AIR®
AMERICAN INSTITUTES FOR RESEARCH®

Impact Evaluation of the Nigeria For Women Project

Inception Report

APRIL 2021

Thomas de Hoop | Adria Molotsky | Christopher Paek | Garima Siwach
Rosa Castro-Zarzur | Olayinka Adegbite | Iyabo Adeoye

MAKING RESEARCH **RELEVANT**

Impact Evaluation of the Nigeria For Women Project

Inception Report

APRIL 2021

Thomas de Hoop | Adria Molotsky | Christopher Paek | Garima Siwach
Rosa Castro-Zarzur | Olayinka Adegbite | Iyabo Adeoye

1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
202.403.5000

www.air.org

Contents

Page

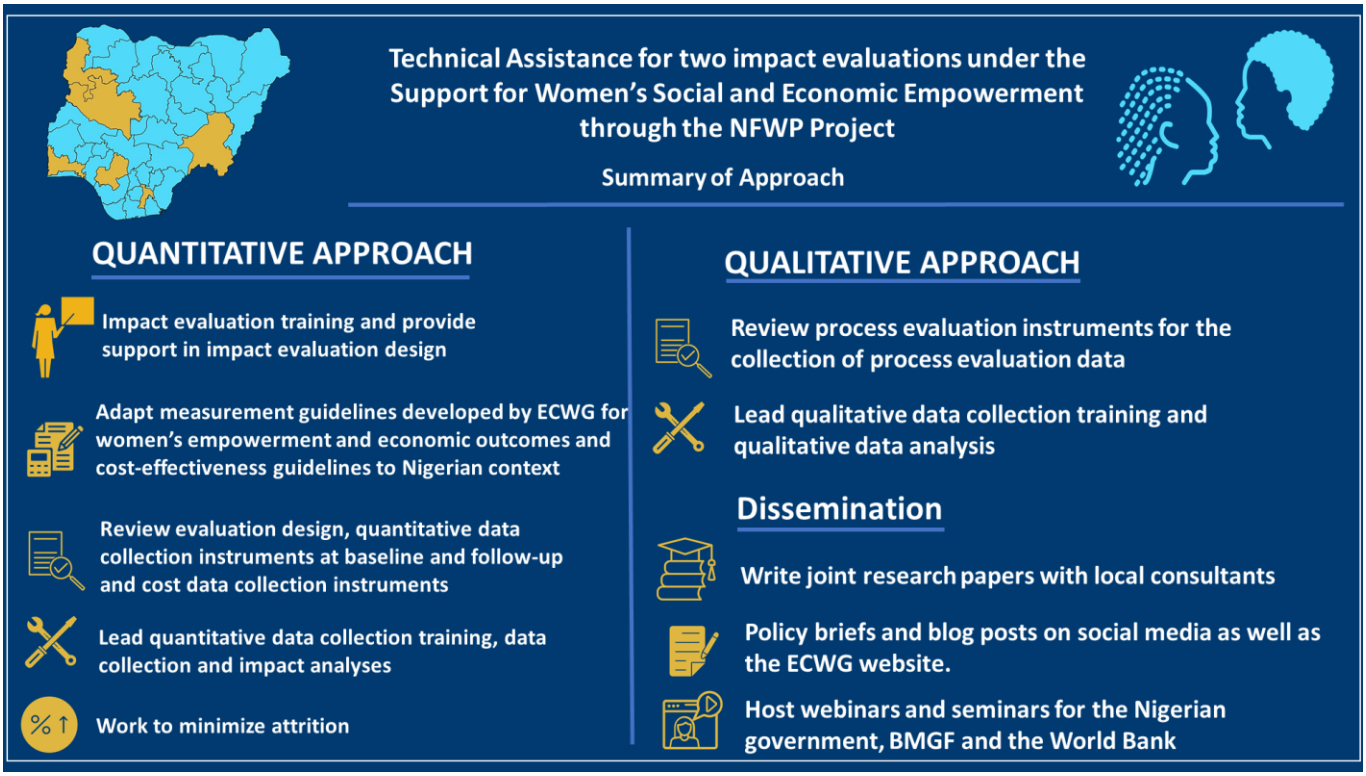
Introduction

The World Bank, in partnership with the Nigerian Ministry of Women's Affairs and Social Development (MOWASD) and the Bill & Melinda Gates Foundation (BMGF), is supporting the implementation of the Nigeria for Women Project (NFWP). The program aims to facilitate access to economic markets for women using a model of women's affinity groups (WAGs).

The American Institutes for Research (AIR) will conduct quasi-experimental impact evaluations to determine (a) the gendered impacts of the NFWP and (b) the impact of the NFWP on group functioning and inclusiveness. We will combine the two impact evaluations with a formative assessment, a process evaluation, and the collection of data to monitor the intervention. We will also assess both the cost-effectiveness and the return on investment of the NFWP. For both impact evaluations, we will explore opportunities to identify impacts for individual program components, such as messages related to social norms and the layering of health interventions and interventions to reduce gender-based violence (GBV). These evaluations of individual program components could include randomized controlled trials or quasi-experimental studies to determine effects of these interventions relative to a control or comparison group that received access to the NFWP but did not receive social norms, health layering interventions, or layered interventions to reduce GBV. Figure 1 depicts a summary of the evaluation approach.

The rest of this report describes the evaluation approach in more detail. We first provide information about the context and background of the NFWP followed by a description of the program and the theory of change. Then we present a detailed evaluation approach for each of the evaluation components, followed by a discussion of the dissemination strategy. Next, we present our approach to quality assurance, risk mitigation, and our workplan.

Figure 1. Approach Summary



Background

Women and girls in low- and middle-income countries such as Nigeria continue to face societal and structural gender equality challenges in areas that relate to education, employment, access to financial markets, and health that limit their opportunities and well-being. Overall, Nigeria is among the 10% of countries with the highest levels of gender inequality, according to the Social Institutions and Gender Inclusion index of the Organisation for Economic Co-operation and Development (Organisation for Economic Co-operation and Development [OECD], 2019). This index comprises indicators on civil liberties, restricted resources and assets, and son bias, among others. In 2018, only 22.1% of women in Nigeria had an account at a financial institution (DHS, 2018). As a result, women often are unable to make future investments, limiting their ability to respond to emergencies or other negative shocks. Further, in 2018, 29.5% of women between the ages of 15 and 49 reported having experienced intimate partner violence (DHS, 2018) and 64.1% reported having a partner with controlling behavior (Benebo, Schumann, & Vaezghasemi, 2018). Maternal mortality remains a significant challenge in Nigeria; a Nigerian woman has a 1 in 22 lifetime risk of dying during pregnancy or childbirth or postpartum/postabortion (World Health Organization [WHO], 2019).

With respect to intra-household decision-making, by 2018, 44% of married women participated alone or jointly with their husband in decisions regarding their health care, 40% participated in decisions about major household purchases, 60% participated in decisions about visits to their family and relatives, and 34% participate in all decisions (either alone or jointly), a slight increase since 2013 (31%). While the percentage of married women who say that they are not involved in any of the three specified household decisions fell from 48% to 37% between 2013 and 2018, a considerable proportion of married women still do not participate in intrahousehold decisions that affect their well-being (DHS, 2013 and 2018).

Women's groups with economic objectives, such as self-help groups, savings groups, and livelihood groups, have emerged as an important means of increasing gender equality as well as women's well-being and empowerment and their access to opportunities, including in sub-Saharan Africa (e.g. Karlan et al., 2013; Desai et al., 2019; de Hoop et al., 2020). Across sub-Saharan Africa informal women's groups have a long history and come together for many different purposes (Desai et al., 2019; de Hoop et al., 2020). However, women's groups have only recently become institutionalized and implemented at scale in sub-Saharan Africa. For example, savings groups have expanded considerably after experiments with Village Savings and Loan Associations (VSLAs) implemented by CARE in sub-Saharan Africa (Brody et al., 2015). However, the level of institutionalization of women's groups in Nigeria differs by type and location, and currently women's groups do not have the composition, capacity, and structure to support all Nigerian women in need of collective action. Nigeria has a diverse network of women's groups that focus on social cohesion, savings, and credit, including social clubs, religious groups, savings groups, and business organizations. For marginalized women seeking to join more formal groups, such as business and trade organizations, considerable barriers and restrictions exist. Unmarried women, women in northern regions, and poorer and less educated women face particular barriers and are at high risk for being excluded from these groups (Desai et al., 2018). In addition, some women are reluctant to join women's groups because of past negative experiences with microfinance organizations that were strict about repayment or which corruption led to loss of money, or because of restrictive social norms, men's negative opinions about women's groups, and women's lack of self-confidence (Desai et al., 2018).

