly wages—to confirm that their employment meets the conditions above.[13] However, like the approach taken to assess active business, we will collect additional information from grantees to increase the plausibility of their responses, again framing our questions as an effort to understand their experiences following training. Specifically, we will confirm that the grantee can provide all the following details about their role: (1) the date they began their current employment; (2) their sector of employment and job title; (3) the location where they work; (4) their mode of commuting to work; and (5) their typical work schedule. Again, we will use piloting to determine a reasonable level of detail to require to verify their formal employment status.

3. Sample sizes and sampling approach

Measuring the business metric outcome from a sample of grant recipients rather than all recipients maximizes efficient use of evaluation resources while still providing reliable estimates. In each cohort, we will stratify the IGA validation random sample by location, refugee status, gender, and youth status to ensure that the sample is representative along these dimensions. We will allocate the sample across strata

[13] Given the vulnerability of program participants, we do not plan to ask them to show us work contracts or salary slips, as this would likely lead to a reluctance to respond to the survey.

in proportion to the number of grantees in each stratum to avoid the need for sampling weights. In the case of survey non-response, we will randomly replace grantees with others in the same stratum to achieve the targeted sample sizes, using an ordered list of replacements that we will provide to Mindset ahead of the data collection effort. If a replacement in the same stratum is not available, we will have Mindset randomly select a replacement from another stratum in the same location.

We propose to survey 600 grant recipients per cohort, or about half of the expected number of grant recipients. A sample size of 600 grant recipients per cohort will provide relatively narrow confidence bands around the expected point estimates. Specifically, if the average of the business metric is 55 percent (the minimum target for the DIB), the 95-percent confidence interval on the pooled estimate for all cohorts will be 1.6 percentage points on either side of the mean, or 53.4 to 56.6 percent (**Table IV.2**). (Confidence interval margins will be narrower if the mean exceeds 55 percent.) As we describe below, we anticipate that the reported mean of the business metric will be used to trigger payments; however, we will also report confidence intervals to indicate the precision of the mean. Confidence intervals for women, youth, and refugee subgroups would be wider but would be used for descriptive purposes rather than for triggering payments.

Table IV.2. Confidence interval margins for the business metric, either side of a mean of 55 percent

Analysis sample	Per cohort	All three cohorts
Full sample	±2.7%	±1.6%
Women	±3.2%	±1.8%
Youth	±4.6%	±2.7%
Refugees	±5.0%	±2.9%

Note: We assume a sample size of 600 per cohort. The numbers in this table indicate 95 percent confidence interval margins on either side of the mean. The smaller the confidence intervals, the more precise outcome estimates will be. We assume that women will comprise 75 percent of the full sample, youth 35 percent, and refugees 30 percent. The calculations apply a finite population correction. DIB payments will be made based on the mean across all three cohorts.

4. Analytical approach

To estimate the mean business metric, we will divide the total number of grantee survey respondents with active IGAs approximately 10 months after grant disbursement by the total number of grantee survey respondents, across all three cohorts.

5. Implications for DIB payments

The mean business metric described above will determine payments from the funders. The following further clarifies how this estimate will translate into payments:

- **Statistical precision.** Payments will be based on our mean estimate of the business metric (the point estimate), regardless of the precision of this estimate. Our statistical power calculations above suggest that our proposed sample sizes will provide a precisely estimated point estimate. However, this point estimate will not be exact; there will still be some degree of uncertainty, represented by the confidence interval around the estimate. To be clear, payments will be based only on the value of the point estimate and statistical precision will not be considered in the incidence and size of payments.

- **Rounding.** We will round down the estimated business metric to the nearest tenth of a percentage point to determine payments. This follows the rationale that payment cutoffs are thresholds to be met or exceeded before payments are made. For example, if we estimate the DIB business metric at 66.68 percent, the funders will then make payments based on reaching the payment threshold of 66.6 percent.
- **Reporting.** For greater transparency, we will report not only the final business metric but also the components that were used to calculate it: (1) the active business component; and (2) the formal employment component. We will use a reporting template as shown in **Appendix C**.

3. Impact evaluation

The impact evaluation has the dual purpose of (1) determining additional payments to investors based on impacts on the household consumption metric, and (2) assessing impacts on other short- and medium-term outcomes in the program logic. To provide rigorous evidence on the impact of the refugee livelihoods program on household consumption and other outcomes, we will use a matched comparison quasi-experimental design. Specifically, we will leverage the staggered rollout of the intervention across cohorts to compare the outcomes of participants in the first cohort with the situation of those entering the third cohort (henceforth referred to as the "outcomes" of the third cohort), about 24 months after grant disbursement for the first cohort and around the time when the third cohort is starting the program (mid-2024). Although the timing of outcome measurement is linked to grant disbursement for the first cohort, the impact evaluation is intended to include all participants except for no-shows—including those who did not receive grants—to fully capture the impacts of the program. We will use a statistical matching approach to improve the comparability between these two groups in terms of observed characteristics that are expected to be correlated with household consumption and other outcomes.

The matched comparisons group design seeks to limit the possibility that differences in outcomes between the two cohorts are driven by differences in participant characteristics, making it more plausible to attribute these differences to the impact of the program.[14] Specifically, because the first and third cohorts will have been selected for the program using the same approach, we expect them to be broadly similar in terms of vulnerability and unobserved characteristics related to program participation and outcomes (for example, motivation and entrepreneurial spirit). The statistical matching approach will further limit the possibility that differences in outcomes between the two cohorts are driven by differences in observed socio-demographic and socio-economic participant characteristics.

1. Matching approach

Our statistical matching approach will rely on data from the vulnerability assessment used to screen potential participants. An important complication is that the vulnerability assessment for the two cohorts will be administered at different points in time, so time-varying characteristics might not be comparable.

[14] An experimental design would have ensured that the treatment and control groups were statistically similar in both observed and unobserved characteristics so that any differences between the two could be confidently attributed to the impacts of the program. This design would have involved randomly assigning eligible participants to receive the program as part of the first cohort (the treatment group), or to receive it as part of the third cohort or not at all (the control group). However, an experimental design was not feasible in this context because withholding services for eligible participants for an extended period was not in the spirit of NEF's operations in Jordan, was incompatible with participant recruitment plans, and posed a reputational risk to NEF. Nevertheless, our matched comparison design can deliver plausible estimates of the program's impact provided that differences in characteristics between the treatment group (first cohort participants) and the comparison group (third cohort participants) are minimized.

In **Appendix A**, which presents the technical details of the proposed matching approach, we describe our plans to account for this by conducting statistical matching using only characteristics that are effectively time-invariant or measured using the same reference period. The former includes socio-demographic characteristics such as year of birth (translated into age at a fixed point in time), gender, refugee status, and education level. The latter includes housing characteristics and ownership of common durable goods, which we will use to construct an asset index to use for matching (Filmer and Scott 2012). It will be necessary to capture these characteristics retrospectively for the third cohort when they are administered the vulnerability assessment in early 2024, using a reference period of early 2022, which is when the first cohort is administered the vulnerability assessment. We will work with NEF to incorporate the relevant characteristics into the vulnerability assessment for both cohorts.

2. Outcome definitions

We will measure outcomes through an impact evaluation survey (distinct from the IGA validation survey) to be conducted with the first and third cohorts 24 months after grant disbursement for the first cohort. Below, we define the outcomes that we plan to include in this survey.

a. Consumption

Household consumption—the secondary payment metric—will be the focal outcome measure for the impact evaluation. We will calculate overall household consumption from expenditures on four main components of consumption, plus debt repayments:[15]

1. Food items
2. Non-food items (for example, clothing, transportation, recreation, health, and education expenses)
3. Durable goods (for example, household appliances and cars)
4. Housing (rent or implicit rent)
5. Debt repayments

We will measure these items through a detailed consumption and debt repayment module that we will design following guidance from the literature and examples from other validated surveys, especially in the Jordanian context. In particular, the items included in the consumption questions of the 2017-2018 Jordan Household Income and Expenditures Survey (HIES) will serve as our starting point. We will adhere to standard practices when computing overall consumption for households. Specifically, Deaton and Zaidi (2002) provide a general guide for calculating consumption in the categories above, given differences in reference periods, household composition, and prices. We will draw on additional examples from the World Bank's Living Standards and Measurement Study (LSMS) to resolve measurement issues related to individual items.[16] For example, we will follow LSMS conventions on how to address item non-

[15] The expected program participants are heavily indebted, having approximately $2,700 in debt, on average (Keaveney et al. 2018). The cash grants provided by the program could alleviate this debt and indirectly contribute to increased consumption by the household in the timeframe of the evaluation. It was for this reason that KOIS proposed during the DIB design phase, and the funders decided, to include debt repayments in the consumption aggregate. However, because repayments of debt, including any interest payments, are normally excluded from standard measures of consumption (Deaton and Zaidi 2002), we will also compute and report the consumption aggregate without debt repayments for better comparison of program impacts to the literature.

[16] The LSMS, the World Bank's flagship household survey program, has been administered in approximately 40 countries since 1980 to measure welfare and living standards around the world.

response and outliers, as we describe below. **Table IV.3** presents a preliminary list of consumption items that we propose to include or exclude from measurement based on theoretical and practical reasons.

Table IV.3. Preliminary list of consumption items and measurement considerations

Included and excluded items	Measurement considerations
Food items	
Included: • Food and beverages purchased from the market • Food and beverages that are home produced • Food and beverages received as a gift or in-kind payment • Meals consumed in restaurants • Meals consumed at work Excluded: • Meals consumed at school • Meals consumed on vacation	We will omit meals consumed at school because they might be challenging to measure accurately and would likely involve relatively small consumption values. We will exclude meals consumed on vacation as we expect these to be an atypical expense for study participants.
Non-food items	
Included: • Clothing and footwear • Transportation • Communication • Recreation • Household furnishings, utilities, housewares, cleaning materials, home repairs • Health expenditures • Education expenditures • Miscellaneous goods and services (for example, personal care products, other accessories, etc.) Excluded: • Taxes and levies • Marriage and dowry • Funerals • Births • Gifts to other households • Remittances to other households • Charitable contributions	We will exclude taxes and levies because these reduce income but are not consumption. We will exclude relatively large and infrequent purchases related to marriage, dowry, funerals, or births because these do not capture typical consumption of the household during the year, which the consumption aggregate aims to capture. We will exclude gifts, remittances, and charitable contributions because these are transfers of income and not consumption.

Included and excluded items	Measurement considerations
Durable goods	
Included: • Vehicles (for example, car, motorcycle, bicycle) • Household appliances • Computers • Mobile phones	We will calculate the "rental equivalent" or flow of services that a household derives from using durable goods during the year, not the total expenditure of the household to purchase these items. The rental equivalent captures the value of services that the household receives from durable goods. This can be calculated using information on (1) the year the durable good was purchased, (2) how much was paid for it, (3) the household's estimate of its current value, (4) and the real interest rate in the country (see **Appendix B** for more details on the calculation). We will follow the Jordan HIES to capture information on items (1) to (3) in our survey. We will obtain information on the real interest rate from official government sources.
Housing	
Included: • Rent (or implicit rent)	For renters, we will use the monthly rent paid by the household to calculate the value or benefit that a household receives from occupying a dwelling in a year. For non-renters (including those who own their dwelling and those who are living rent free), we will use the estimated market rental rate for the unit. For households living in refugee tents, we will set the consumption value of their housing to zero, as it will not be possible to impute a rental value for these tents.
Debt repayments	
Included: • Payments for debt incurred to formal and informal sources before the first cohort starts the program Excluded: • Payments for debt incurred after the first cohort starts the program	We will exclude repayments on debt incurred after the first cohort starts the program to avoid double counting, because credit used to purchase goods in the measurement period would be counted under other consumption categories. For example, if a participant borrows 100 dinars to purchase a durable good and then repays that debt, they would have consumed 100 dinars' worth and not 200 dinars' worth.

We will address other key issues in survey design and measurement of household consumption as follows:

- **Method of data capture.** We will use the recall method rather than consumption diaries to capture the value of consumption. This is consistent with the 2017-2018 Jordan HIES. Evidence from Beegle et al. (2012) suggests that the recall method performs almost as well as personal consumption diaries for each member of the household with intensive and frequent supervision—the gold standard—without being nearly as burdensome to respondents. The same research shows that the recall method can perform better than unsupervised household consumption diaries.
- **Reference period.** We will vary the recall period across different types of items to improve accuracy, following the recall periods used in the Jordan HIES. We will use a short reference period for frequently purchased items (for example, 7 days for food, 30 days for utilities, and 90 days for health

expenditures) and a longer reference period for infrequently purchased items (for example, 12 months for educational expenditures and household appliance purchases). We will scale reported consumption of these items to obtain annualized values.