Program Description

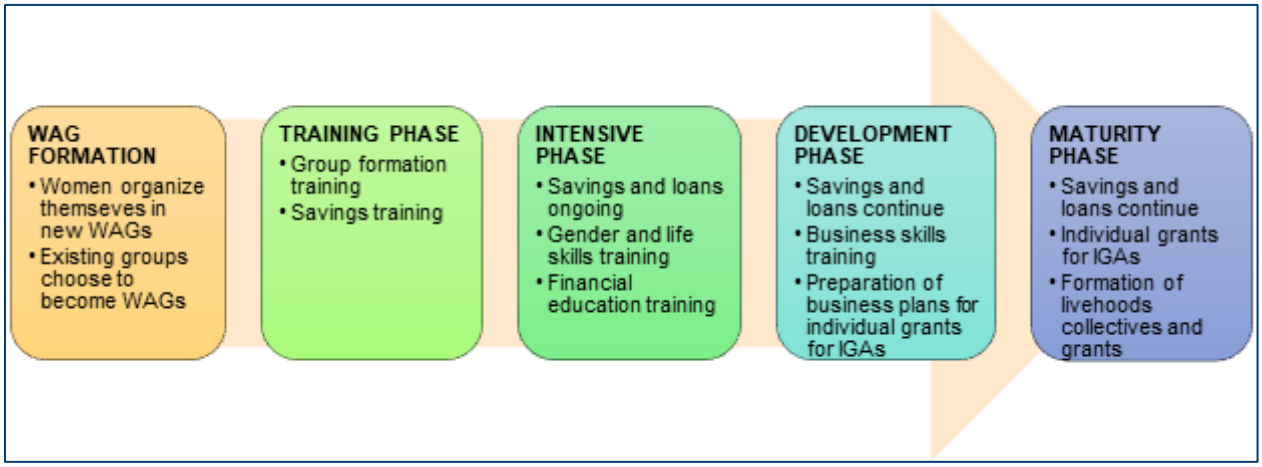
The World Bank, in partnership with the Nigerian MOWASD and the BMGF, is supporting the implementation of the NFWP. The program aims to improve women's livelihood opportunities and facilitate their access to economic markets using a model of women's affinity groups (WAGs). By providing opportunities in four components—the development of social capital, the building of livelihoods, the creation of partnerships, and messaging about gender and other social norms—the program seeks to overcome institutional and social barriers that currently

restrain economic outcomes for women, in addition to stimulating the development of social capital, supporting the building of livelihoods, and influence attitudes and behaviors related to gender equality and discriminatory social norms. The project also has an integrated feedback component of monitoring and evaluation that allows for constant learning and improvement.

The NFWP works with new and existing women's groups comprised of women over the age of 18, and targets women who are considered part of the "Missing Middle". This group of women consists of women who are economically active, but live close to the poverty line, and are thus vulnerable to shocks, which negative effects may bring these women below the poverty line in the absence of savings or other women's groups. Recent studies on women's groups and COVID-19 indicates that participation in savings and other women's groups could limit the negative consequences of shocks, such as COVID-19, for example because participation in such groups could enable women to make use of past savings and access to credit to cope with negative shocks (Namisango et al., 2021; Walcott et al., 2021). It is more challenging for women who are not economically active to participate in savings groups because participation in such groups requires regular savings. Programs such as cash transfers may bring larger benefits for economically inactive women. However, economically active women could benefit from participation in savings and women's groups because this participation limits their vulnerability to negative shocks. By focusing on including women in existing women's groups, the project seeks to help women benefit from existing structures and knowledge and increase diversity in these groups. Once the women are part of the group, they could access knowledge and resources available within the group, such as individual livelihood grants and trainings, which can help women start or further expand their economic activities.

The first year of NFWP implementation focuses on the formation and strengthening of WAGs through a five-phase process (see Figure 2 below). Throughout these five phases, groups will receive trainings in savings and credit, financial literacy, gender and life skills, and business skills. In addition, the program will select WAG members to receive individual grants to create or expand their income generating activities, after the development of business plans. In addition, the program aims to create and expand livelihoods collectives to form livelihoods partnerships at the end of the formation process. The livelihood collectives will also be eligible to receive grants from the project. Concurrently, the program will carry out a series of behavior change, and awareness raising-related activities targeted to gatekeepers as well as all women and men in the community to influence social norms and gender beliefs and behaviors at the community-level.

Figure 2: Formation and Strengthening of WAGs.



Source: Community Operating Manual (COM) – Nigeria For Women Project

From year 2 onwards the program will also introduce health and GBV layering activities to leverage the WAGs as a platform to deliver health and GBV programming to many women at once. This approach is in line with the suggestion by Diaz-Martin et al. (2020) that groups could deliver benefits at a lower cost per program participant than programs that focus on delivering health information at the individual level. While the WB and the Nigerian MOWASD have not yet made decisions about the specific contents of the health programming, the layering activities will likely focus on maternal and child health and may include an existing curriculum developed by the Nigerian Ministry of Health or adapt successful programs implemented by NGOs.

The NFWP is a federal program (supported by a $100 million loan from the World Bank) that is implemented gradually, starting with six states and 18 local government areas (LGAs) in Nigeria. During the first phase of the program, the NFWP seeks to expand participation of women in women's groups by reaching 324,000 women through approximately 21,600 WAGs in Ogun, Taraba, Kebbi, Abia, Niger, and Edo state.[1] This initial phase of implementation seeks to institutionalize the WAG model in Nigeria while simultaneously providing the opportunity to test and evaluate the model in various state and LGA contexts. The evaluation will assess the mid- to long-term impacts of the introduction of WAGs on group-level outcomes and on the livelihoods and the economic, social, and psychological empowerment of women and households in order to inform future project investments. In addition, we aim to assess the impacts of the program and layering activities related to social norms, health and GBV outcomes to the extent possible.

[1] It is, however, not certain yet that program implementation will take place in Edo state.

Even though the initial phase seeks to provide lessons for potential future expansions, we anticipate that implementation will vary widely across the six states. The states and LGAs differ significantly, and each state government will likely adapt the programming to its specific context, particularly around social norms. First, entry points for mobilizing women might differ by region because of varying social and cultural norms. For example, in the northern regions, women have few opportunities to socialize outside the home other than through informal networks and social ceremonies. The program will likely have to leverage these informal networks and social ceremonies to find local leaders that can support WAGs. In southern Nigeria, however, the program could approach women through formal networks, such as through community leaders or existing women's groups. Second, WAGs may set different priorities. While all WAGs will engage in savings and credit, savings and credit amounts may differ by WAG. Because of differences in familiarity with women's groups and opportunities to participate in the past, women may also have various levels of confidence, skills, and understanding of concepts of banking, savings, credit, organization, and other economic matters (Desai et al., 2018).

Theory of Change

To inform our study design, we developed a theory of change (ToC) in consultation with the World Bank and the NFWP Implementation Team at the national and state levels focused on the implementation of WAGs and the impact of those groups on group-, household-, individual-, and community-level outcomes (see Figure 3 on page 15). Below we describe the pathways through which we hypothesize changes will occur. We link these pathways to the NFWP components targeted at the group level (Savings and Livelihoods Trainings, Social Norms Messages and Trainings, Health and GBV Layering) and at the community level (Social Norms Messages). The description focuses on a generic ToC model for Nigeria and does not distinguish between groups that existed before the intervention and transformed into WAGs (e.g., social groups, savings groups, livelihoods groups) or newly formed WAGs, because all groups will receive the full package of trainings regardless of their prior status or type. In this way, the ToC focuses on the pathways through which the intervention components are expected to generate impacts at different points in time, differentiating between the unit of measurement of the changes (i.e., group, individual and household, and community). However, we did include potential moderators to generate hypotheses about how WAGs with different program components, different membership composition, differences in the group's implementation before the transformation to WAGs, and in other contexts may show different effects along the causal chain of the ToC.

The ToC suggests that the NFWP can achieve improvements in women's economic empowerment through several mechanisms. Savings and livelihoods trainings can lead to greater access to savings mechanisms and increases in collective savings, which can in turn enable women to gain access to individual or group credit. Combining increased access to credit with improved business skills can then help women start or expand their income generating activities across high-productivity agricultural and non-agricultural sectors. The creation and expansion of women's individual and collective businesses could increase women's income and asset ownership, which could stimulate their bargaining power within the household.

Social norm messages at the group- and community-level, which aim to change discriminatory social and gender beliefs and practices held by both men and women, could also lead to changes in women's empowerment and social norms at the community level. In particular, increased social support within groups can help WAG members reflect on gender stereotypes and discriminatory dynamics, change group members' attitudes, and improve women's self-confidence and well-being. These changes in women's attitudes could later result in improvements in women's bargaining power within the household, as women could leverage the economic and social opportunities brought to them through the NFWP. Social norm messages at the community level could lead to additional lessons for men and other community members and could lead to changes in gender and other social norms following changes in women's attitudes. However, changes in social norms generally are non-linear and women may have to transgress social norms in order to change them. Such transgressions could lead to community backlash, which may create dynamics in which women return to their level of bargaining power before the NFWP because of social sanctions (de Hoop et al., 2014; Brody et al., 2015). The NFWP aims to limits such backlash by sending social norms messages at the community level as well engaging through community dialogues, which could lead to changes in social and gender beliefs and practices held by men and other non-WAG members.