- **Level of detail.** We will carefully select the items included in the survey to provide a reliable proxy for full consumption while limiting respondent burden. Beegle et al. (2012) shows that including more items can improve accuracy with relatively little added respondent burden. However, the Jordan HIES gathers consumption information on more than 500 items, and our survey instrument will need to capture data on secondary outcomes as well; capturing all these items will not be feasible. Thus, we will consider dropping items from the Jordan HIES that represent a very small share of consumption based on the 2017-2018 data (for example, Jordanians reported almost no expenditures in CD/DVD/tapes in 2017-2018). We will also consider collapsing some items into broader categories as appropriate (for example, "yellow cheese packed," "yellow cheese canned," and "yellow cheese large" can be collapsed).

After data collection, we will compute the consumption aggregate for the household, adjusting for household size. When calculating the aggregate, we will minimize the influence of outliers and missing values arising from item non-response (answers of "don't know") in the data by replacing these with the median value for households in the same study group and location, following the LSMS approach.[17] We discuss the formulas and technical details for computing the consumption aggregate in **Appendix B**.

b. Other impact metrics

An increase in household consumption at 24 months after grant disbursement—the second payment metric—is a key medium-term outcome in the program logic. However, there are also many other outcomes that reflect various dimensions of progress towards the program's long-term outcomes. To provide a deeper understanding of program impacts that might inform broader learning, we will also measure several of these outcomes alongside the consumption metric. In this section, we describe the additional outcomes we plan to measure and how we plan to measure them.

Because accurately measuring the household consumption payment metric will require a long and time-consuming module, we carefully weighed the utility of measuring each potential additional outcome with the increase in survey administration time to limit respondent fatigue and ensure high-quality data. To determine the appropriate measurement approach for each outcome, we drew on relevant surveys in the Jordan and Syrian refugee contexts. These include the Jordan Demographic and Health Survey (DHS), the Jordan Syrian Refugees and Host Communities Survey, the NEF vulnerability assessment, and the World Food Programme's Renewed Efforts Against Child Hunger survey. We supplemented measures from these context-specific surveys with other widely used metrics that have been validated in other contexts. **Table IV.4** lists the measures we recommend including—subject to survey piloting—organized by domain, and the rationale for including them.[18]

[17] This approach to missing values was taken by almost all LSMS surveys we have been able to review. See for example the documentation provided in this link for Bosnia and Herzegovina, Ethiopia, and Nigeria: https://microdata.worldbank.org/index.php/catalog/lsms.

[18] We determined that qualitative work would be better able to capture several outcomes—including social cohesion between refugees and the local community and child labor—discussed at the proposal stage, and we therefore propose excluding questions related to these outcomes from the household survey.

Table IV.4. Additional outcome metrics

Domain	Recommended measure(s)	Rationale for inclusion
Household savings	• Total household savings (NEF vulnerability assessment), amount of money household holds in savings	• Straightforward measurement: a single number at a single point in time • Debt will be captured separately as part of the consumption module • A medium-term outcome in the program logic
Self-confidence	• Rosenberg Self-esteem Scale (Rosenberg 1965), a 10-item scale measuring positive and negative feelings towards the self	• Classic measure in use for decades, enabling comparisons across studies • Validated across cultures (Schmitt and Allik 2005; Baranik et al. 2008; Haushofer and Shapiro 2016) • Also used in the Middle East and in Syrian refugee populations (Ceylan et al. 2017; Miller et al. 2020) • Short—10 questions, each on a Likert scale • A short-term outcome in the program logic
Women's social and economic empowerment (female participants only)	• Extent of women's influence in household spending decisions (DHS, Donald et al. 2017), a series of questions about spending money that ask about influence on a 3-point scale • Spousal respect (project-level Women's Empowerment in Agriculture Index), questions asking to what extent each spouse respects the other, trust one another to act in best interest of the other, and feels comfortable expressing disagreement, each on a 4-point scale • Women's freedom of movement (Yount et al. 2016), a series of questions about ability to independently visit places (market, doctor, neighbor, etc.) on a 4-point scale • Phone and bank account ownership, internet use (DHS)	• A medium-term outcome in the program logic, women's social and economic empowerment is a complex set of concepts that requires a multi-faceted measurement approach • Women's influence in decisions on spending money and spousal respect are linked to the medium-term program logic outcome of increased female bargaining power in the household; because influence in decisions is non-binary, it is important to capture influence on a scale • Women's freedom of movement is linked to mobility and independence, a medium-term outcome in the program logic • DHS questions on phone, bank, and internet are straightforward measures linked to the medium-term program logic outcome of women's independence
Coping strategies and food security	• Consumption-based Coping Strategy Index, measuring strategies to adjust food consumption to bridge limited availability of food in the short term (Maxwell et al. 2008) • Livelihood-based Coping Strategy Index (World Food Programme), measuring broader strategies to meet basic food needs (for example, spending savings, buying food on credit, selling belongings or assets, begging)	• Consumption-based index can be interpreted as a measure of food security, a long-term outcome in the program logic • Livelihood-based index provides a broader measure of harmful coping strategies, a reduction in which is a long-term outcome in the program logic

Domain	Recommended measure(s)	Rationale for inclusion
Sense of safety and well-being	• Diener's Satisfaction with Life Scale (Diener et al. 1985), a short 5-item instrument designed to measure global cognitive judgments of satisfaction with one's life	• Improved well-being is a long-term outcome in the program logic • Validated across cultures and has an official Arabic translation • Used for refugee populations (Hussam et al. 2021)
School enrollment and attendance	• Whether child is enrolled in school and if so, how many days they attended in the past week when school was open (NEF vulnerability assessment)	• Links to long-term program logic outcome of increased resilience and ability to meet basic needs
Receipt of assistance and social protection	• Types of assistance received, from whom, and whether received assistance in past month (Syrian Refugees and Host Communities Survey)	• Linked to short-term program logic outcome on increased knowledge of other types of support that could contribute to increased household income

3. Data collection approach

We will survey households in the treatment and comparison groups in mid-2024, about 24 months after the midpoint of grant disbursement for the first cohort. Given the sensitivity of some of the outcomes we plan to measure (for example, related to women's agency) and the complexity of others (for example, consumption), we will conduct in-person surveys. We will also take advantage of these visits to ask about active IGAs among the first cohort. This will provide descriptive longer-term evidence of this outcome, although given survey length limitations we will likely use a less rigorous approach (focused on simple self-reports) than that used for IGA validation at 10 months. If feasible, we will also include simple questions on business revenues and costs for the first cohort—like those we plan to include in the IGA validation—to help us better understand the impacts on consumption. For example, if we find positive impacts on consumption, examining business profits will help us assess whether these were likely to be driven by the respondent's business or by other sources of income.

4. Sample sizes and sampling approach

For the impact evaluation, we estimate that our proposed sample of 930 individuals in the treatment group (first cohort) and 930 individuals in the comparison group (third cohort) will enable us to detect an impact of 0.12 standard deviations or larger on a standardized household consumption measure.[19] This is below the minimum target of 0.22 standard deviations beyond which additional payments based on the household consumption metric would be triggered. The proposed sample size will also enable us to assess program impacts on key subgroups of interest—women (0.14 standard deviations or larger), youth (0.20 standard deviations or larger), and refugees (0.22 standard deviations or larger)—although our ability to detect impacts for these groups will be less precise.[20] (These subgroup impacts are not linked to DIB payments, but will help inform the broader evaluation.)

[19] This minimum detectable effect is for a 95 percent confidence interval and 80 percent power. We will use replacements to achieve the targeted sample size in the case of non-response. Based on Ginn's (2020) analysis of refugee data in Jordan, our calculations assume that covariates will account for 15 percent of the variation in outcomes.

[20] We assume that women will comprise 75 percent of the sample, youth 35 percent, and refugees 30 percent.

As described in **Appendix A**, our sampling approach will vary slightly depending on the matching procedure selected, but broadly entails selecting a random sample of treatment and comparison participants to which we will apply the selected matching procedure. In the case of survey non-response, Mindset will randomly replace participants with others with the same treatment status and similar characteristics to achieve the targeted sample sizes, using a list of replacements that we will provide to them ahead of time.

5. Analytical approach

We will employ regression analysis to estimate the impact of the refugee livelihoods program by comparing outcomes in the matched treatment and comparison groups. For each outcome, we will estimate the following regression equation using ordinary least squares:

$$(1) \quad Y_i = \alpha + \beta_1 T_i + \beta_2 X_i + \varepsilon_i$$

where Y_i is the outcome for participant i; T_i is an indicator for treatment, equal to 1 for the treatment group (first cohort) and 0 for the comparison group (third cohort); and ε_i is a random error term. Although our matching approach will minimize any pre-existing differences between treatment and comparison groups, we will include a set of control variables, X_i, in the model to account for any remaining imbalance between the two groups (Ho et al. 2007; Stuart 2010). X_i will consist of all matching variables.[21] The coefficient, β_1, will provide a regression-adjusted estimate of the impact of the program, which is the difference in outcome means between the treatment and comparison groups after controlling for potential confounders that may contribute to the difference. We will estimate equation (1) for the full sample and for subgroups of interest, including by gender, youth status, and refugee status of the participants. We will incorporate appropriate weights in the analysis, depending on our chosen matching procedure.

We will also conduct similar analyses to estimate impacts on subgroups of treatment group participants who benefitted from various combinations of program interventions—for example, those who received the core training and grants, or those who received the core training, grants, and vocational training. To be credible, these analyses must account for potentially important unobserved differences across participants who received different combinations of program interventions. We will therefore restrict the comparison group for a given combination of interventions to third cohort participants in the impact evaluation sample who eventually received the same combination of interventions, and redo the matching using this restricted comparison group. (Given the evaluation timeline discussed in Chapter IV, there will be enough time to observe the program participation of the third cohorts and conduct these analyses before the final evaluation report.) The combinations of interventions that we will be able to examine using this approach will depend on the available sample sizes; if only a small fraction of the treatment group experience a certain combination, the impact estimates will be imprecise.

6. Implications for DIB payments

The impact estimate of β_1 in equation (1) for the household consumption aggregate (measured in Jordanian dinars), the secondary payment metric, will determine additional payments to outcome investors. To compute the standardized mean effect size, we will divide β_1 by the standard deviation of

[21] To avoid dropping any observations when estimating equation (1), we will apply the standard approach of replacing missing covariate data with the mean or median value and including missing data indicators for each covariate.

the comparison group's consumption aggregate. This provides the exact effect size that determines payments under the DIB's Outcomes Agreement.[22]

The following further clarifies how our impact estimate will translate into payments:

- **Statistical precision.** Like the business metric, payments will be based on our point estimate of the program's impact, β_1, regardless of the precision of this estimate.
- **Rounding.** We will round the standardized impact estimate down to the nearest hundredth to determine payments, following the same rationale as for the business metric. For example, if we estimate an impact of 0.257, we will report that the impact reached the payment threshold of 0.25.
- **Reporting.** For greater transparency, we will report not only the final impact estimate but also the components that were used to calculate it: (1) the regression-adjusted mean outcome of the treatment group; (2) the unadjusted mean outcome of the comparison group; (3) the difference between these two means (which corresponds to β_1 in the regression above); (4) the standard deviation of the outcome in the comparison group; and (5) the rounded standardized mean effect size (used to determine DIB payments) . We will use the reporting template shown in **Appendix D**.

4. Process evaluation

As described earlier, the process evaluation will seek to document the programmatic context, explore participants' experiences with and perceptions of the program, and identify facilitators and barriers to achieving the envisaged outcomes. It will draw primarily on qualitative data, comprising FGDs with participants and semi-structured interviews with program implementers. We will supplement the qualitative data analysis with a descriptive analysis of relevant program monitoring and evaluation data collected by NEF, which will likely include information on the characteristics of program participants, program attendance, and outcomes collected at baseline and at regular post-program intervals.[23] Below we describe the data collection and analysis approach in more detail.

1. Data collection approach

The FGDs and semi-structured interviews will be divided across the first two cohorts. This will provide an opportunity to assess the program both when it is initially rolled out and after CBO staff have more experience implementing it. Together, the FGDs and interviews will cover all five implementation locations, which will help us capture variation across CBOs and local conditions.

We will hold the FGDs and interviews soon after the IGA validation for each of the first two cohorts—about 13 months after the midpoint of grant disbursement. This timing will ensure that participants have had sufficient time to apply the skills and learnings of the training and mentorship, as well as sufficient experience in the marketplace. The FGDs and interviews for the first cohort will occur before we finalize

[22] The Outcome Agreements stipulate that the effect size will be calculated as the difference between the mean outcome of the treatment group and the mean outcome of the comparison group ($\mu_T - \mu_C$), divided by the standard deviation of the outcome for the comparison group (σ_c). β_1 in the analysis provides the regression-adjusted estimate of ($\mu_T - \mu_C$).

[23] Mathematica is coordinating with NEF as they develop their Monitoring, Learning, and Evaluation Plan to avoid duplication and maximize synergies between program data and data collected for the independent evaluation. Because this plan has not yet been finalized, the description of the process evaluation in this section focuses on the qualitative data.

the survey instrument for the impact evaluation, potentially enabling us to identify additional outcomes of interest to include in the impact evaluation (for example, specific dimensions of women's empowerment).

2. Sample sizes, sampling approach, and key themes

Across the two cohorts, we will hold 17 FGDs with program participants to understand their experience in the program, gathering their insights regarding the quality and content of the training, how they applied the skills from the training to IGAs, and how the grants were invested to support their livelihoods. We will also explore the challenges of starting or growing a business in the contexts where they live and work, and what aspects of the program functioned best to overcome these challenges. We will employ homogeneity sampling when selecting FGD participants, forming FGDs comprising specific sub-groups of grant recipients, which will reduce variation, simplify analysis, and facilitate group interviewing (Palinkas et al. 2015). For the first cohort, we will form FGDs comprising women, men, and youth (young women or men) in each of three locations. We will also conduct one FGD with participating women who did not receive grants and another with participating men who did not receive grants; these FGDs will help us assess barriers and challenges to moving through the program. (We will conduct these FGDs in one location— the location with the lowest grant recipient rate—but will assess whether it is logistically feasible to include participants from other nearby locations, as well.) For the second cohort, we will likewise conduct focus groups with women, men, and youth in each of the two remaining locations. To deepen our understanding of the programmatic context and program implementation—and how this might have affected participant outcomes—we will also conduct 15 semi-structured key informant interviews with representatives from NEF and all 5 implementing CBOs. **Table IV.5** summarizes the FGDs and semi-structured interviews that we plan to conduct, and the key themes we expect them to cover.

Table IV.5. Qualitative data collection for the process evaluation

Sample: cohort 1	Sample: cohort 2	Illustrative key themes
Focus group discussions (FGDs) with program participants[a]		
11 FGDs in total: • 3 with female grant recipients (3 locations) • 3 with male grant recipients (3 locations) • 2 with female youth grant recipients (2 locations) • 1 with male youth grant recipients (2 locations) • 1 with female non-grant recipients (1 location) • 1 with male non-grant recipients (1 location)	6 FGDs in total: • 2 with female grant recipients (2 locations) • 2 with male grant recipients (2 locations) • 1 with female youth grant recipients (1 location) • 1 with male youth grant recipients (1 location)	• Motivation for enrollment in the program, including timing of enrollment • Perceived value of various program components and suggestions for program improvement • Perceptions of the grant award process • Use of grants (or, for non-grant recipients, barriers to receiving grants) • Local market context • Barriers and facilitators to business growth and sustainability • Program effects on various aspects of social and economic wellbeing • Barriers to women's social and economic empowerment and program effects • Availability of other types of assistance and social protection in the community and program effects • Perceived level of social cohesion in the host community and program effects • Plans and expectations for livelihoods in the future
Semi-structured interviews with the NEF program manager and Jordan-based staff		
--	5 interviews in total: • Program manager • Local program coordinator • MEL manager and MEL officer (group interview) • Business development officers (group interview) • Capacity building officer	• Successes and challenges of implementation (including coordination between NEF and local CBOs) • Effects of NEF engagement on CBO capacity • Nature of and reasons for deviations from original implementation plans • Barriers and facilitators to business growth and sustainability • Lessons for future implementation

Sample: cohort 1	Sample: cohort 2	Illustrative key themes
Semi-structured interviews with NEF field officers and CBO staff		
2 group interviews with NEF field officers and 2 group interviews with CBO staff (2 locations)	3 group interviews with NEF field officers and 3 group interviews with CBO staff (3 locations)	• The beneficiary recruitment and admission process, including timing of recruitment • Successes and challenges of implementation (including coordination between NEF and local CBOs) • Effects of NEF engagement on CBO capacity • Nature of and reasons for deviations from original implementation plans • Perceptions of the grant award process • Perceived differences in program engagement by participant type • Local market context • Barriers and facilitators to business growth and sustainability • Lessons for future implementation

Note: Each focus group will comprise between 8 and 10 participants.

[a]If it proves infeasible to gather participants together for FGDs we will conduct interviews instead. Locations for the first and second cohorts will be different.

3. Analytical approach

We will analyze the qualitative data using a grounded theory approach, which allows for hypotheses to emerge from the data as they are analyzed rather than testing hypotheses that are developed in advance. This approach is appropriate given the open-ended nature of the research questions that the process evaluation is intended to answer, as well as the limited research on similar programs in similar contexts.

We will conduct the analysis of transcripts from the FGDs and interviews using MAXQDA software, which offers an efficient and flexible way to code and organize qualitative information. The initial coding scheme will be informed by the topics covered by the qualitative data collection, as well as by initial themes that emerged during data collection. Using an iterative process, we will expand and refine our coding approach to accommodate additional emerging themes as we examine the data in more detail. We will also include codes for respondent subgroup, such as program participants' gender, age, and refugee status. As we examine the data, we will carefully read answers to specific questions given by different respondents but will also read transcripts as a whole to ensure that context of specific answers is not lost. After grouping codes together into categories related to the key research questions, we will triangulate across respondents to identify commonalities and differences in experiences and perspectives, including by subgroup. This will enable us to develop to develop and interpret the key qualitative findings relevant to the research questions.

5. Data quality assurance

Data collection for the evaluation will be led by Mathematica's local evaluation partners, Mindset and Integrated, with oversight from Mathematica. Mindset will lead the survey data collection for IGA validation and the impact evaluation, and Integrated will lead the qualitative data collection (and analysis) for the process evaluation. Below we describe our plans to ensure data quality for these data collection efforts.

a. Survey data collection

Mathematica and Mindset will work together to implement quality assurance procedures throughout the data collection preparation, planning, and implementation processes for the IGA validation and impact evaluation surveys:

- During questionnaire development and programming, Mindset will pretest the draft questionnaire internally; conduct cognitive interviews with a small number of potential respondents to better understand how respondents are interpreting and responding to specific survey questions; modify the survey accordingly before programming the questionnaire into the survey software; and conduct an extensive piloting process in the field once programmed.
- Mathematica staff will travel to Jordan to attend training and observe several interviews for the first round of the IGA validation survey and for the impact evaluation survey.
- During survey implementation, Mindset's supervisors will accompany enumerator teams to ensure adherence to data collection protocols. Supervisors will regularly debrief with enumerators to address problems that arise (for example, unusual situations that are unclear how to record in the survey program) and provide feedback for improvement.
- Mindset will conduct data processing and cleaning to identify and address errors on an ongoing basis as the data are collected. They will conduct two kinds of backchecks: (1) randomly calling back selected interviewees to verify their responses to a subset of key questions and to monitor the performance of all enumerators; and (2) re-contacting respondents when the data processing team identifies problematic responses (for example, logically inconsistent responses or outliers).
- Mathematica will also assess pretest, pilot, and preliminary datasets to identify any emerging challenges. We will request that Mindset address these challenges either through additional guidance to field enumerators during data collection or post-survey processing and cleaning.

To provide further independent quality assurance oversight of the IGA validation and impact evaluation survey data collection conducted by Mindset, Integrated staff will attend the piloting and training for these surveys. Further, Integrated staff will observe several interviews in each location for each round of survey data collection (three rounds of IGA validation surveys and one round of impact evaluation survey) to assess adherence to the data collection protocols and identify any necessary course corrections.

b. Qualitative data collection

Before beginning the data collection, Mathematica staff will conduct remote briefing sessions with the two senior Integrated staff who will be conducting the FGDs and qualitative interviews to ensure that we have a common understanding of the goals of each question in the qualitative protocols. The first FGDs and interviews will serve as pilots; Mathematica staff will closely review the transcripts from these initial FGDs and interviews and debrief with Integrated to make any necessary adjustments. Mathematica will ultimately review all the transcripts and provide feedback to Integrated as necessary.

6. Data privacy protection

The evaluation will follow internationally accepted guidelines for protecting the privacy of respondents to the IGA validation survey, impact evaluation survey, and qualitative interviews and FGDs. Mathematica will prepare and submit institutional review board (IRB) applications for formal approval of the research

and data collection plans to registered IRBs in both the United States and Jordan.[24] The IRB applications will clearly describe our plans for obtaining informed consent from respondents and limiting the risk of their confidentiality being compromised. We will coordinate closely with our local evaluation partners to ensure that the data collection protocols, as well as mechanisms for data storage and transmission, are consistent with IRB requirements. As part of the IRB applications, we will also request that DIB parties be granted access to survey data with all personally identifying information removed to enable them to replicate our findings. This will be made clear to respondents as part of the informed consent statement. To obtain access to these data, the DIB Parties will be required to commit in writing to comply with IRB requirements and the applicable data protection laws.

[24] In the United States, Mathematica works with the Health Media Lab IRB. In Jordan, Mindset works with the IRB at the King Hussein Cancer Center.

V. Limitations, risks, and mitigation

The evaluation of the refugees DIB program faces some important limitations and risks that we will work to mitigate to the extent possible:

1. **Lack of baseline data for matching.** Ideally, the matching approach used for the impact evaluation would incorporate baseline information on consumption—the main outcome of interest—and other outcomes for the first and third cohorts, measured just before the first cohort was admitted to the program (that is, in early 2022). This would enable us to confidently assess the baseline equivalence of the treatment and comparison groups, and statistically adjust for any baseline differences when estimating impacts. However, these baseline data are not available because NEF and its partner CBOs will only identify the third cohort in early 2024. As described earlier, we will attempt to mitigate this by working with NEF to retrospectively capture information on early 2022 characteristics likely to be correlated with consumption for the third cohort (such as asset ownership) and using these along with time-invariant characteristics as matching variables. Although this is not as ideal as having a full baseline for both cohorts, achieving balance on these variables will help increase confidence in the validity of the impact estimates.

2. **Possibility of a pre-program dip in outcomes for the comparison group.** Although participants in the first and third cohorts will all be vulnerable by virtue of having been selected for the program, the timing of their vulnerability might differ. Specifically, some third cohort participants might have experienced a negative shock shortly before entering the program in early 2024, which made them more likely to apply for it at that specific time. We cannot be sure that first cohort participants would have experienced an equivalent negative shock at the same time, absent the program. This pre-program dip in outcomes for the third cohort might lead to us overestimating the impacts of the program (Ashenfelter 1978). To address this concern to the extent possible, we will use the semi-structured interviews with program implementers and FGDs with program participants to better understand the motivation for and timing of participant selection.

3. **Exposure of the comparison group to the program.** For the impact evaluation, the final implementation schedule implies that the impact evaluation survey would be conducted midway through the implementation period for the third cohort. This increases the risk that the comparison group would start to be affected by the program, which could dilute the impact estimates. We therefore propose slightly reducing the follow-up period for the impact evaluation from 24 to 23 months after grant disbursement for the first cohort, which would enable us to complete the survey earlier into the implementation period for the third cohort.[25] Impacts on consumption at 24 months versus 23 months are unlikely to be substantively different, and this change will help maintain the integrity of the comparison group.