The first step in the pathway of change is community sensitization and mobilization. This process will include meeting various stakeholders such as leaders and women in their safe spaces with the aim of introducing the project to the community and getting the approval to operate. Carefully selected and trained Ward Facilitators will then identify the women targeted by the program who are either not yet part of any women's group or are part of existing women's groups with limited governance structure and support. Community leaders, peers, and existing women's groups will then encourage vulnerable women living in poverty or marginalized for other reasons to join a WAG in their community.

Once women are part of a group, they could benefit from livelihood support through trainings, the preparation of business plans, livelihood grants, and increased credit from their WAG. The

grants and group credit could enable women to invest in existing or new individual or collective enterprises, which could generate profits and additional household income and could enable women to accumulate assets, especially after the preparation of business plans and livelihoods training help them to make productive investments across diversified agricultural and non-agricultural income-generating activities. The intervention also stimulates partnerships with outside actors, such as the private sector, civil society organizations, NGOs, and individuals, to identify innovations for the groups through a development marketplace. These partnerships provide women with the opportunity to strengthen networks and increase their knowledge and skills (e.g., further improve their business skills and complement what they would have learned through group trainings). The partnerships, networks, and skills development can also give the participating women a comparative advantage and can lead to increased access to financial markets, greater economic opportunities, and more income generation. This market access could result in synergies that could lead to additional effects of the livelihood support, business plans, and livelihood grants.

From Year 2, individual WAG members could also benefit from the health-layering component through health trainings and health-related group activities. The specific focus on health may contribute to women's knowledge about topics such as nutrition and maternal or neo-natal health. We hypothesize that this knowledge could result in changes in health behavior, including changes in health care spending, health-seeking behavior, usage of preventative and pre-natal care, and infant and young child feeding practices.

The generation of social capital through group support could improve trust, social networks, and social cohesion by enabling women to collaborate with peers in their community and discuss economic and social issues. We hypothesize that improved social cohesion and networking could have both individual- and group-level effects. For instance, discussions about social norms may affect women's attitudes toward empowerment and their psychological empowerment. The sense of confidence, dignity, and self-esteem that comes from women's empowerment could help women achieve "power within" and could positively influence intra-household dynamics—for example, by enabling women to gain bargaining power within the household with regard to financial decision and decisions on education and health (Barooah et al., 2019; Diaz-Martin, Gopalan, Guarnieri, & Jayachandran, 2020).

Social norms trainings can bring additional improvements in different women's empowerment domains because of their specific focus on raising awareness about gender stereotypes and gender-based violence (GBV). In addition, the community-level social norms intervention is designed to further strengthen the program's impacts on intra-household decision-making and GBV, as household and village members (men and women) engage in community activities and

receive targeted messages (via various channels including but not limited to radio programming, household and community dialogues, theater for change etc.), thereby reflecting upon these issues and possibly adopting better attitudes and behaviors.

The WAGs can also generate impacts at the group level. First, the composition of groups may change considerably after encouraging membership of marginalized women. These changes in group composition may lead to increased diversity either within groups or across groups, which may strengthen social cohesion within the community and may provide women with opportunities to pool risks and resources, thereby improving their resilience — particularly the resilience of marginalized women. Importantly, however, increased within-group diversity may limit the ability of women's groups to pursue joint goals; women with different characteristics likely have different objectives and different means to contribute to savings, which could limit the ability of groups to pool savings and risks. In contrast, the number of women's groups may increase if marginalized women form their own groups to pursue joint financial objectives.

The within-group structures of rotating leadership, group-based decision-making, and regularly planned meetings can help build trust, shared responsibility, and a sense of community. These group-level changes can lead to collective action, such as increasing the frequency of planned meetings or of interactions with third parties regarding training and innovation. Ultimately, these changed group behaviors may result in an increase in individual or collective savings, greater access to formal credit, and a reduction in fraud.

With the support of the Ward Facilitators, WAGs could spark the formation of livelihoods collectives, such as cooperatives, farm and non-farm producer organizations, social enterprises, producer-governed private limited companies, especially after groups receive livelihoods training and livelihoods grants. These trainings and grants can help groups set up livelihoods collectives that include all or a subset of WAG members. By leveraging the groups' increased human capital and members' social networks, and by gaining access to grants (for which collectives are eligible), women's livelihoods collectives could break into higher-productivity agricultural and non-agricultural sectors, improve their access to markets and value chains, and negotiate better transaction conditions (e.g., prices and quantities). Successful livelihoods collectives may also lead to improvements in individual and collective asset ownership and profits.

Increases in the number of WAGs can lead to general equilibrium effects and changes in prices and wages. For example, WAGs may increase demand for credit from women's groups, which likely will become available at lower interest rates than in informal settings. The increased demand for credit from WAGs may come at the expense of demand for informal credit, which

could in turn result in a reduction in informal interest rates. Hoffmann et al. (2020) provide evidence for such a mechanism based on a cluster randomized controlled trial of a self-help group program in India, and research by the American Institutes for Research under the Evidence Consortium on Women's Group (ECWG) shows that such general equilibrium effects are sometimes of greater significance for the cost-effectiveness of women's groups than the direct individual- or household-level impacts of women's groups (Siwach et al., 2021). The formation of collective enterprises may also result in increases in the demand for labor, which could in turn lead to increases in wages.

Achieving impacts through the described pathways depends on several critical factors. First, the successful mobilization of women will depend on the motivation and incentives of community leaders to identify and target marginalized women who do not yet participate in any well-functioning women's group. In addition, women should have sufficient time and a minimum level of resources to enable their participation in meetings, create a network, and contribute to collective savings. Using accumulated savings to make productive investments in the future further depends on women's ability to expand livelihood activities. In addition, women's savings can only increase sustainably if they can avoid having to take out a large portion of the money in the interim. Furthermore, not all changes will occur at the same time, and behavioral and changes in social norms in the community are likely slow and highly dependent on context and on the existing social norms. Finally, group-level impacts depend on the social cohesion among group members, regular meetings and savings, and the return on investment of investments in collective enterprises.

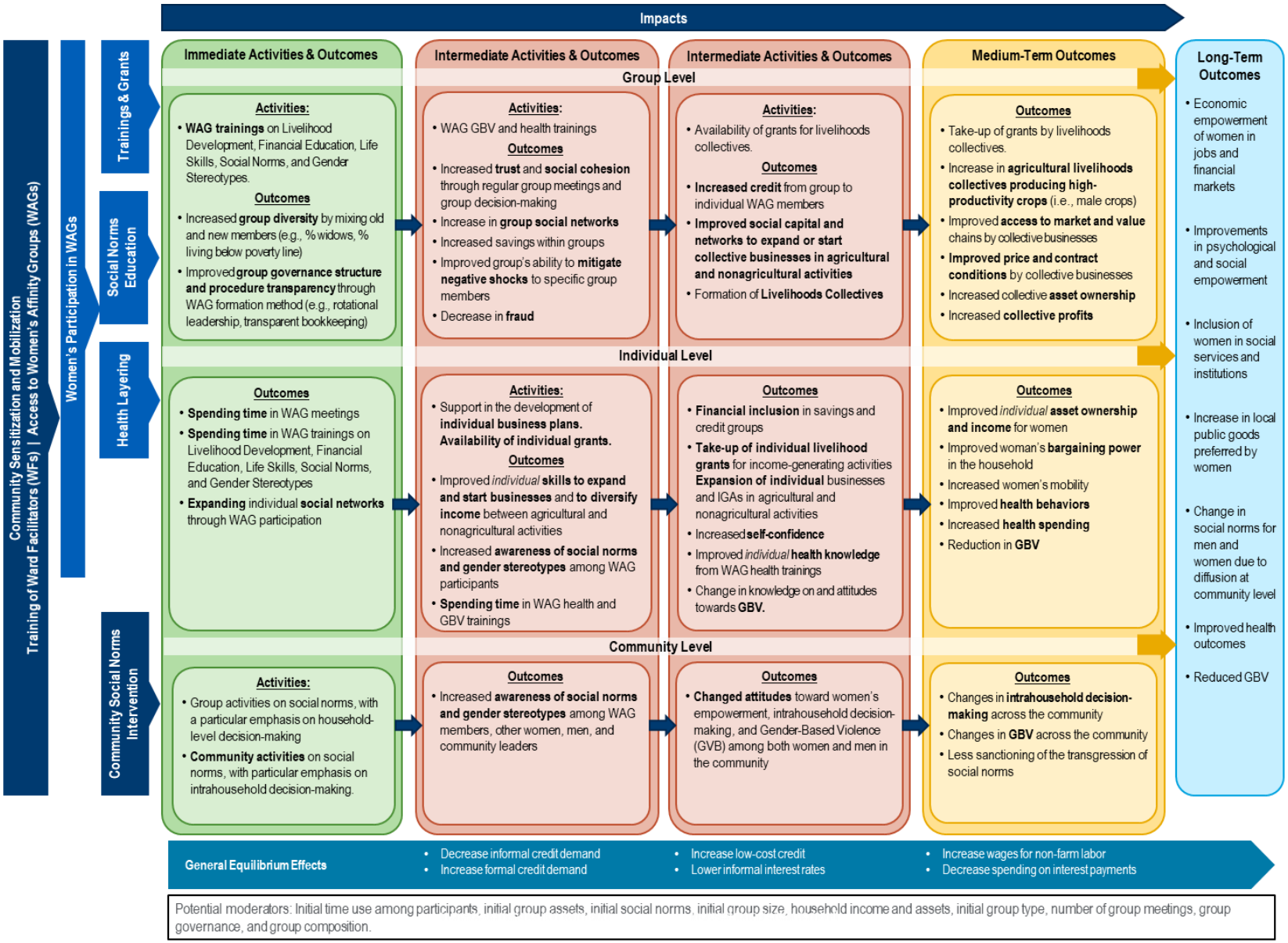
Although improvements in women's bargaining power in the household can reduce GBV, the current evidence on this mechanism is mixed. Some studies suggest that greater bargaining power can lead to backlash and possibly an increase in GBV. Nonetheless, in the long term, any backlash may reduce because of changes in social norms at the community-level (Brody et al., 2015). We hypothesize that women in communities benefiting from the social norm messages at the community-level and the WAGs may therefore have a larger likelihood of seeing reductions in GBV than women in communities with WAGs without social norms messages. This is particularly likely for women benefiting from the layering of additional GBV-interventions.