4. **Potential for unit and item non-response.** Two types of non-response might affect the evaluation findings, including the estimated payment metrics: (1) unit non-response, or attrition, when some members of the sample cannot be contacted or refuse to respond to the survey (relevant to IGA validation and the impact survey); and (2) item non-response, when survey respondents do not recall or refuse to provide answers to specific questions (relevant mostly to the impact survey). Both types

[25] Participants are identified through the vulnerability assessment during the inception period and enter the program on a rolling basis once the implementation period begins. Because the data collection effort will take several weeks, we will explore whether it is feasible to prioritize first surveying respondents who enter (or are expected to enter) the program earlier, so that we can survey them as close as possible to the start of their program.

of non-response can introduce bias if those who do not respond are systematically different from those who do. For example, refugee grant recipients whose businesses have failed might be more likely to migrate and hence not response to the IGA validation survey, which could bias upwards the business metric. (This example might be an especially salient risk if there is large-scale repatriation or resettlement of Syrian refugees in the coming years.) Further, both types of non-response reduce statistical power, because the analysis sample size—overall or for specific outcomes—is reduced. We will use several strategies to maximize response rates in order reduce the potential for bias and maintain statistical power:

- Our local evaluation partner Mindset will make **multiple attempts to contact initial non-respondents** using their primary contact information, including on different days and at different times of day. For those who still cannot be contacted, Mindset will use detailed secondary contact information (such as the phone number of a relative or friend) to attempt to locate them; we will work with NEF to gather this information.
- To avoid a loss of statistical power, we will achieve our originally targeted sample sizes for both IGA validation and the impact evaluation by **replacing non-respondents with new sample members with similar characteristics**, as described in Chapter IV. This will help maintain the representativeness of the sample in terms of observable characteristics, although it does not rule out the potential for bias if respondents are systematically different from non-respondents in unobserved characteristics related to outcomes.
- To account for item non-response in the consumption metric—which might be expected given the long list of items that we will include—we will **impute missing values** using the approach described in **Appendix B**. This will avoid having to drop from the analysis respondents who are unable to recall consumption for a small number of items out of the full list.
- Mindset will strive to **use female interviewers to interview female respondents** to help build rapport with respondents to mitigate survey non-response and item non-response for potentially sensitive questions (for example, around women's empowerment).

5. **Challenges collecting data during Ramadan.** For the IGA validation for the first and second cohorts, the final implementation schedule implies a data collection timeline that falls during the Ramadan holiday in 2023 and 2024, respectively. Based on their extensive experience in Jordan, Mindset has recommended avoiding data collection during Ramadan because of the increased risk of poor data quality due to respondent and interviewer fatigue, as well as low response rates. Our proposed timeline, which we discuss in further detail in Chapter VI, accounts for this by adjusting the timing of IGA validation for the first and second cohorts to avoid collecting data during Ramadan. These adjustments have two main implications. First, the validation for the first cohort will occur 11 rather than 10 months after the expected midpoint of grant disbursement, and validation for the second cohort will occur 9 rather than 10 months after disbursement. However, on average across all three cohorts, the validation will occur 10 months after disbursement, as originally envisaged. Second, for the first cohort, our proposed 60-day reference period for measuring business activity will cover Ramadan. Although it is possible that some business activities will be affected by Ramadan, depending on the sector and geography, NEF has suggested that they do not believe that there will be a substantial net effect on the business metric.

6. **Receipt of other services.** In the 23 months between the midpoint of grant disbursement and the impact evaluation survey, first cohort participants might receive services from CBOs (through the

Siraj centers), other non-governmental organizations, refugee agencies, or the government, potentially contributing to increases in consumption levels. In contrast, participants in the third cohort (the comparison group) would generally not have received these types of services over the same period. (If they had, they would be less likely to qualify as vulnerable and be included in the third cohort in the first place.) Therefore, the estimated impacts on consumption could reflect both the effects of the program on the first cohort *and* the effects of other services received by them between early 2022 and mid-2024. We do not plan to attempt to quantitatively disentangle the impacts of these other services. The study is not designed or powered to do so, and the program logic suggests that uptake of other services might be in part attributable to the project (for example, by increasing the self-confidence of participants). Nevertheless, we plan to conduct a descriptive analysis of the types of other services received drawing on data collected through the impact evaluation survey, and we will explore the complementarity of these services with the program as part of the process evaluation.

7. **Challenges collecting data on business activity from vulnerable populations.** The vulnerable populations targeted by the refugee livelihoods DIB program might be wary of strangers asking for detailed information about their business activity. For example, they might be reluctant to provide accurate information about their business activity in case it is used by the tax authorities, or for fear of jeopardizing their refugee status, social benefits, or eligibility for future surveys. To address this, we will work with NEF and its partner CBOs to sensitize program participants to the evaluation data collection effort ahead of time. We will also ensure that we frame the data collection to respondents not as an audit, but rather as research to help understand their businesses and inform program improvements for future recipients. The informed consent statement will emphasize that the data will be used for research purposes only and assure respondents of their confidentiality.

VI. Evaluation team, timeline, and reporting

In this final chapter, we describe the evaluation team, present the workplan and reporting schedule for the evaluation, and discuss our plans for disseminating the evaluation findings.

1. Evaluation team roles and responsibilities

Mathematica has assembled a skilled and experienced evaluation team, with roles and responsibilities allocated across the team to maximize efficiency, leverage individuals' skills and experience, and ensure a high-quality evaluation. **Dr. Evan Borkum** is the project director and principal investigator. He will provide technical and managerial leadership and oversight of the entire evaluation and will be the primary evaluation point of contact for the IKEA Foundation, NEF, and other DIB partners. **Dr. Paolo Abarcar** is the impact evaluation lead. He will lead the development of the impact evaluation survey instrument, the impact evaluation analysis, and the reporting of the impact evaluation findings. **Ms. Laura Meyer** serves as both the IGA validation lead and data collection lead. As IGA validation lead, she will lead the survey instrument design, analysis, and reporting for that component of the evaluation. As data collection lead, she will provide close coordination with and oversight of Mathematica's local partners. She will also lead the development of the qualitative protocols for the process evaluation. **Mr. Matt Spitzer** is the team's statistical programmer and research analyst, and will support data collection, quantitative data analysis, and reporting. Mathematica also has two Jordanian research partners, **Integrated** and **Mindset**. Integrated will lead the qualitative data collection and analysis for the process evaluation, support Mindset in conducting high-quality quantitative data collection, and provide on-the-ground coordination between Mathematica and the program implementers as needed. Mindset will lead the quantitative data collection for IGA validation and the impact evaluation.

2. Workplan and reporting schedule

Figure VI.1 presents the proposed monthly workplan and reporting schedule for the evaluation, which is closely linked to the implementation timeline for the three cohorts. The workplan covers data collection for the three components of the evaluation (IGA validation, the impact evaluation, and the process evaluation), reporting of intermediate findings related to the payment metrics, and reporting of the final evaluation findings.

a. IGA validation

To validate the business metric, we will prepare and pretest the data collection instrument and protocols by late February 2023. We will use the approved instrument and protocols to collect data on business activity and formal employment for each of the three cohorts. As described in Chapter IV, data collection for each cohort would ideally occur in a window centered about 10 months after the end of the fifth month of implementation, which is expected to be the midpoint of grant disbursement. However, for the first and second cohorts this implies a data collection window centered around the end of March, which would overlap with the Ramadan holiday in 2023 and 2024, respectively.[26] As mentioned earlier, collecting these data during Ramadan is problematic because of challenges with data quality. For the first cohort, we therefore recommend starting the IGA data collection after the end of Ramadan (two or three weeks later, to provide some time for training and mobilization of the data collection team.) This implies that the data collection window will be centered 11 rather than 10 months after the end of the fifth month

[26] Ramadan falls between March 22 and April 21, 2023, and between March 10 and April 9, 2024.

of implementation for the first cohort. For the second cohort, we propose conducting the IGA data collection for these cohorts in a short four-week window before Ramadan. This implies that the data collection window will be centered 9 rather than 10 months after the end of the fifth month of implementation for the second cohorts.[27] (The data collection window would still be centered 10 months after the fifth month of implementation for the third cohort.) We will summarize the findings on the business metric for the first two cohorts in two separate memos that we will prepare soon after the data are collected for each cohort (**Appendix C** provides a template for these memos); we will incorporate the findings for the third cohort into the revised final evaluation report, which we discuss below.

b. Impact evaluation

To conduct the impact evaluation, which will include assessing impacts on the household consumption metric, we will prepare and pretest the data collection instrument and protocols by late November 2023. As soon as recruitment is completed for the third cohort (immediately before the start of the program) we will conduct statistical matching and sampling to identify the sample for data collection. We will then collect the survey data from sampled individuals in the treatment group (first cohort) and comparison group (third cohort). As described in Chapter IV, the window of data collection for the impact evaluation would ideally be centered 24 months after the end of the fifth month of implementation for the first cohort, which is at the end of June 2024. However, because of the risk of program exposure for the comparison group, we are proposing centering it 23 months after the end of the fifth month of implementation for the first cohort, which is at the end of May 2024. Data collection would therefore occur in May and June 2024. (Because Ramadan 2024 ends at the beginning of April, we are not concerned that our retrospective 7- or 30-day consumption measures will be affected by unusual consumption patterns during Ramadan.) We will summarize the findings on the consumption metric in a memo that we will prepare by late August 2024 (**Appendix D** provides a template for this memo).

c. Process evaluation

As described in Chapter IV, the process evaluation will draw on qualitative data from FGDs with program participant and semi-structured interviews program implementers, conducted soon after the IGA validation for the first two cohorts. It will also draw on NEF's program monitoring data and more comprehensive information on grantees' business operations from our IGA validation surveys. We do not plan to separately report on the process evaluation findings before the final report (described below), although we will communicate preliminary findings that might inform program improvement to NEF as they become available.

d. Final reporting

We will produce a preliminary final evaluation report by mid-December 2024, which will include the findings on the business metric for the first two cohorts (drawing on the respective memos), the impact evaluation (drawing from the consumption metric memo and analysis of other outcomes), and the process evaluation. We will update the preliminary final report with findings on the business metric for the third cohort as soon as they are available. We will produce a revised final report by the mid-August 2025, followed by an evaluation brief and a set of PowerPoint slides on the findings by the end of the month.

[27] We prefer not to delay IGA data collection for the second cohort until after Ramadan because the data collection and reporting would then coincide with that for the impact evaluation, requiring Mathematica and Mindset to divide resources across these two efforts.

Figure VI.1. Proposed workplan and reporting schedule

Program month	1 Oct 2021	2 Nov 2021	3 Dec 2021	4 Jan 2022	5 Feb 2022	6 Mar 2022	7 Apr 2022	8 May 2022	9 Jun 2022	10 Jul 2022	11 Aug 2022	12 Sep 2022
Program implementation (C1)									*			
IGA instrument/protocols												

Program month	13 Oct 2022	14 Nov 2022	15 Dec 2022	16 Jan 2023	17 Feb 2023	18 Mar 2023	19 Apr 2023	20 May 2023	21 Jun 2023	22 Jul 2023	23 Aug 2023	24 Sep 2023
Program implementation (C1)												
Program implementation (C2)									*			
IGA instrument/protocols					▲							
IGA validation (C1)												
IGA memo (C1)											▲	
Impact instrument/protocols												
FGDs and KIIs (C1)												

Program month	25 Oct 2023	26 Nov 2023	27 Dec 2023	28 Jan 2024	29 Feb 2024	30 Mar 2024	31 Apr 2024	32 May 2024	33 Jun 2024	34 Jul 2024	35 Aug 2024	36 Sep 2024
Program implementation (C2)												
Program implementation (C3)											*	
IGA validation (C2)												
IGA memo (C2)								▲				
Impact instrument/protocols		▲										
Impact survey												
Consumption memo											▲	
FGDs and KIIs (C2)												
Preliminary evaluation report												

Program month	37 Oct 2024	38 Nov 2024	39 Dec 2024	40 Jan 2025	41 Feb 2025	42 Mar 2025	43 Apr 2025	44 May 2025	45 Jun 2025	46 Jul 2025	47 Aug 2025	48 Sep 2025
Program implementation (C3)												
IGA validation (C3)												
Preliminary evaluation report			▲									
Revised evaluation report											▲	
Evaluation brief and slides											▲	

* = expected midpoint of grant disbursement; ▲ = draft deliverable

Note: Yellow shading indicates training and grants; blue shading indicates supplementary services; green-bordered boxes indicate "measurement periods" as defined in the DIB agreements (these periods begin at the start of each data collection effort and end with the draft deliverable that reports the respective findings).

C1/2/3 = cohorts 1/2/3

3. Dissemination plans

We plan to disseminate the findings through several channels to inform potential future investments in refugee livelihoods through DIBs or other mechanisms. First, we will post the revised final report permanently on Mathematica's website and publicize it through our social media channels. Second, we will prepare a short user-friendly and visually attractive policy brief summarizing the evaluation and its findings for a non-technical audience; we will post and publicize this brief online and encourage other DIB parties to do so too. Third, we will prepare a set of PowerPoint slides on the findings, which we will present to stakeholders in person in Jordan (including, potentially, NEF, government stakeholders, and other donors) and potentially to the broader development community through an online event. We will also make these slides available to DIB parties for broader dissemination. Finally, we will explore the possibility of developing publications based on our final report for peer-reviewed journals.

References

Al-Hawarin, Ibrahim, Ragui Assaad, and Ahmed Elsayed. "Migration Shocks and Housing: Evidence From the Syrian Refugee Crisis in Jordan." Economic Research Forum Working Paper Series, no. 1213, 2018.

Alkire, Sabina, Ruth Suseela Meinzen-Dick, Amber Peterman, Agnes R. Quisumbing, Greg Seymour, and Ana Vaz. "The Women's Empowerment in Agriculture Index." *World Development*, vol. 52, 2013, pp. 71–91. doi:10.1016/j.worlddev.2013.06.007.

Alzúa, Maria Laura, Soyolmaa Batbekh, Altantsetseg Batchuluun, Bayarmaa Dalkhjav, and José Galdo. "Demand-Driven Youth Training Programs: Experimental Evidence from Mongolia." *The World Bank Economic Review*, vol. 35, no. 3, 2021, pp. 720–744. https://doi.org/10.1093/wber/lhaa013.

Anderson, Stephen J., Rajesh Chandy, and Bilal Zia. "Pathways to profits: The impact of marketing vs. finance skills on business performance." *Management Science*, vol. 64, no. 12, 2018, pp. 5559-5583.

Anderson, S., P. Chintagunta and N. Vilcassim. "Do Marketers Matter for Entrepreneurs? Evidence from a Field Experiment in Uganda." *Journal of Marketing*, vol. 85, no. 3, 2021, pp. 78–96.

Anderson, S., L. Iacovone, S. Kankanhalli and S. Narayanan. "Modernizing Retailers in Emerging Markets: Investigating Externally-Focused and Internally-Focused Approaches." Stanford University Graduate School of Business Research Papers, no. 3920, 2020.

Anderson, S. and D. McKenzie. "Improving Business Practices and the Boundary of the Entrepreneur: A Randomized Experiment Comparing Training, Consulting, Insourcing and Outsourcing." World Bank Policy Research Working Paper, no. 9502, 2020.

Ashenfelter, Orley. "Estimating the Effect of Training Programs on Earnings." *The Review of Economics and Statistics,* vol. 60, no. 1, 1978, pp. 47-57. https://doi.org/10.2307/1924332.

Aziz, Sahar F. and Gardner, Joanna and Anaie, Tamara and Rana, Omar. "Towards Sustainability and Empowerment: Reforming America's Syrian Refugee Policy." Newark, New Jersey: Rutgers Law School Research Paper, January 29, 2019. http://dx.doi.org/10.2139/ssrn.3324792.

Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman. "Six Randomized Evaluations of Microcredit: Introduction and Further Steps." *American Economic Journal: Applied Economics*, vol. 7, no. 1, 2015 1-21. https://doi.org/10.1257/app.20140287.

Banerjee, Abhijit and Breza, Emily and Duflo, Esther and Kinnan, Cynthia. "Can Microfinance Unlock a Poverty Trap for Some Entrepreneurs?" NBER Working Paper, no. 26346, 2019. https://doi.org/10.3386/w26346.

Baranik, Lisa E., Adam W. Meade, Chad E. Lakey, Charles E. Lance, Changya Hu, Wei Hua, and Alex Michalos. "Examining the differential item functioning of the Rosenberg self-esteem scale Across eight countries 1." *Journal of Applied Social Psychology*, vol. 38, no. 7, 2008, pp. 1867-1904.

Bedoya, Guadalupe, Aidan Coville, Johannes Haushofer, Mohammad Isaqzadeh, and Jeremy Shapiro. "No Household Left Behind: Afghanistan Targeting the Ultra Poor Impact Evaluation." World Bank Policy Research Working Paper, no. 8877, 2019.

Beegle, Kathleen, Joachim De Weerdt, Jed Friedman, and John Gibson. "Methods of household consumption measurement through surveys: Experimental results from Tanzania." *Journal of Development Economics*, vol. 98, no. 1, 2012, pp. 3-18.

Berge, Lars Ivar Oppedal, Kjetil Bjorvatn, and Bertil Tungodden. "Human and financial capital for microenterprise development: Evidence from a field and lab experiment." *Management Science*, vol. 61, no. 2015, pp. 707-722.

Blattman, Christopher, Nathan Fiala, and Sebastian Martinez. "Generating Skilled Self-Employment in Developing Countries: Experimental Evidence from Uganda." *The Quarterly Journal of Economics*, vol. 129, no. 2, 2014, pp. 697–752.

Blattman, Christopher, Eric P. Green, Julian Jamison, M. Christian Lehmann, and Jeannie Annan. "The Returns to Microenterprise Support among the Ultrapoor: A Field Experiment in Postwar Uganda." *American Economic Journal: Applied Economics*, vol. 8, no. 2, 2016, pp. 35-64. https://doi.org/10.1257/app.20150023.

Blattman, Christopher, Nathan Fiala, and Sebastian Martinez. "The Long-Term Impacts of Grants on Poverty: Nine-Year Evidence from Uganda's Youth Opportunities Program." *American Economic Review: Insights*, vol. 2, no. 3, 2020, pp. 287-304. https://doi.org/10.1257/aeri.20190224.

Blattman, Christopher and Laura Ralston. "Generating Employment in Poor and Fragile States: Evidence from Labor Market and Entrepreneurship Programs." Social Science Research Network Working Paper, July 2015. Available at http://dx.doi.org/10.2139/ssrn.2622220.

Borkum, Evan, Arif Mamun, and Malik Khan Mubeen. "Evaluation of the Vocational Training Grant Fund in Namibia: Final Report." Washington, DC: Mathematica Policy Research, March 2017.

Brookings. "Social and development impact bonds by the numbers." December 2021. Available at: https://www.brookings.edu/research/social-and-development-impact-bonds-by-the-numbers/ Accessed December 6, 2021.

Brooks, Wyatt, Kevin Donovan, and Terence R. Johnson. "Mentors or Teachers? Microenterprise Training in Kenya." *American Economic Journal: Applied Economics*, vol. 10, no. 4, 2018, pp. 196-221. https://doi.org/10.1257/app.20170042.

Buvinić, Mayra and Rebecca Furst-Nichols. "Promoting Women's Economic Empowerment: What Works?" *The World Bank Research Observer*, vol. 31, no. 1, February 2016, pp. 59–101. https://doi.org/10.1093/wbro/lku013.

Calderon, Gabriela, Jesse M. Cunha, and Giacomo De Giorgi. "Business Literacy and Development: Evidence from a Randomized Controlled Trial in Rural Mexico." NBER Working Paper, no. 19740, 2013. https://doi.org/10.3386/w19740.

Campos, Francisco, Michael Frese, Markus Goldstein , Leonardo Iacovone , Hillary C. Johnson, David McKenzie and Mona Mensmann. "Teaching Personal Initiative Beats Traditional Training in Boosting Small Business in West Africa." *Science*, vol. 357, no. 6357, 2017, pp. 1287-1290. https://doi.org/10.1126/science.aan5329.

Ceylan, Vedat, Coşkun Algan, Mehmet Yalçin, Esra Yalçin, A. K. I. N. Ercan, and K. Ö. S. E. Samet. "An Investigation of Syrian Refugees and Asylum Seekers in Terms of Various Psychological Factors: Batman Sample." *International Journal of Social Sciences and Education Research*, vol. 3, no. 2, 2017, pp. 595-606.

Chakravarty, Shubha, Mattias Lundberg, Plamen Nikolov, Juliane Zenker. "Vocational Training Programs and Youth Labor Market Outcomes: Evidence from Nepal." *Journal of Development Economics*, vol. 136, 2019, pp. 71-110. https://doi.org/10.1016/j.jdeveco.2018.09.002.

Chinen M., T. de Hoop, L. Alcázar, M. Balarin, J. Sennett. "Vocational and Business Training to Improve Women's Labour Market Outcomes in Low- and Middle-Income Countries: A Systematic Review." *Campbell Systematic Reviews*, vol. 16, 2017. https://doi.org/10.4073/csr.2017.16.

Clemens, Michael, Cindy Huang, Jimmy Graham, and Kate Gough. "Migration Is What You Make It: Seven Policy Decisions That Turned Challenges into Opportunities." Washington, DC: Center for Global Development, 2018.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. "Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand." NBER Working Paper, no. 0330, 2006. https://doi.org/10.3386/t0330.

Culbertson, Shelly, Olga Oliker, Ben Baruch, and Ilana Blum. "Rethinking Coordination of Services to Refugees in Urban Areas: Managing the Crisis in Jordan and Lebanon." Santa Monica, California: RAND, 2016.

Dahal, Mahesh and Nathan Fiala. "What Do We Know About the Impact of Microfinance? The Problems of Statistical Power and Precision." *World Development*, vol. 128, 2020, pp. 104773. https://doi.org/10.1016/j.worlddev.2019.104773.

Deaton, Angus, and Zaidi, Salman. "Guidelines for Constructing Consumption Aggregates for Welfare Analysis." LSMS Working Paper. No. 135. Washington DC: World Bank, 2002.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. "Business Training and Female Enterprise Start-Up, Growth, and Dynamics: Experimental Evidence from Sri Lanka." *Journal of Development Economics*, vol. 106, 2014, pp. 199-210. https://doi.org/10.1016/j.jdeveco.2013.09.005.

Diener, E. D., Robert A. Emmons, Randy J. Larsen, and Sharon Griffin. "The Satisfaction with Life Scale." *Journal of Personality Assessment*, vol. 49, no. 1, 1985, pp. 71-75.

Donald, Aletheia, Gayatri Koolwal, Jeannie Annan, Kathryn Falb, and Markus Goldstein. "Measuring Women's Agency." World Bank Policy Research Working Paper, no. 8148, 2017.

Doocy S. and H. Tappis. "Cash-Based Approaches in Humanitarian Emergencies: A Systematic Review." *Campbell Systematic Reviews*, vol. 17, 2017, https://doi.org/10.4073/csr.2017.17.

Department of Statistics (DOS) and ICF. "Jordan: Population and Family Health Survey 2017-2018." Amman, Jordan and Rockville, Maryland, USA: DOS and ICF. 2019. Available at: https://dhsprogram.com/pubs/pdf/FR346/FR346.pdf.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael W. Walker. "General Equilibrium Effects of Cash Transfers: Experimental Evidence from Kenya." NBER Working Paper, no. 26600, 2019.

Fallah, Belal, Caroline Krafft, and Jackline Wahba. 2018. "The Impact of Refugees on Employment and Wages in Jordan." Economic Research Forum Working Paper Series, no. 1189, 2019.

Fiala, Nathan. "Access to Finance and Enterprise Growth: Evidence from an Experiment in Uganda." Geneva, Switzerland: ILO, 2015. Available at: http://ilo.ch/wcmsp5/groups/public/--ed_emp/documents/publication/wcms_432287.pdf.

Filmer, Deon, and Kinnon Scott. "Assessing asset indices." *Demography*, vol. 49, no. 1, 2012, pp. 359-392.

Giné, Xavier and Ghazala Mansuri. "Money or Ideas? A Field Experiment on Constraints to Entrepreneurship in Rural Pakistan." World Bank Policy Research Working Paper, no. 6959, 2014.

Gonzalez-Uribe, Juanita and Michael Leatherbee. "The Effects of Business Accelerators on Venture Performance: Evidence from Start-Up Chile." *The Review of Financial Studies*, vol. 31, no. 4, April 2018, pp.1566–1603. https://doi.org/10.1093/rfs/hhx103.

Harder, Valerie S., Elizabeth A. Stuart, and James C. Anthony. "Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research." *Psychological Methods*, vol. 15, no. 3, 2010, pp. 234-249.

Hassan, Billow, Stephen Mutiso, and Munshi Sulaiman. "More Cash, Bigger Impacts? A Field Experiment of Cash Transfers in Somalia." London, UK: Save the Children, 2018. Available at: https://somalia.savethechildren.net/sites/somalia.savethechildren.net/files/library/UCT_IGA_Somalia.pdf.

Haushofer, Johannes and Jeremy Shapiro. "The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya." *The Quarterly Journal of Economics*, vol. 131, no. 4, 2016, pp. 1973–2042. https://doi.org/10.1093/qje/qjw025.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis*, vol. 15, no. 3, 2007, pp. 199-236.

Huis, Marloes, Robert Lensink, Nhung Vu, and Nina Hansen. "Impacts of the Gender and Entrepreneurship Together Ahead (GET Ahead) Training on Empowerment of Female Microfinance Borrowers in Northern Vietnam." *World Development*, vol. 120, 2019, pp. 46-61.

Hussam, Reshmaan N., Erin M. Kelley, Gregory V. Lane, and Fatima T. Zahra. "The Psychosocial Value of Employment." NBER Working Paper, no. 28924, 2021.

Iacus, Stefano M., Gary King, and Giuseppe Porro. "Causal inference without balance checking: Coarsened exact matching." *Political Analysis*, vol. 20, no. 1, 2012, pp. 1-24.