Different initial group types may lead to different impacts as they may condition the way WAGs evolve after the first WAG cycle ends (i.e., after trainings and Ward Facilitator support end and groups decide whether to continue saving or not). WAGs may have different effects depending on what components they include, and WAGs could evolve as part of their institutional evolution. For example, some groups may continue to focus on savings and credit after the first cycle, whereas other groups may choose to focus on social activities. We hypothesize that the

former group type may have larger impacts on women's financial inclusion, which may result in improvements in women's asset ownership and women's bargaining power in the household. Secondary benefits may include improvements in women's agency, and business opportunities. We hypothesize that the latter group type (i.e., social groups) may have larger effects on women's psychological empowerment if discussions about social issues focus on building women's self-confidence. Secondary benefits of these groups may include improvements in women's bargaining power, which could in turn result in improvements in women's agency, and economic outcomes, such as asset ownership and income.

Different group types also impose different costs, both program-level costs and opportunity costs related to the time women spend in group meetings and other group activities. Adding components will add ingredients to the intervention, which will increase the program costs. Furthermore, adding intervention layers and trainings may increase the time group members have to spend in group meetings and thus the opportunity costs of participation in WAGs. This may lead to changes in the composition of the group, for example by reducing the participation of women who face time constraints or have higher opportunity costs. Adding new topics to a group will either increase the time group members spend in meetings or reduce the time the groups can spend on other topics. For example, it may not be possible to spend an hour on livelihoods training if a group dedicates additional time to health education unless the group increase the number of meetings.

Figure 3: Theory of Change



Research Questions

The evaluation of the NFWP is guided by the following key research questions, which are based on the theory of change:

Impact Evaluation Questions

1. What is the impact of the NFWP on group-level outcomes?
    a. What is the impact of the NFWP on collective savings and credit?
    b. What is the impact of the NFWP on the set-up, sales, and profits of collective enterprises?
    c. What is the impact of the NFWP on group composition?
    d. What is the impact of the NFWP on group meetings and governance?
    e. What is the impact of the NFWP on social cohesion in groups?
2. What is the impact of the NFWP on individual-level and household-level outcomes?
    a. What is the impact of the NFWP on financial inclusion (access to formal credit, savings, and demand for informal credit)?
    b. What is the impact of the NFWP on economic outcomes (women's and men's asset ownership, household consumption, women's and men's income)?
    c. What is the impact of the NFWP on women's empowerment and agency (women's decision-making power, women's mobility, women's self-confidence)?
    d. What is the impact of the NFWP on gender-based violence?
3. What is the impact of the NFWP on community-level outcomes?
    a. What is the impact of the NFWP on social norms?
    b. What is the impact of the NFWP on informal interest rates?
    c. What is the impact of the NFWP on agricultural and non-agricultural wages?

Cost and Cost-Effectiveness Questions

1. What are the costs of the implementation of WAGs in Nigeria?
2. What are the drivers of the costs of WAGs in Nigeria?
3. What is the cost-effectiveness of WAGs in achieving group-level outcomes?
4. What is the cost-effectiveness of WAGs in achieving women's empowerment outcomes?
5. What is the cost-effectiveness of WAGs in achieving economic outcomes, such as consumption, asset ownership, and income?
6. What is the return on investment of WAGs?

Formative and Process Evaluation Questions

1. What are the barriers and facilitators to participating in women's groups and specifically WAGs?
2. How have contextual factors influenced program implementation?
3. What are the perceptions of WAG and other women's group participants and nonparticipants about the benefits and costs of participating in women's groups?
4. How do women's group participants and nonparticipants perceive community gender norms and their relationship with WAGs?
5. How do WAGs interact with nonmembers, including men and Ward facilitators?
6. To what extent can WAGs serve as a platform to deliver health and social protection programs?

The following section details our approach to addressing these research questions using a mixed-methods approach.

Study Design

AIR developed a comprehensive, mixed-methods quasi-experimental approach to determine the impact of the NFWP at the individual- and group-level, as well as the fidelity of its implementation and cost effectiveness. In the following sections, we describe the details of our

mixed-methods design starting with the quantitative methods followed by the qualitative methods.

Quantitative Study Design

In the first phase of the program, the WB and the MOWASD will implement the WAG model in six states and 18 LGAs based on agreed-upon criteria. The project will target 54,000 individual beneficiary women per state—that is, a total of 324,000 women aged 18 and above who are members of approximately 21,600 WGAs. We designed the impact evaluation of the program based on this proposed rollout as well as methodological principles that enable us to establish a counterfactual. This design requires rigorous methodologies to address the following question: What would have happened in the absence of the intervention?

Identification Strategy to Estimate Gendered Impacts of the NFWP at the Individual and Household Levels. We will use a quasi-experimental design to determine the impact of the NFWP by combining a difference-in-differences (DID) method (which compares the average change over time for the treated group to the average change over time for the comparison group) and a three-stage matching approach. However, it may be feasible to design RCTs for components that will be added to the NFWP at later stages, such as the layering of health and GBV interventions and the social norm interventions. We provide more details on different options to determine the additive effects of the social norm messages, and the health and GBV layering interventions below.

To ensure that treatment and comparison households or groups are similar, we propose to design and implement a matching process in three stages. We will first match the treatment LGAs to similar comparison LGAs, followed by the matching of treatment wards to comparison wards based on a listing survey in the second stage, and finally the matching of treatment households and groups (including but not exclusively WAGs) in the selected treatment wards to comparison households and groups in the selected comparison wards in the third stage. In the first stage, we will match treatment LGAs to comparison LGAs based on geographic characteristics. In the second stage, we will match households and groups in 126 treatment wards (an LGA typically has 10-15 wards) to households and groups in a select number of comparison wards based on variables relevant to the targeting and potential outcomes of the program as well as variables to characterize households and groups (we will gather these data using a listing survey or short census). In the third stage, we will match treatment households and groups to comparison households and groups based on the data from the full baseline survey.

In the first stage, we will match each of the 18 treatment LGAs to four neighboring comparison LGAs. We will work with the World Bank and the Nigerian government in selecting four potential comparison neighboring LGAs for each treatment LGA using nearest neighbor

matching so that we can achieve a sample size of 18 treatment LGAs and 72 comparison LGAs. Within each LGA we will then implement a listing survey in a minimum of seven but preferably 10 randomly selected wards (total sample size to be determined in close consultation with the data collection firm) to collect data on a small but important number of variables to match on.

In the second stage, we will match 126 wards in 18 treatment LGAs to a select number of wards in 72 comparison LGAs based on the data from the listing survey. We will conduct the matching using a nearest neighbor approach in which we will match each treatment ward to a minimum of one comparison ward based on aggregate data from the listing survey.[2]

In the third stage, we will match 1,170 treatment households and 252 treatment groups (including but not exclusively WAGs) to 4,680 comparison households and 1,008 comparison women's groups using a combination of individual-level data for both women and men, household-level data, and group-level data. We will prioritize outcome variables and variables that are highly correlated with the outcome variables in a Mahalanobis distance matching approach. These variables could include women's empowerment indicators; women's and household income, expenditures, and asset ownership; the education level and age of the household head and his spouse; food security; and various other individual- and household-level variables for the matching at the household level. For the group-level matching, we could include variables related to the group size, number of meetings, group type and objectives, and group composition (e.g., whether members speak a certain language or are part of a specific ethnic group). In addition, we could include aggregate indicators of women's group members, such as their average age, and education level. However, we will include variables that are not plausibly affected by the introduction of WAGs before the baseline survey. It is our understanding that some WAGs will already operate before the baseline survey. For that reason, we will collect recall data on group-level outcomes, savings, and credit to use in the matching process. For some other outcome variables (such as women's empowerment) it is not plausible that the WAGs affected them in such a short period of time, however, so we will include these variables in the matching process. It is also important to keep in mind that we may include non-WAGs in the treatment group sample, because it is not likely that all women's groups will transform to WAGs. Excluding non-WAGs in the treatment group could result in a bias in the impact estimates because the comparison group will also include WAGs that would not have transformed to WAGs even when they would have had the opportunity.