Iacus, Stefano M., Gary King, and Giuseppe Porro. "A theory of statistical inference for matching methods in causal research." *Political Analysis*, vol. 27, no. 1, 2019, pp. 46-68.

International Labour Organization (ILO). "Annual Report 2021: An Industry and Compliance Review Better Work Jordan." Geneva, Switzerland: ILO, 2021. Available at: https://betterwork.org/wp-debtcontent/uploads/2021/04/BWJ_12-annual-report_EN_.pdf.

Imai, Kosuke, and Marc Ratkovic. "Covariate balancing propensity score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. vol. 76, no. 1, 2014, pp. 243-263.

Imbens, Guido W., and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Keaveney, Erika, Mike Duthie, and Andrea Hur. "Development Impact Bond (DIB) for Improving Livelihood for Syrian Refugees and Vulnerable Populations in Jordan and Lebanon: Evaluation Design Report." Arlington, VA: Social Impact, 2018.

Kebede, Tewodros Aragie, Svein Erik Stave and Maha Kattaa. "Facing Double Crises: Rapid Assessment of the Impact of Covid-19 on Vulnerable Workers in Jordan."
Geneva, Switzerland: International Labour Organization, May 2020. Available at: https://www.ilo.org/wcmsp5/groups/public/---arabstates/---ro-beirut/documents/publication/wcms_743391.pdf.

Kebede, Tewodros Aragie, Svein Erik Stave, Åge A. Tiltnes, Maha Kattaa, Michaela Prokop, Ramzi Maaytah, and Silke Handley. "Impact of Covid-19 on Enterprises in Jordan: One Year into the Pandemic." ReliefWeb, 2021. Available at: https://reliefweb.int/sites/reliefweb.int/files/resources/ILO%20FAFO%20UNDP%20Covid%20Report%20July%202021.pdf.

King, Gary, and Richard Nielsen. "Why propensity scores should not be used for matching." *Political Analysis*, vol. 27, no. 4, 2019, pp. 435-454.

Kluve J., S. Puerto, D. Robalino, J. M. Romero, F. Rother, J. Stöterau, F. Weidenkaff, M. Witte. "Interventions to Improve the Labour Market Outcomes of Youth: A Systematic Review of Training, Entrepreneurship Promotion, Employment Services, and Subsidized Employment Interventions." *Campbell Systematic Reviews*, vol. 12, 2017. https://doi.org/10.4073/csr.2017.12.

Krafft, Caroline, Susan Razzaz, Caitlyn Keo, and Ragui Assaad. "The Number and Characteristics of Syrians in Jordan: A Multi-Source Analysis." Economic Research Forum Working Paper Series, no. 1288, 2019.

Legrain, Philippe, and Andrew Burridge. "Seven Steps to Success: Enabling Refugee Entrepreneurs to Flourish." Sydney, Australia: Cities and Settlement Initiative, 2019.

Lehmann, Christian and Daniel T. R. Masterson. "Impact Evaluation of a Cash-Transfer Programme for Syrian Refugees in Lebanon." *Field Exchange,* no. 48, November 2014. https://www.ennonline.net/fex/48/impactevaluation.

Lenner, Katharina and Lewis Turner. "Making Refugees Work? The Politics of Integrating Syrian Refugees into the Labor Market in Jordan." *Middle East Critique*, vol. 28, no. 1, 2019, pp. 65–95. doi:10.1080/19436149.2018.1462601.

Maitra, Pushkar, and Subha Mani. "Learning and Earning: Evidence from a Randomized Evaluation in India." *Labour Economics*, vol. 45, 2017, pp.116-130. https://doi.org/10.1016/j.labeco.2016.11.007.

Malaeb, Bilal, and Jackline Wahba. "Impact of Refugees on Immigrants' Labor Market Outcomes." Economic Research Forum Working Paper Series, no. 1194, 2018. Available at: https://erf.org.eg/app/uploads/2018/05/1194_Final1.pdf.

Maxwell, Daniel and Richard Caldwell. "The Coping Strategies Index Field Methods Manual, Second Edition." Rome, Italy: World Food Programme, 2008.

McKenzie, David. "How Effective Are Active Labor Market Policies in Developing Countries? A Critical Review of Recent Evidence." World Bank Policy Research Working Paper, no. 8011, 2017.

McKenzie, David, and Susana Puerto. "Growing Markets through Business Training for Female Entrepreneurs: A Market-Level Randomized Experiment in Kenya." *American Economic Journal: Applied Economics*, vol. 13, no. 2, 2021, pp. 297-332. https://doi.org/10.1257/app.20180340.

McKenzie, David and Christopher Woodruff. "Business Practices in Small Firms in Developing Countries." *Management Science*, vol. 63, no. 9, 2016, pp. 2967-2981. https://doi.org/10.1287/mnsc.2016.2492.

McKenzie, David, Christopher Woodruff, Kjetil Bjorvatn, Miriam Bruhn, Jing Cai, Juanita Gonzalez-Uribe, Simon Quinn, Tetsushi Sonobe, and Martin Valdivia. "Training Entrepreneurs." VoxDevLit, vol. 1 no. 2, August 2021. Available at: https://voxdev.org/lits/training-entrepreneurs.

Meager, Rachael. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics*, vol. 11, 2019, pp. 57–91.

Miller, K. E., Koppenol-Gonzalez, G. V., Jawad, A., Steen, F., Sassine, M., & Jordans, M. J. "A Randomised Controlled Trial of the I-Deal Life Skills Intervention with Syrian Refugee Adolescents in Northern Lebanon." *Intervention*, vol, *18*, no. 2, 2020, p. 119. Available at: https://www.interventionjournal.org/temp/Intervention182119-6242422_172024.pdf.

Morduch, Jonathan. "Why RCTs failed to answer the biggest questions about microcredit impact." *World Development*, vol. 127, 2020, pp. 104818.

Near East Foundation (NEF). "Vulnerability Assessment." 2020.

Oxford Government Outcomes Lab. "Impact Bond Dataset." December 2021. Available at: https://golab.bsg.ox.ac.uk/knowledge-bank/indigo/impact-bond-dataset-v2/ Accessed December 6, 2021.

Palinkas, Lawrence A., Sarah M. Horwitz, Carla A. Green, Jennifer P. Wisdom, Naihua Duan, and Kimberly Hoagwood. "Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research." *Administration and Policy in Mental Health and Mental Health Services Research*, vol. 42, no. 5, 2015, pp. 533-544.

REACH 2014. "Livelihoods, Employment and Tensions in Jordanian Communities Hosting Syrian Refugees: Thematic Assessment Report." Geneva, Switzerland: REACH, 2014. Available at: https://www.impact-repository.org/document/reach/f9d117b8/reach_jor_report_employmentandtensionsinjordaniancommunitieshostingsyrianrefugees.pdf.

ReliefWeb. "The Jordan Response Plan for the Syria Crisis 2016-2019." 9 August 2017. Available at: https://reliefweb.int/report/jordan/jordan-response-plan-syria-crisis-2016-2019.

Rosenberg, Morris. "Rosenberg Self-Esteem Scale (RSE)." *Acceptance and Commitment Therapy. Measures Package*, vol. 61, no. 52, 1965, pp. 18.

Rubin, Donald B. "Using propensity scores to help design observational studies: application to the tobacco litigation." *Health Services and Outcomes Research Methodology*. vol. 2, no. 3, 2001, pp. 169-188.

Schmitt, D. P., & Allik, J. "Simultaneous Administration of the Rosenberg Self-Esteem Scale in 53 Nations: Exploring the Universal and Culture-Specific Features of Global Self-Esteem." *Journal of Personality and Social Psychology,* vol. 89, no. 4, 2005, pp. 623–642. https://doi.org/10.1037/0022-3514.89.4.623.

Stuart, Elizabeth A. "Matching methods for causal inference: A review and a look forward." *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, vol. 25, no. 1, 2010, pp. 1-21.

Tiltnes, Åge A., Huafeng Zhang and Jon Pedersen. "The Living Conditions of Syrian Refugees in Jordan: Results from the 2017-2018 Survey of Syrian Refugees Inside and Outside Camps." ReliefWeb. 2019. Available at: https://reliefweb.int/sites/reliefweb.int/files/resources/67914.pdf.

Ubfal, Diego, Diether Beuermann, Michael Frese, Alessandro Maffioli, Irani Arráiz, and Daniel Verch. "The Impact of Soft-Skills Training for Entrepreneurs in Jamaica." IDB Working Paper no. 1011, June 2020.

United Nations Children's Fund (UNICEF). "Geographic Multidimensional Vulnerability Analysis - Jordan." New York, New York: UNICEF, February 2020. Available at: https://reliefweb.int/sites/reliefweb.int/files/resources/Summary%20English.pdf.

United Nations (UN). "Unpacking Gendered Realities in Displacement: The Status of Syrian Refugee Women in Jordan." Cairo, Egypt: United Nations Women Regional Office for Arab States, 2018. https://arabstates.unwomen.org/en/digital-library/publications/2018/12/unpacking-gendered-realities-in-displacement.

United Nations High Commissioner for Refugees (UNHCR). "Vulnerability Assessment Framework (VAF) population report." Geneva, Switzerland: United Nations High Commissioner for Refugees, April 2019. Available at: https://data2.unhcr.org/en/documents/download/68856.

Wall, Melissa, Madeline Otis Campbell, and Dana Janbek. "Refugees, Information Precarity, and Social Inclusion: The Precarious Communication Practices of Syrians Fleeing War." *The Handbook of Diasporas, Media, and Culture*, 2019, pp. 503-514. https://doi.org/10.1002/9781119236771.ch34.

World Bank. "Purchasing Power Parities and the Size of World Economies: Results from the 2017 International Comparison Program." Washington, DC: World Bank, 2020.

World Bank. "The World Bank In Jordan." October, 2021. Available at: https://www.worldbank.org/en/country/jordan/overview#1. Accessed November 4, 2021.

World Bank. "Jordan - Survey of Syrian Refugees and Host Communities, 2015-2016 (SRHCS 2015-16) | Data Catalog." n.d. Available at https://datacatalog.worldbank.org/dataset/jordan-survey-syrian-refugees-and-host-communities-2015-2016. Accessed May 15, 2020.

Yount, Kathryn M., Kristin E. VanderEnde, Sylvie Dodell, and Yuk Fai Cheong. "Measurement of Women's Agency in Egypt: A National Validation Study." *Social Indicators Research*, vol. 128, no. 3, 2016, pp. 1171–1192. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5010232/

Appendix A: Statistical matching for the impact evaluation

In this appendix we describe the technical details of the statistical matching between the first cohort (treatment group) and third cohort (comparison group) that we will conduct as part of the quasi-experimental matched comparison group design. Below, we describe the matching procedures and matching characteristics we will consider, how we will assess the quality of the matches we obtain, and how we will conduct the sampling.

1. Matching procedure

There are several methods for conducting matching, with tradeoffs between methods. Based on the literature and our experience with other evaluations, we propose to use either Coarsened Exact Matching (CEM) or Inverse Propensity Score Weighting (IPSW):

- **Coarsened Exact Matching (CEM).** Under CEM (Iacus et al. 2012), we would divide members of the potential treatment and comparison groups into mutually exclusive strata, with each stratum defined by a combination of participant characteristics (matching variables). We would then reweight the comparison group so that its distribution across strata is identical to that of the treatment group. Thus, all comparison group observations in the same stratum would receive the same weight. In this way, the comparison group is adjusted so that it is as similar as possible to the treatment group in terms of the matching characteristics, enabling us to produce credible impact estimates. For continuous or multi-valued matching variables (for example, age), using the variable directly to form strata would lead to many small strata that have only treatment or control observation(s). These strata would not contribute to the impact estimates and would have to be dropped, reducing the representativeness of the sample and hence the generalizability of the results. Therefore, under CEM we would temporarily coarsen continuous or multi-valued matching variables into broader categories (for example, age groups instead of age in years). Combinations of these coarsened matching variables, together with other categorical matching variables (for example, location) and binary matching variables (for example, refugee status), would define the strata for the analysis.[28]
- **Inverse Propensity Score Weighting (IPSW).** Under this method, we would estimate the likelihood of being in the treatment group (the first cohort), based on a set of pre-program participant characteristics likely to be correlated with treatment (that is, correlated with being vulnerable and referred to the program in early 2022, when the first cohort is selected) and with outcomes (especially household consumption, in mid-2024). We would then use this estimated likelihood, known as the propensity score, to assign weights to each participant in the treatment and comparison cohorts for the impact analysis. Specifically, treatment observations with a low propensity score—that is, a lower predicted likelihood of being in the treatment group—would be given more weight than treatment observations with a high propensity score, while the opposite would apply for units in the comparison group.[29] To estimate the propensity score we will use the covariate-balancing propensity score approach (Imai and Ratkovic 2014) that models treatment assignment while also optimizing balance

[28] Some matching variables may have missing values because participants did not provide the information for the vulnerability assessment. When defining strata, we will consider missing values as a separate category.