Lastly, we will combine DID analysis and matching in two ways as a robustness check. First, we could use matching to identify the comparison group and then compare changes in outcomes over time between NFWP women and the comparison group using DID analysis. Second, we could

[2] It is possible that we may have to sample enumeration areas as opposed to wards if maps are available for enumeration areas but not for wards. We will discuss the different options with the data collection firm.

pursue a reweighting approach in which we weight the comparison group by members' estimated chance of receiving treatment (Abadie, 2005), which is based on the likelihood that a woman or household would be part of NFWP. We propose to use both methods to explore the robustness of the findings. In refining this matching approach, we will work with the World Bank and the data collection firm to take advantage of their on-the-ground knowledge of the program implementation and the Nigerian context.

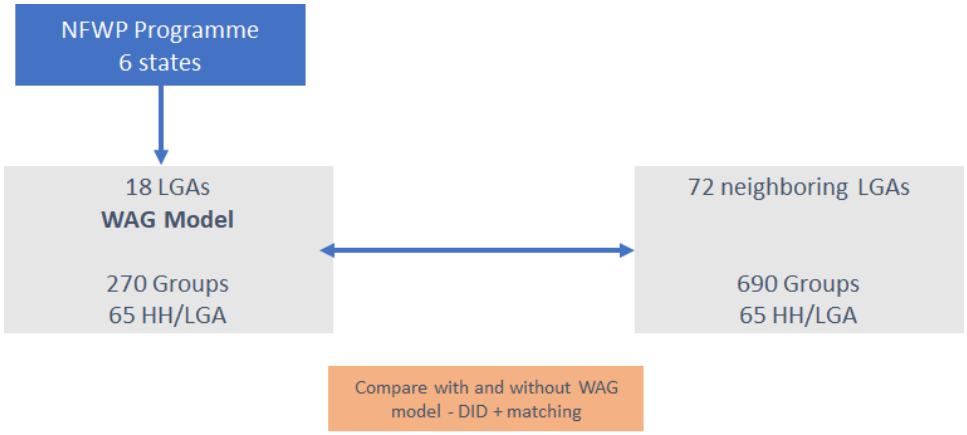
Identification Strategy to Estimate the Impact of Specific Interventions at the WAG Level. We will explore opportunities to design and implement experimental and quasi-experimental evaluations of specific interventions that are targeted to a select number of WAGs, groups, or group members. Such interventions may include messaging about social norms, and the layering of health and GBV interventions. These impact evaluations could serve to estimate the additional effects that social norms messaging may have on women's agency, men's attitudes toward women's empowerment, and the transgression of norms (for social norm messaging), the additional effects that health layering could have on health behaviors and health outcomes, and the additional impacts that GBV layering could have on attitudes to GBV and the likelihood of GBV.

However, it is unlikely that we can estimate the effects of all intervention components because we will then no longer have the statistical power to differentiate between the effects of the social norm intervention, health, and GBV layering. Distinguishing between the effects of all different components will either require a larger number of LGAs where the program is implemented or a combination of treatment assignment at the LGA level, the ward level, and the women's group level.

It is our understanding that some of the intervention components, such as social norms interventions, may happen before the planned midline survey, while others may only happen after the midline survey. Based on these considerations, we anticipate that we will estimate the impacts of the NFWP with some of the additional components during the midline survey, the impacts of the NFWP with and without the additional components during the endline survey, and the additional effects of specific components during the endline survey.

Figure 4 summarizes the impact evaluation design to determine the impact of the full intervention, including the number of LGAs, groups, and wards that will be required for the impact evaluation. The sample sizes are based on power calculations that we describe in the next section.

Figure 4. Evaluation Design for Quasi-Experimental Study to Determine the Effects of the Full Intervention



Identification Strategy to Estimate the Impact of social norm messages: We already discussed options to determine the effects of social norm messages with the WB and the NFWP team because this intervention component will be introduced in the first year of implementation. The program would have to consider limited or staggered implementation of social norms interventions and undertake significant additional data collection to estimate the effects of the social norm intervention. We present two options below to credibly assess the additive effects of the social norm intervention.

Option 1: Randomized Controlled Trial with 8 treatment LGAs and 7 control LGAs in five States

- With this option 8 LGAs in Ogun, Taraba, Kebbi, Abia, and Niger would be randomly selected to receive the social norm messages at the community level.[3]
    - Each state will have LGAs that do and that do not receive the social norms messages intervention.
- 7 LGAs will be randomly selected to not receive the social norm messages at the community level. If possible, they would not receive the social norms messages during the entire evaluation period. Alternatively, they would receive the social norms messages 1.5-2 years after the start of the implementation in the other LGAs. In the latter case, we would, however, only be able to estimate impacts of the intervention after 3 years relative to a counterfactual where the intervention is implemented for 1.5-2 years. This may reduce our ability to detect statistically significant effects.
- This option will require two extra data collection rounds (with baseline, midline, and endline data collection already planned) about social norms and other key outcomes:
    1. Baseline (May 2021)

[3] With this option, we assume that only 15 LGAs (in Ogun, Taraba, Kebbi, Abia, and Niger) will participate in the NFWP project given the ongoing issues in Edo state. The second option assumes implementation in one additional state.

2. Midline (May 2022)
3. *Second Midline for Social Norms (July 2022)*
4. Endline (May 2023)
5. *Second Endline for Social Norms (July 2023)*

- We would collect data in 10 wards per LGA. [4]

Option 2: Randomized Controlled Trial with 9 treatment LGAs and 9 control LGAs in six States

- With this option 9 LGAs in Ogun, Taraba, Kebbi, Abia, Niger and another additional state would be randomly assigned to receive the social norm messages at the community level.
- Again, each state will have LGAs that do and that do not receive the social norms messages intervention. If possible, they would not receive the social norms messages during the entire evaluation period. Alternatively, they would receive the social norms messages 1.5-2 years after the start of the implementation in the other LGAs. In the latter case, we would, however, only be able to estimate impacts of the intervention after 3 years relative to a counterfactual where the intervention is implemented 1.5-2 years. This may reduce our ability to detect statistically significant effects.
- This option will require <u>two additional</u> data collection rounds about social norms and other key outcomes:

    1. Baseline (May 2021)
    2. Midline (May 2022)
    3. *Second Midline for Social Norms (July 2022)*
    4. Endline (May 2023)
    5. *Second Endline for Social Norms (July 2023)*

- We would collect data in 10 wards per LGA.[5]

It is likely that further postponing social norms community messaging in the control group (until after the evaluation period) would increase statistical power. For this reason, this is the preferred option from an evaluation perspective. While it is unclear how much further postponing the intervention will help in gaining additional statistical power, it is possible the evaluation will only detect statistically significant effects if the social norms community messaging is postponed until after the evaluation period. This will depend on the actual effect of the intervention, however.

Identification Strategy to Estimate the Impact of health and GBV layering: We propose several options to estimate the effects of additional components, such as the health and GBV layering:

[4] LGAs have between 10 and 11 wards.
[5] LGAs have between 10 and 11 wards.

- A randomized controlled trial based on a phased roll-out of the program at the ward level.
- A randomized design with different levels of dosage at the ward level.
- A quasi-experimental regression discontinuity design.
- A quasi-experimental matching approach.
- A randomized encouragement design.

Randomized Controlled Trial of the health and/or GBV layering: For this option, we propose randomly assigning wards that would receive the additional intervention first or to a delayed control group that will receive the intervention later (after the data collection to estimate effects). An RCT permits us to directly attribute any observed differences between the treatment and control groups to the health and/or GBV layering because of the random assignment of participants to health and/or GBV layering (Duflo, Glennerster, & Kremer, 2007). We suggest randomizing at the ward level to limit the risk of bias caused by spillover effects of the health layering.

Importantly, randomization of the additional components is a means of achieving unbiased impact estimates and is not a goal in itself. While we recommend RCTs where feasible and desirable, we recognize that they may not be feasible or desirable for health and/or GBV layering. For this reason, we also propose other options to determine the impact of the health and/or GBV layering.

Randomized Design with different levels of dosage: A second option is to allocate the health and/or GBV layering to all treatment wards, but with different dosage. Evidence from India indicates that current health layering efforts are often not sufficient to achieve positive impacts on health outcomes, which may be because of the low dosage of the intervention (Desai et al., 2020).

We propose a randomized design where a sub-sample of wards would be randomly assigned to receive health and/or GBV layering or with a higher dosage and a sub-sample of wards would be randomly assigned to receive health and/or GBV layering with a lower dosage. Specifically, we suggest that treatment wards receive substantially longer health and/or GBV trainings than control wards.

While this design does not allow for a direct comparison between groups who do not receive the intervention, and those that do, it would allow us to examine the effects of longer health and/or

GBV trainings relative to shorter health and/or GBV trainings. This design would also allow the WB and Nigerian government to rollout the layered interventions to all groups at the same time and eliminate the need for a delayed control group. However, the design would require significant coordination with NGOs to ensure that some NGOs provide health and/or GBV layering with a low dosage and other NGOs provide health and/or GBV layering with a high dosage.

Quasi-Experimental Regression Discontinuity Design: Our third option proposes a quasi-experimental regression design. Similar to the generic RCT, this option relies on a phased rollout of the layered interventions by an arbitrary geographical boundary such as a ward. Wards or villages would be randomly assigned to receive the interventions starting in year two while the neighboring comparison wards would not receive the intervention until after the endline evaluation. The underlying premise of the RDD is that individuals directly on either side of the ward boundary (cut-off) are likely very similar in their characteristics and differ only in that they happen to reside on either side of the boundary. We would work with the World Bank and the Nigerian government in determining the exact bandwidth (in this case, distance from the geographical border between wards) to use for our analysis such that the we restrict the analysis to only those we would expect to have similar characteristics.

Quasi-Experimental Matching Approach: Another quasi-experimental approach to assess the additive effect of the layered interventions is through a matching approach. In this instance, we would work with the World Bank and the Nigerian government to purposively select wards to receive the interventions due to group readiness, ward health status (need), or local NGO availability to deliver trainings, for example. Then, like our matching of treatment and comparison wards, households, and groups for the overarching impact evaluation, we would use matching techniques to identify groups and group members to form the comparison group from NFWP treatment wards. This approach also requires a phased rollout of the interventions as we would need to ensure a viable comparison group (i.e., WAGs under the NFWP who were not receiving the layered interventions).

Randomized Encouragement Design: If the World Bank or Nigerian government are unable or unwilling to implement a phased rollout of the layered interventions, we could randomly select individuals to receive additional "nudges" or encouragement to attend the health and/or GBV layering. For example, individuals could receive SMS messages or short phone calls to encourage their participation in the health and/or GBV layering. Insights from behavioral economics suggest that small nudges and reminders can induce behaviors and persuade individuals to attend trainings or events that they otherwise may have skipped. Our analysis would then assess the additional benefit of receiving more information and encouragement to attend trainings on behavior change.

Importantly, it may not be feasible to precisely estimate the impact of the health and/or GBV layering when both social norm messages and health/and or GBV layering are randomly assigned at the LGA level. Distinguishing between the effects of the social norm messages and the health and/or GBV layering will require variation at the ward level for at least one of the program components if the other program component is randomly assigned at the LGA level.

Regardless of the evaluation design, we suggest using a monitoring approach whereby we would employ shorter and more frequent data collection via short surveys administered through phone calls or SMS. To ensure the greatest value, we would align these data collection points with the timing of the trainings. We would then collect data for a sub-sample immediately following the trainings as well as at regular intervals into the longer-term to identify immediate and as well as sustained changes in individuals' KAP related to health and social norms. This rapid monitoring approach would also facilitate instant feedback on the implementation of the interventions for the World Bank and Nigerian government to simplify the process of adjusting implementation for improved outcomes, as needed.

Sampling

Since we lack sufficient roster data on households and women in the treatment and comparison LGAs, we will let the data collection firm conduct a listing exercise in a select number of randomly chosen wards in each of the treatment and comparison LGAs, followed by a sampling of groups, households and women in these wards. We anticipate that we will be able to select five wards in each of the treatment and potential comparison LGAs though this will depend on available resources of the data collection firm.

The data collection firm can then conduct a short census (the listing survey) in each of the selected wards to identify all women's groups and households with women who are at least 18 years old. This approach will enable the research team to identify all women who are eligible for enrolling in the groups and allow the impact evaluation firm to collect basic descriptive data and a small number of outcome variables for each of the women's groups and households in the randomly selected villages. Next, we can select comparison wards that are similar in observable characteristics to the treatment wards using the matching approach described above. We aim to then sample women's groups, households, and women from each selected ward using the proposed three-stage matching approach.

Power to Detect Effects

We have also conducted power calculations to identify the number of individual women and groups that will be required for the impact evaluations. It is vital to have a sample size that is sufficiently large to detect small but meaningful effects of the intervention.

Power Calculations to Estimate Gendered Effects of the NFWP at the Individual and Household Levels. Power calculations at the individual level suggest that interviewing 1,170 treatment women and 4,680 comparison women across 18 treatment LGAs and 72 comparison LGAs will be sufficient to detect small but meaningful effects of the NFWP. We would have an 80% chance of detecting an intention-to-treat (ITT) effect of 0.21 standard deviations when we assume an intra-class correlation (ICC) of 0.089 for individuals clustered in LGAs and an R-squared of 0.25. This effect size is aligned with previous systematic reviews on the impact of self-help groups and vocational and business training on women's labor market outcomes (Brody et al., 2015; Chinen et al., 2017). The minimum detectable treatment effect of 0.21 standard deviations is equivalent to an impact of 8 percentage points on women's ownership of livestock (baseline mean = 20%). For these power calculations, the ICC is based on an asset index score we derived from the 2018 LSMS survey.

We would be able to detect similar or smaller effects (0.19 SD) for indicators related to women's control over income for which the ICC is 0.073 based on the 2018 LSMS Survey. This effect corresponds to a 9 percentage points increase from a mean value of 55 percent. The ICC for women's decision-making power over agricultural activities is 0.156, suggesting that we may need a larger sample size for detecting impacts on this variable with sufficient precision; with the current sample size we would be able to detect a minimum treatment effect of 0.27 standard deviations.

Power Calculations to Estimate Group-Level Effects. It is more challenging to come up with appropriate ICCs for group-level outcomes. The LSMS includes survey questions about membership in savings groups and women's groups, but the questions and associated data are not sufficiently detailed for the estimation of ICCs. For the purpose of the power calculations, we assume an ICC of 0.10. Power calculations at the group level suggest that a sample size of 270 treatment women's groups and 690 comparison women's groups across 18 treatment LGAs and 46 comparison LGAs will be sufficient to detect program impacts of 0.275 standard deviations with 80% power when we assume an ICC of 0.10 for groups clustered in LGAs.

Power Calculations without the state of Edo. While the WB plans to pilot the implementation of the NFWP in six states, currently the state of Edo is yet to confirm whether it will be implementing the program during the piloting stage. For this reason, we have also assessed the study's power to detect statistical differences at the individual and group levels without the state of Edo under two scenarios:

1. **Without** sampling additional comparison LGAs in the remaining five states, failing to compensate for the reduction in sample size.

Power calculations at the individual level suggest that interviewing 975 treatment women and 3,900 comparison women across 15 treatment LGAs and 60 comparison LGAs would give us an 80% chance of detecting an intention-to-treat (ITT) effect of 0.23 standard deviations - keeping the assumptions of an intra-class correlation (ICC) of 0.089 for individuals clustered in LGAs and an R-squared of 0.25.

Power calculations at the group level indicate that a sample size of 225 treatment women's group and 450 comparison women's groups across 15 treatment LGA and 30 comparison LGA would allow us to detect program impacts of 0.31 standard deviations with 80% power, assuming an ICC of 0.10 for groups clustered in LGAs.

The calculations above indicate that Edo not participating in the program moderately diminishes our ability to detect meaningful impacts. Standardized program impacts for any outcome would have to increase by 14% relative to a scenario that includes the State of Edo to maintain an 80% chance of detecting statistically significant NFWP effects, both at the individual level.

2. Sampling additional comparison LGAs in the remaining five states to compensate for the reduction in sample size.

At the individual level, power suggest that interviewing 975 treatment women and 3,900 comparison women across 15 treatment LGAs and 60 comparison LGAs would give us an 80% chance of detecting an intention-to-treat (ITT) effect of 0.23 standard deviations. This minimum detectable effect size is very close to the 0.21 SD we calculated for the scenario in which all six states are part of the pilot.

Keeping the ICC assumption constant, to reach the 0.21 SD minimum detectable effect size calculated for the scenario in which all six states are part of the study, we would need to interview 965 treatment women and 4,875 comparison women across 15 treatment LGA and 75 comparison LGA.

Power calculations at the group level indicate that a sample size of 210 treatment women's group and 840 comparison women's groups across 15 treatment LGA and 65 comparison LGA would allow us to detect program impacts of 0.29 standard deviations with 80% power, assuming an ICC of 0.10 for groups clustered in LGAs. This minimum detectable effect size is also close to the 0.275 SD we calculated for the scenario in which all six states implement the NFWP during the piloting stage.