[29] A common alternative approach using propensity scores is matching, which seeks to match treatment and comparison observations based on the similarity in their propensity scores and estimates impacts for the matched sample. The main advantage of weighting over matching is that the former typically uses the full sample (or close to the full sample), whereas matching methods drop unmatched observations; this enables us to estimate the average treatment effect for the fully representative treatment sample (Stuart 2010).

in participant characteristics; this approach reduces the risk of bias due to misspecification of the propensity score relative to the typical approach of using a simple logistic regression.[30]

Our final choice of method will be data dependent and will be driven by: (1) balance on observable characteristics between treatment and comparison groups, and (2) representativeness of the sample. CEM has been shown to possess attractive statistical properties and produce less bias compared to other methods (Iacus et al. 2012; Iacus et al. 2019; King and Nielsen 2019). CEM is also an intuitive approach as it is equivalent to comparing participants in the same strata, who by definition have the same observable characteristics; in contrast, propensity score methods rely on statistical modeling approaches and assumptions that are less transparent. However, CEM runs the risk of having several units unmatched, resulting in a less representative sample, which can threaten the generalizability of results. If there are many unmatched units, we would consider IPSW as an alternative. The advantage of IPSW is that we can retain the full sample while achieving acceptable balance between treatment and comparison groups. The disadvantage of IPSW is its potential sensitivity to a few observations with very large weights, which could lead to bias and imprecise impact estimates. We will address this concern by employing stabilization weights and top coding the weights as necessary (Harder et al. 2010).[31]

2. Matching characteristics

To conduct matching, we intend to draw on data collected through the vulnerability assessment that NEF and its CBO partners administer to screen potential participants. The vulnerability assessment scores candidates in each of six domains: (1) household characteristics (20 percent); (2) shelter (10 percent); (3) household finances (25 percent); (4) food consumption and coping strategies (25 percent); (5) access to a safety net (10 percent); and (6) decision-making and mentality (10 percent). From these scores, NEF calculates an overall vulnerability score between 0 and 100 for each candidate, with a higher score indicating greater vulnerability. Those who are deemed to be the most vulnerable based on this score and meet other eligibility criteria are admitted to the program.

The matching literature suggests that it is important to include in the matching procedure pre-program characteristics that are plausibly related to both treatment assignment and outcomes (Stuart 2010). Many characteristics included in the vulnerability assessment for the first cohort in early 2022 are good candidates for matching because they are likely to be associated with household consumption in early 2022. This in turn is likely to be correlated with household consumption in mid-2024—the key outcome of interest.

However, an important challenge is that many of the characteristics in the vulnerability assessment are time varying and will only be measured for third cohort participants in early 2024, shortly before the implementation period for the third cohort begins. To be valid for matching, these characteristics would have to be measured at the same point in time as for the first cohort, in early 2022. Otherwise, participants in the treatment and comparison groups could be dissimilar even if well-matched on these characteristics. For example, consider matched participants in the first and third cohorts who report the same income

[30] To avoid dropping any observations in our sample when estimating the propensity score, we will apply the common approach of replacing missing covariates with mean or median values and including missing data indicators for each of these covariates.

[31] Stabilization is achieved by multiplying the treatment or comparison weights by the expected value of being in the treatment or comparison group, respectively. This does not change the impact estimate but decreases its variance. Top-coding sets extreme weights equal to prespecified cutoff values (for example, it sets weights greater than 10 to 10). This procedure minimizes the influence of a few outlier weights.

levels in the early 2022 and early 2024 vulnerability assessments, respectively. If the first cohort experiences negative macro-economic effects of the COVID-19 pandemic and the third cohort does not, first cohort participants might be different to the matched participants in the third cohort despite the similarity in income (for example, they could be more resilient or with access to more stable sources of income, given that they achieved that level of income during a crisis). We would therefore expect the treatment group to have higher consumption levels than the comparison group in mid-2024, when the two groups are facing the same external conditions, even absent the program. This could lead us to overstate the impacts of the program.

Therefore, we propose to match on characteristics that are (1) time invariant, or (2) can be captured retrospectively for the third cohort when they are administered the vulnerability assessment in early 2024. We will work with NEF to incorporate these characteristics into the vulnerability assessment. The potential matching characteristics include the following (**Table A.1**):[32]

- **Socio-demographic characteristics**. These characteristics (for example, gender, year of birth, refugee status, and educational attainment) are largely time-invariant and are plausibly correlated with household consumption. However, because household size might change over time, we will capture the third cohort's early 2022 household size retrospectively.[33]
- **Housing characteristics**. The type of construction materials used in participants' housing, the number of rooms, and access to services (water, electricity, and sanitation) are standard characteristics used in estimating asset or wealth indices in Jordan (for example, Department of Statistics and ICF 2019 and Tiltnes et al. 2019).[34] These indices are typically strongly correlated with consumption and are often used as an alternative measure of consumption in many developing country settings (Filmer and Scott 2012). Although housing characteristics might change over time, we believe that the third cohort will be able to retrospectively provide information on early 2022 housing characteristics. (For the first cohort, some of these characteristics are already captured in the vulnerability assessment, but others will need to be added.) We will use these data on these characteristics, together with the ownership of durable goods described below, to construct an asset index using principal components analysis for use in matching (Filmer and Scott 2012).
- **Ownership of durable goods**. Although not included in the vulnerability assessment, ownership of durable goods is strongly correlated with consumption and is typically included in asset indices (Filmer and Scott 2012). We therefore propose to add a short list of durables to the vulnerability assessment for both the first and the third cohorts, asking the questions retrospectively about early 2022 for the third cohort. We selected durable goods that are commonly owned by Jordanian households and refugees in the country based on recent surveys of these populations (Department of Statistics and ICF 2019 and Tiltnes et al. 2019), but excluded goods owned by almost all households

[32] This list of matching variables is preliminary. We will determine the appropriate variables to include based on (1) variation in the sample, and (2) correlation with the rough measures of consumption and income included in the vulnerability assessment for the first cohort. The most useful matching variables will be those that are variable across the sample and are highly correlated with consumption and income.

[33] To assess the reliability of this and other retrospective measures described here, we will also measure them retrospectively for the first cohort in the impact evaluation survey (mid-2024) and compare to what the first cohort reported when they completed the vulnerability assessment (early 2022). Because of concerns about the length of the impact evaluation survey, we will test the reliability of each measure for a randomly selected subset of first cohort respondents to the impact evaluation survey.

[34] The precise list of characteristics in Department of Statistics and ICF (2019) that were used to calculate the wealth index for Jordan is available through the following link: https://dhsprogram.com/topics/wealth-index/Wealth-Index-Construction.cfm.

because these will not contribute useful information for matching. We will include these variables, together with the housing characteristics described above, to construct the asset index for matching.

Table A.1. Proposed characteristics for matching

Participant characteristic	Adjustments required to vulnerability assessment
Socio-demographic characteristics	
Gender	None
Year of birth/age	None; calculate age of respondent in early 2022
Refugee status	None
Location	Ask cohort 3 retrospectively about early 2022[a]
Educational attainment and/or literacy	None
Household size	Ask cohort 3 retrospectively about early 2022
Housing characteristics	
Type of housing (e.g., apartment, house, or tent)	Ask cohort 3 retrospectively about early 2022
Type of occupancy (e.g., rented, shared, or owned)	Ask cohort 3 retrospectively about early 2022
Main flooring material	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Main roof material	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Main wall material	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Number of rooms	Ask cohort 3 retrospectively about early 2022; use to estimate ratio of household size to rooms
Whether housing has a separate kitchen	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Access to electricity	Ask cohort 3 retrospectively about early 2022
Access to clean water	Ask cohort 3 retrospectively about early 2022
Type of toilet facility	Ask cohort 3 retrospectively about early 2022
Durable good ownership	
Car	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Computer	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Air conditioner	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Gas oven/cooker	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Microwave	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Electric/gas heater	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Water cooler/heater	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Freezer	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Electric fan	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Vacuum cleaner	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Electric iron	Add to cohort 1 and ask cohort 3 retrospectively about early 2022
Radio	Add to cohort 1 and ask cohort 3 retrospectively about early 2022

[a]We will match on the location of participants when they were selected for the program. We will also consider restricting third cohort refugee participants to those who were already in Jordan in early 2022, or even matching

based on when refugee participants arrived in Jordan. Third cohort participants who were not in Jordan in early 2022 might be especially poor comparisons because their trajectory of outcomes in subsequent years might not be a good counterfactual for those who were already in Jordan by that time.

3. Assessing balance and representativeness

To finalize the matching approach, we will assess the quality of the matched treatment and comparison groups in terms of balance and representativeness. We will quantify balance by calculating the standardized difference between treatment and comparison groups for each matching variable using the calculated matching weights. (Under CEM, balance will mechanically be perfect on all coarsened matching variables, but not necessarily on the uncoarsened versions of these variables.) As a rule-of-thumb, a standardized difference of less than 0.25 for matching variables is usually a good indication of balance (Harder et al. 2010; Rubin 2001; Imbens and Rubin 2015). To assess the extent to which the matched treatment sample (matched treatment units under CEM, or the weighted treatment sample under IPSW) is representative of the full treatment sample, we will compare the characteristics of the two samples using the same approach to assess treatment-comparison balance.

Because our sampling approach for the impact evaluation survey (described below) depends on first matching *all* participants in the first and third cohorts, we will have to conduct the matching rapidly in a short window between when the third cohort is identified and when we conduct the survey. To facilitate this, we will test potential matching approaches using program intake data for the first and second cohorts, which will help identify the most promising approach ahead of time for the first and third cohorts. (This approach will require NEF to implement the modifications to the vulnerability assessment proposed above for the second cohort too.) Finalizing the matching approach and matched sample for the first and third cohorts before we conduct the impact evaluation survey will also help avoid concerns that the matching approach was selected to achieve desirable impact estimates.

4. Sampling approach

The sampling approach will depend on the matching procedure used. Under CEM, our sample frame will comprise the full sample of treatment and comparison participants, except those in strata where all participants have the same treatment status. We would then randomly sample treatment and comparison participants in each stratum in proportion to the respective number in each stratum to achieve our target sample size. This will preserve the overall distribution of treatment and comparison participants across strata, ensuring that it is representative of the full matched sample without the need for sampling weights. In the case of survey non-response, we will replace the participant with another in the same stratum and treatment status unless there are none, in which case, we will randomly sample a participant of the same treatment status from other strata. Under IPSW, our initial sample frame will comprise the full sample of treatment and comparison participants; we will estimate the propensity score using this full sample frame. We will then draw a simple random sample of the targeted number of treatment and comparison participants. To improve the comparability between the two groups while largely maintaining the representativeness of the sample, we will consider first excluding participants with very high or very low propensity scores (Crump et al. 2006). In the case of survey non-response, we will randomly replace participants with another with the same treatment status that has a propensity score in the same decile as the non-respondent.

Appendix B: Technical details for calculating the consumption aggregate

The consumption aggregate is the monetary value of goods and services consumed by the household. To calculate it, we will follow the standard advice, procedures, and formulae in Deaton and Zaidi (2002). For the evaluation, we define the consumption aggregate for each household as the sum of the four main components of consumption plus debt repayments after adjusting for household size and price differences across locations. We will compute the aggregate using the following equation for household *h* residing in location *l*:

$$ConsAgg_h = \frac{Food_h + Nonfood_h + Durables_h + Housing_h + Debt_h}{HHsize_h}$$

Where:

$ConsAgg_h$ is the spatially adjusted estimated annual consumption aggregate per capita for the household;

$Food_h$ is the estimated annual food consumption of the household;

$Nonfood_h$ is the estimated annual non-food consumption of the household;

$Durables_h$ is the estimated annual durable goods consumption for the household;

$Housing_h$ is the estimated annual housing expenses (rent) for the household;

$Debt_h$ is the annual debt repayments of the household; and

$HHsize_i$ is the household size at the time of data collection.

1. Food consumption

We will calculate food consumption as the total value of food purchased from the market, food that is home-produced, and food that is received as a gift or in-kind payment. Our survey will first ask whether the household consumed each type of food item in the last seven-day period and if so, the total physical quantity consumed that was purchased, home-produced, or received as a gift or in-kind payment. We will also ask for the share that was purchased and the amount paid for the purchased share, which will enable us to estimate a per-unit price that we can use to value non-purchased food.