Power Calculations under Low Take-Up (Imperfect Treatment Compliance). Imperfect treatment compliance would also significantly reduce statistical power. In particular, we consider

the hypothetical scenario in which just one third (33.3%) of the treatment group takes up the intervention and there is 100% compliance in the comparison group. Under such a situation, power calculations at the individual level suggest that the minimum detectable program effect size is 0.342 standard deviations, or 50% higher than in the scenario in which we have perfect compliance (i.e., 0.22 SD). Similarly, at the group level, a take-up rate of 33% would increase the minimum detectable effect size by 57% from 0.275 to 0.431 standard deviations, when we keep everything else constant.

We propose to increase statistical power by oversampling women who are more likely to participate in the NFWP, such as women's group members. Specifically, we propose a sample, in which 4 out of 5 respondent households includes a woman who is currently a member of a WAG or other women's group. We expect that by oversampling women for whom the take-up rate would be much higher and by involving community leaders in the program, take-up rates in our proposed sample would be no lower than 80%. Under such a scenario, power calculations at the individual level suggest that the minimum detectable program effect size is 0.24 standard deviations, or 8% higher than in the scenario in which we have perfect compliance (i.e., 0.21 SD). Similarly, at the group level, a take-up rate of one third would increase the minimum detectable effect size by 9% from 0.275 to 0.30 standard deviations, everything else constant. However, it remains important to also sample non-members to enable the estimation of intention-to-treat effects.

Power Calculations to Estimate Impacts of Social Norm Messages. We conducted power calculations for the RCT options to determine impacts of the social norm messages. Under option 1, we would collect data for 2,250 households in 10 wards per LGA during all data collection rounds. We would then be able to detect minimum effect sizes of around 0.22 SD with a likelihood of 80%; if the ICC is about 0.09, we would be able to detect effect sizes of 0.19 SD with a likelihood of 80% (see Table 1 below). This is when we assume an autocorrelation for social norm measures of 0.70. Under option 2 we would again collect data for 2,250 households in 10 wards per LGA during all data collection rounds. Under this scenario, we should be able to detect minimum effect sizes of around 0.2 SD If the ICC is 0.125; if the ICC is 0.09, we would be able to detect effect sizes of 0.18 SD with an 80% chance (see Table 2 below).

Table 1: Power Calculations with 5 States

MDES (in standard deviations)	Total sample size (individuals)	# of sampled wards per LGA	# of individuals sampled per ward	# of treatment LGA	# of control LGA	ICC	Autocorrelation Coefficient
0.28	1575	7	15	8	7	0.13	0.7
0.25	1800	10	12	8	7	0.125	0.7
0.22	2250	10	15	8	7	0.125	0.7
0.22	1800	10	12	8	7	0.09	0.7
0.19	2250	10	15	8	7	0.09	0.7

Table 2: Power Calculations with 6 States

MDES (in standard deviations)	Total sample size (individuals)	# of sampled wards per LGA	# of individuals sampled per ward	# of treatment LGA	# of control LGA	ICC	Autocorrelation Coefficient
0.23	1890	7	15	9	9	0.13	0.7
0.2	2250	10	12.5	9	9	0.125	0.7
0.18	2250	10	12.5	9	9	0.09	0.7

Power Calculations to Estimate Impacts of Additional Components. We also conducted preliminary power calculations for the proposed RCTs or quasi-experimental studies to estimate the impact of health and/or GBV layering that may be added to the NFWP at later stages. There are approximately 225 wards in the 18 treatment LGAs. Interviewing a sample of 930 women in 62 wards receiving the secondary treatment and 930 women in 62 control wards would allow us to detect a treatment effect of 0.2 standard deviation when using a conservative ICC of 0.15.

Outcome Measures

We plan to measure the impact of the NFWP approach in Nigeria using a mixed-methods approach. We will conduct surveys with women, men, and group representatives. We will use the measurement guide developed by the ECWG, led by AIR, to aid in designing the survey instruments. The Women's Empowerment and Economic Outcomes Measurement Guide section, which immediately follows, describes these instruments in more detail.

In addition, we will conduct a rigorous costing and cost-effectiveness analysis, including estimation of overall program costs and cost-effectiveness based on the guidelines developed by the ECWG. We provide more details on the costing and cost-effectiveness analysis in the Costing and Cost-Effectiveness Guidelines section.

Women's Empowerment and Economic Outcomes Measurement Guide

To fill a measurement gap, AIR led the development of guides for measuring women's empowerment, economic outcomes, and group-level outcomes under the ECWG. The guides seek to explicitly link the foundational work on how to measure women's empowerment and economic outcomes to theories of change that connect women's groups to these outcomes (de Hoop, Peterman, and Anderson, 2019). Although evidence on the impact of women's groups is growing, measurement challenges limit the ability to produce reliable and comparable estimates across impact evaluations. For example, studies often use inconsistent definitions of women's empowerment as well as different approaches to measure empowerment; evidence on the reliable measurement of consumption is scant and inconsistent; most impact evaluations estimate impacts on household-level asset ownership as opposed to ownership at the individual level; and group qualities that lead to success are difficult to quantify (e.g., Alkire et al., 2013; Beegle, De Weerdt, Friedman, & Gibson, 2012; Barooah et al., 2019). The measurement guides developed by the ECWG aim to address this measurement gap.

The measurement guides provide a collection of field-tested survey instruments and questions for measuring women's empowerment and economic outcomes (including consumption, poverty, assets, labor market participation or livelihoods, entrepreneurship, agriculture, and savings and credit) in quantitative impact evaluations and mixed-methods studies of women's groups, along with recommendations about ways to use these tools. For all outcome measures, AIR will provide quality control and ensure the technical soundness and contextual appropriateness of quantitative measures for these evaluations in Nigeria. For example, the evaluation team will measure women's income through self-reports of income from wage employment and self-employment. We will use the Women's Empowerment in Agriculture Index (WEAI) to develop measures of women's asset ownership, and questions from the Social Observatory of the World Bank to measure women's access to formal or informal credit and adapt national expenditure surveys to measure household expenditures. For the measurement of women's empowerment, we will include questions about decision-making regarding economic resources in the household based on the WEAI, and questions about self-efficacy from the New General Efficacy Scale. Table 3 lists key outcomes and tools used to guide the survey development. The complete document can be retrieved from https://womensgroupevidence.org/sites/default/files/2020-03/Guide-for-Measuring-Women-s-Empowerment-and-Economic-Outcomes-in-Impact-Evaluations-of-Women-s-Groups.pdf.

Table 3. Examples of Outcome Measures and Survey Tools

Outcome measure	Survey tool
Women's income	Self-reported income from wage employment and self-employment
Women's and men's asset ownership	Women's Empowerment in Agriculture Index

Women's financial inclusion	Social Observatory Questionnaires
Household expenditures	National expenditure surveys
Women's economic empowerment	Women's Empowerment in Agriculture Index
Women's psychological empowerment	New General Self-Efficacy Scale
Social norms	Tools based on validated social norms measurement adapted to Nigerian context

Costing and Cost-Effectiveness Guidelines

To facilitate a comparison of a project's components, outcomes, and locations, it is critical to go beyond impact measures and compute program costs and cost-effectiveness. However, research on the costs and cost-effectiveness of women's groups is scarce, possibly because of the lack of consistent cost data linked to different activities and outcomes.

Considering the scant evidence on the costs and cost-effectiveness of women's groups, the ECWG prepared guidelines for collecting cost data on programs that focus on women's groups, including the layering of interventions. These guidelines offer key methods, suggestions, and tools to support cost data collection, including suggestions on ways to use the data to conduct an economic evaluation that compares program costs to program benefits and ways to use the data to guide resource allocation decisions. The guidelines can be found at https://womensgroupevidence.org/our-work/guidance-for-cost-effectiveness. Using these guidelines and drawing on our experience with cost-effectiveness analysis (CEA) and cost-benefit analysis (CBA), AIR will collect data on costs throughout the evaluation and carry out the analysis following the guidelines.

Cost-Effectiveness Analysis

The first step in conducting CEA estimation will be to conduct a cost analysis of WAGs and then combine cost measures from this analysis with aligned WAGs' outcomes and effect sizes. We will produce cost-effectiveness ratios (per-dollar impact on a given outcome) for most standardized empowerment indices and economic outcomes to enable a direct comparison against other programs. When conducting the cost analysis, we will need to determine the appropriate perspective for economic evaluation. We assume that costs to all stakeholders (including group members) are relevant. Group members incur costs, in terms of time spent in group activities (e.g., on meetings or trainings), and real costs (e.g., costs related to travel). We will incorporate questions about costs incurred and time spent by group members into household surveys. Relatedly, we will work with the World Bank, Nigeria, to collect costs from the implementers' perspective, including accounting costs as well as costs of resources not directly paid for by the program (e.g., volunteer time or shared resources).