We will thus calculate the annual food consumption for a household *h* as the sum of reported values of food items purchased and imputed values of food not purchased (as described below) across all food items *i* consumed in the last seven days, which we will then annualize:

$$Food_h = \sum_{i=1}^{I} value\ of\ food\ item_{ih} * \frac{365}{7}$$

We will impute the total monetary value of non-purchased food by multiplying the physical amount that was home-produced or received as a gift or in-kind payment by its unit price. We will derive the unit price from the total monetary value and physical amount that was purchased if the household also purchased the item in the last seven days. If this is not available, then we would use the median unit price

paid by other households in the same location (across both the treatment and comparison groups) or from the whole study sample if there were no purchases of the item by households in the same location.[35]

2. Non-food consumption

We will calculate non-food consumption as the total value of expenditures on a wide range of non-food items and services. Following standard practice, our survey will only record the value of non-food items and services purchased and not the quantities consumed. Given that these expenses generally take place with different frequencies, we will ask households to recall their expenditure on non-food items using different recall periods: 30 days, 90 days, or 12 months, following the recall periods for each group of items captured in the Jordan HIES. We will calculate annual non-food consumption for a household *h* as the sum of reported non-food expenditures across all non-food items *i*, which we will then annualize:

$$Nonfood_h = \sum_{i=1}^{I} value\ of\ nonfood\ item_{ih} * annualization\ factor$$

We will apply an annualization factor of 365/30 for 30-day recall items, 365/90 for 90-day recall items, and 1 for 12-month recall items.

3. Durable goods consumption

We will calculate durable goods consumption as the value of services that the household receives from all durable goods in its possession over the year. For each durable good *i*, we will calculate the rental equivalent using the following formula (Deaton and Zaidi 2002):

$$rental\ equivalent\ of\ durable_i = V_{i,t}\ (\ r_t\ +\ \delta_i\ )$$

Where:

$V_{i,t}$ is the current value of durable good in the present time period *t*;

r_t is the real interest rate;[36] and

δ_i is the depreciation rate.

Intuitively, an individual would be willing to rent a good if they were compensated for the alternative option of selling the good and investing the money, as well as the depreciation of the good due to its use by the renter.

To compute the rental equivalent, our survey will gather information on (1) the year the durable good was purchased, $t - T$; (2) how much was paid for it, $V_{i,t-T}$; and (3) the household's estimate of its current value $V_{i,t}$. Using this information, we can calculate the depreciation rate for each item using the following formula:

[35] This approach to missing values was taken by almost all LSMS surveys we have been able to review. See for example the documentation provided in this link for Bosnia and Herzegovina, Ethiopia, and Nigeria: https://microdata.worldbank.org/index.php/catalog/lsms.

[36] The real interest rate represents the value one could have earned if the money was invested in a financial asset instead of a consumer good.

$$\delta_i = 1 - \left(\frac{V_{i,t}}{V_{i,t-T}}\right)^{\frac{1}{T}}$$

We can then apply the prevailing real interest rate in Jordan for the year, r_t, to compute the rental equivalent for each durable good. We will thus calculate the total annual durable goods consumption for household *h* as the sum of the rental equivalent of all goods owned by the household:

$$Durables_h = \sum_{i=1}^{I} rental\ equivalent\ of\ durable_{ih}$$

4. Housing

Like durable goods consumption, we will calculate the housing sub-aggregate as the value or benefit that a household receives from occupying a dwelling. For renters, we will use the monthly rent paid by the household. For non-renters, we will use the implicit rental value of the dwelling by asking the household for the estimated monthly market rental value. This follows the way the 2017-2018 Jordan HIES and many LSMS surveys capture implicit rental value. For households living in refugee tents, we will set the consumption value of their housing to zero.[37] We will thus calculate the annual housing expense for household *h* using the following equation:

$$Housing_h = monthly\ rent\ (or\ implicit\ rent)_h * 12$$

5. Debt repayments

Debt repayments will include the total value of payments made by the household during the previous year towards debt incurred prior to the start of the livelihood program for the first cohort (that is, by the end of 2021).[38] We will cover repayments on debt obtained across all types of formal and informal sources. Formal sources of credit include banks, microfinance institutions, and other private lenders, while informal sources include moneylenders, shopkeepers, friends, and neighbors. Thus, the debt repayment component of the consumption aggregate will be the value of debt repaid in the past year for household *h* across all sources *s*:

$$Debt_h = \sum_{s=1}^{S} debt\ repayment_{sh}$$

6. Other measurement considerations

a. Adjusting for price differences across locations

To ensure the comparability of consumption across households, it is common to account for differences in price levels of food and other goods across locations by constructing a location-level price index and using this to deflate the consumption aggregate (Deaton and Zaidi 2002). However, we propose not

[37] It is unclear how to compute the implicit value of rent for households living in tents. However, recent estimates suggest that most refugees in Jordan (73 percent) live in proper housing instead of tents (Tiltnes et al. 2019), and we expect refugees to comprise only about 30 percent of the sample. Therefore, assuming a zero implicit value of rent for these households will not substantially affect our consumption measure and impact estimate.

[38] As described in Chapter IV, we plan to exclude payments made towards debt incurred after the end of 2021 from the consumption aggregate to avoid double counting. Nevertheless, we will capture the total level of debt at the time of the impact evaluation survey in mid-2024 as an additional measure of financial wellbeing.

conducting price adjustments for this study, for two main reasons. First, based on information from our local evaluation partners, we understand that although prices can vary across locations in Jordan, the variation is not substantial and would not justify the additional analytical complexity required to conduct price adjustments. Second, in the context of a matched comparison group design where the treatment and comparison groups are balanced by location, price differences across locations would to a large extent average out.

b. Adjusting for household size

Because larger households tend to have higher consumption without necessarily being better off, we will adjust the consumption aggregate to account for household size, as measured at the time of data collection. We will do this by dividing the consumption aggregate for each household by the number of members in the household, transforming the measure into per capita terms. This measure of consumption assumes that all individuals in the household have the same needs and that consumption is shared equally among household members. An alternative approach would be to calculate an adult equivalent household size accounting for differences in consumption requirements by age and/or gender (for example, because children might have lower consumption needs than adults). However, for simplicity we will adjust the consumption aggregate only by the household size; we do not expect this to substantively affect the impact estimates because household size should be similar in the treatment and comparison groups.

c. Handling missing values and outliers

Missing values in the data may arise because of item non-response (answers of "don't know" or refusals in the survey). For households that report consumption of a good or service but are unable to provide the specific amount consumed, we will impute missing values by replacing them with the median amount consumed of the good or service by households in the same study group (treatment or comparison) and location. For households who are unable to provide information about whether they consumed a good or service, we will set consumption of the item to zero because we assume that household who cannot remember consuming an item is likely not to have purchased it; we will treat refusals in the same manner.

We will identify outliers in the data at the level of the four main consumption categories, plus debt repayments, after aggregating the components within each category. It could be challenging to identify outliers at a finer level such as at the item level, especially in the case of less common consumption items. We will define outliers as values greater than the 99th percentile of the category for households with the same treatment status and will top-code these values to value associated with the 99th percentile.

Appendix C. Template for IGA memos

Estimated business metric:

- Cohort 1: X percent (to one decimal place)
- Cohort 2: Y percent (to one decimal place, second memo only)
- Cumulative for cohorts 1 and 2: Z percent (to one decimal place, second memo only)

1. Introduction
2. Sampling approach
3. Data collection approach (including timing, response rates, and analysis sample sizes)
4. Definition and calculation of business metric
5. Analysis approach
6. Findings

Table C.1. Sample characteristics, percentages unless otherwise noted

Characteristic	Sample size	Mean
Women		
Youth		
Refugees		
Location:		
Zarqa		
Amman		
Aidon		
Kuforsum		
Russifieh		
Other socio-demographic characteristics[a]		

[a] These characteristics, which will be drawn from the vulnerability assessment, could include number of household members, head of household status, age, literacy, education level, disability status, and so on.

Table C.2. Business metric, overall and by subgroup

Sample	Sample size	Active business: at least one sales transaction in past 60 days (percent)	Active business: no sales transaction but at least one purchase transaction in past 60 days (percent)	Active business: no sales or purchase transaction but productive activity in past 60 days (percent)	No active business but formally employed (percent)	Total value of business metric (percent)	95 percent confidence interval (percent)
Full sample:							
All							
By gender:							
Women							
Men							
By age:							
Youth							
Non-youth							
By refugee status:							
Refugees							
Jordanians							
By location:							
Zarqa							
Amman							
Aidon							
Kuforsum							
Russifieh							

Note: For the full sample, among those validated as having at least one sales transaction in the past 60 days, X percent were validated by observing a business sales transaction, Y percent by reviewing transaction records, and Z percent were self-reported with details of the transaction provided. Among those validated as having no sales transactions but at least one purchase transaction in the past 60 days, X percent were validated by observing a business purchase transaction, Y percent by reviewing transaction records, and Z percent were self-reported with details of the transaction provided. The memo for the first cohort will present a single table; the memo for the second cohort will present a table for each of the first two cohorts separately and a combined table.

Table C.3. Additional outcomes, overall and by subgroup

Sample	Sample size range	Mean: outcome 1 (units)	Mean: outcome 2 (units)	Mean: outcome 3 (units)	Mean: outcome 4 (units)	Mean: outcome 5 (units)	Mean: outcome 6 (units)
Full sample:							
All							
By gender:							
Women							
Men							
By age:							
Youth							
Non-youth							
By refugee status:							
Refugees							
Jordanians							
By location:							
Zarqa							
Amman							
Aidon							
Kuforsum							
Russifieh							

Note: The memo for the first cohort will present a single table; the memo for the second cohort will present a table for each of the first two cohorts separately and a combined table.

Appendix D. Template for consumption memo

Estimated household consumption metric: X standard deviations (to two decimal places)

1. Introduction
2. Matching and sampling approach
3. Data collection approach (including timing, response rates, and analysis sample sizes)
4. Definition and calculation of household consumption metric
5. Analysis approach
6. Findings

Table D.1. Equivalence of the treatment and comparison samples after matching

Characteristic	Treatment mean	Comparison mean	Difference
A			
B			
C			

*/**/*** Difference significantly different from zero at the 1/5/10 percent level, two-tailed test.

Note: Sample size is X for treatment and Y for comparison.

Table D.2. Impacts on consumption per capita, overall and by subgroup

Sample	Treatment sample size	Comparison sample size	Treatment mean, μ_T (dinars)	Comparison mean, μ_C (dinars)	Difference, $\mu_T - \mu_C$ (dinars)	Standard deviation in comparison, σ_C (dinars)	Standardized difference, $(\mu_T - \mu_C)/\sigma_C$ (standard deviations)
Full sample:							
All							
By gender:							
Women							
Men							
By age:							
Youth							
Non-youth							
By refugee status:							
Refugees							
Jordanians							
By location:							
Zarqa							
Amman							

Sample	Treatment sample size	Comparison sample size	Treatment mean, μ_T (dinars)	Comparison mean, μ_C (dinars)	Difference, $\mu_T - \mu_C$ (dinars)	Standard deviation in comparison, σ_C (dinars)	Standardized difference, $(\mu_T - \mu_C)/\sigma_C$ (standard deviations)
Aidon							
Kuforsum							
Russifieh							

*/**/*** Difference significantly different from zero at the 1/5/10 percent level, two-tailed test.

Note: Treatment means and treatment-comparison differences are regression adjusted, with controls for A, B, and C.

Table D.3. Impacts on consumption per capita for the full sample, by consumption category

Sample	Treatment mean, μ_T (dinars)	Comparison mean, μ_C (dinars)	Difference, $\mu_T - \mu_C$ (dinars)	Standard deviation in comparison, σ_C (dinars)	Standardized difference, $(\mu_T - \mu_C)/\sigma_C$ (standard deviations)
Total:					
All items					
By category:					
Food items					
Non-food items					
Durables					
Housing					
Debt					

*/**/*** Difference significantly different from zero at the 1/5/10 percent level, two-tailed test.

Note: Treatment means and treatment-comparison differences are regression adjusted, with controls for A, B, and C. Sample size is X for treatment and Y for comparison.

Mathematica Inc.

Princeton, NJ • Ann Arbor, MI • Cambridge, MA
Chicago, IL • Oakland, CA • Seattle, WA
Tucson, AZ • Woodlawn, MD • Washington, DC

EDI Global, a Mathematica Company

Bukoba, Tanzania • High Wycombe, United Kingdom

mathematica.org