In addition to defining a cost perspective, a critical issue concerning "construct validity" arises from the difference between overall costs and marginal costs. The general approach of a cost

evaluation is to include all labor, capital, and ongoing costs in the model. Some of the cost elements are fixed costs (for example, costs of setting up a monitoring system or physical infrastructure), which are not dependent on the activities of the program's scale of operations, at least in the short-run. Variable costs depend on program activities and can vary significantly over time and scale (Siwach et al., 2020). We will draw on data from resources with both fixed and variable costs from multiple sources to ensure that the total overall costs are considered. This approach is especially important when considering scaling up WAGs or applying the WAG model in a new state. Scaling up WAGs focuses on expanding a program within the same target group, adapting a program to suit a different target group, or collaborating with a different implementing agency, such as the government. To inform scalability decisions, these complexities should be added into the CEA, especially because cost estimates can be misleading if they are based on pilot programs. So-called "pilot bias" may work through a number of channels, leading to over- or underestimation of results (Evans & Popova, 2014). A scaled-up program will benefit from higher economies of scale, thus reducing costs, but it may also require more resources as the scope of the program broadens. To understand the scaled-up cost-effectiveness, we will analyze how these costs scale up when the target number increases.

We will combine estimated program costs with impact estimates to generate the overall cost-effectiveness of the program. We plan to estimate two measures of program cost-effectiveness – (1) the Cost-effectiveness Ratio (CER), which measures the relative efficacy of the program by looking at costs per unit of impact achieved for one outcome; and the (2) Benefit-to-Cost Ratio (BCR), which measures the return on investment by looking at multiple outcomes simultaneously, monetizing those outcomes, and generating a common financial unit of "benefits." The CER is useful for conducting a comparative analysis, because the ratio can be compared against other programs that intend to move the same outcomes. Additionally, CERs are estimated separately for different outcomes, allowing different stakeholders to consider the cost-effectiveness of outcomes that they place highest weight on. The analysis also will explore options for translating women's empowerment into disability-adjusted life years (DALYs) based on Calvi (2020), who examined the ways in which women's empowerment can result in reductions in maternal mortality. This allows for a more direct comparison with cost per DALY averted of other programs, given the widespread use of this measure. Finally, in addition to the overall cost-effectiveness of the program, we will estimate incremental cost-effectiveness of additive program components, such as messages related to social norms, layering of health interventions, and livelihoods grants.

Despite their widespread use, CERs are estimated for specific outcomes and therefore do not represent the holistic program cost-effectiveness when multiple outcomes hold value. Therefore, we will also estimate BCRs, given the range of outcomes considered in the impact evaluation—such as women's income, assets ownership, and decision-making power—the

value of non-economic outcomes may be considered relatively subjective, and we will estimate these through a stated preference approach where stakeholders are asked to state their willingness to pay to achieve the impact they receive.

Before estimating the ratios (CER or BCR), we will adjust all costs and benefits for exchange rates, inflation rates, and time value. We will use the Market Exchange Rates to convert all local prices to U.S. dollars, and adjust for inflation using the Consumer Price Index approach. After estimating the costs and benefits in standardized currency and year, they will be adjusted to base-year prices to reflect different time preferences using a discount rate. Studies in the international development literature generally use the social opportunity cost of capital to determine discount rates, and recent guidelines have suggested that 10 percent is considered as a reasonable rate for discounting (Dhaliwal et al., 2013). Our analysis will leverage the average discount rates used in cost-effectiveness ratios of similar programs that will be considered for benchmarking our findings, but we will also present sensitivity analyses with varying discount rates ranging from 5 to 10 percent.

Analysis

AIR will assess the impact of the WAG model using a DID regression framework while controlling for time-varying individual and group characteristics. We will compare changes in outcomes over time between the treatment and comparison group. DID entails calculating the change in outcomes, such as women's income, between the baseline survey and midline or endline survey for treatment and comparison group units and comparing the magnitude of these changes between the treatment and the comparison groups. In this study, the DID model will use individual-, household-, and group-level data from before and after the NFWP's components are implemented to compare the total change in income, asset ownership, household expenditures, and economic empowerment for women participating in NFWP with the total change in the same outcomes for women in LGAs in which NFWP is not implemented. The key assumption underpinning the DID approach is that there is no systemic, unobserved, time-varying difference between the treatment and comparison groups.

Two key features of the DID estimator are particularly attractive for deriving unbiased program impacts. First, using pre- and post-treatment measures enables us to "difference" out unmeasured fixed (i.e., time-invariant) household and group characteristics that may affect outcomes, such as education level and household composition. The approach also enables us to benchmark the change in the indicator against its value in the absence of treatment. Second, using the change in a comparison group as a counterfactual enables us to account for general trends in the value of the outcome. We will use cluster-robust standard errors to account for a lack of independence across observations due to clustering of households when the intervention is allocated at a level of

aggregation above the household level, such as LGAs. For secondary outcomes, we will apply corrections for multiple comparisons to multiple outcome measures within the same outcome domain using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

The DID design will provide us with the intent-to-treat (ITT) effect of the WAG model; in other words, the average treatment effect for those women or groups assigned to a treatment condition regardless of take-up of treatment. Results from the intervention-monitoring evaluation will inform the fidelity of implementation of the program and will allow us to determine if there were discrepancies in adherence to treatment conditions. If so, we will further assess the effect of treatment on the treated; that is, we will evaluate the effect of the program in those LGAs assigned to treatment that actually implement the treatment as planned (i.e., full fidelity of implementation).

To estimate the effect of specific layering interventions and social norm messages in addition to the WAG model, AIR anticipates using an analysis of covariance (ANCOVA) empirical approach using both baseline and endline data if an RCT design is selected. The ANCOVA approach uses a regression specification that includes the baseline measures of outcome variables as an additional explanatory variable. This empirical approach can improve statistical power by exploiting information and variation contained in the baseline data (McKenzie, 2012). In other words, the use of ANCOVA increases the likelihood that statistically significant effects will be detected if the program indeed causes statistically significant effects. In our regression, we will use cluster-robust standard errors to account for a lack of independence across observations due to clustering of women in wards. For secondary outcomes, we will apply corrections for multiple comparisons to multiple outcome measures within the same outcome domain using the Benjamini-Hochberg method.

We will also create subgroups to assess heterogeneous effects. For example, we will examine heterogeneous effects for women in different age categories, and women of different ethnic groups, as well as different regions. In addition, we will examine heterogeneous effects for various moderators that we identified in theory of change. These moderators include women and households with different levels of asset ownership, different social norms, and different time constraints at baseline. While we will examine these heterogeneous effects, we recognize that we may not have the statistical power to detect these heterogeneities with sufficient precision unless we significantly increase the sample size. In addition, we will apply corrections for multiple comparisons in the estimation of the standard errors for heterogeneous effects to limit the risk of fishing for statistical significance.

Analysis of Potential Spillover Effects

AIR's approach to the impact evaluation is to focus on the ITT effect, which is a weighted average of direct effects for women's group members and spillover effects for nonmembers. We will select a random sample of households with adult women in each selected ward, based on the listing survey and before the start of the baseline survey. This sample of households will include members and nonmembers of WAGs. It is also possible that some women will be WAG members at some point and nonmembers at another point. While we anticipate oversampling households with women who are more likely to participate in WAGs, the inclusion of nonmembers should allow us to measure spillover effects.

To achieve this objective, we will compare nonmembers in the treatment LGAs with a matched group of nonmembers in the comparison LGAs. Such an approach would enable the identification of spillover effects in the absence of differences in unobservable characteristics. However, it is very likely that unobservable characteristics will bias the estimates of spillover effects identified through such an approach because of self-selection in the program.

For this reason, we will use an alternative mixed-methods approach to identify spillover effects. First, we will explore the option of using social network analysis to identify spillover effects. For example, we will ask women about their friends and other contacts in the community and whether they are members of women's groups. These data will allow us to map social networks and may enable us to link the responses of women's group members during baseline to the likelihood of their friends' participation in WAGs during future data collection rounds. In addition, we will ask survey questions about how women heard about WAGs, and about their reasons for joining or not joining WAGs. The inclusion of such questions and analyses will not allow us to fully assess spillover effects. However, they will provide some evidence of the presence or absence of spillover effects.

Such analyses can be particularly powerful when combined with qualitative data on interactions between members and nonmembers. Mixed-methods analyses will help in shaping the narrative around possible positive spillover effects of the program. Both KIIs and FGDs can probe members and nonmembers about their interactions with community members outside of the WAGs; the information they share with their friends about their participation in WAGs, and the benefits of that participation; and their perceptions of spillovers into their communities.

Qualitative Study Design

We will combine the impact evaluations with rigorous qualitative research, which is comprised of two main components: 1) A formative assessment conducted at the baseline phase; and 2) a process evaluation conducted at the midline phase. Moreover, we will periodically utilize rapid qualitative assessment (RQA) techniques throughout the study to probe and gain insight on

emerging practices, trends, and dynamics that warrant a deeper inquiry. For example, we used RQA interviewing techniques prior to the start of the baseline phase in order to learn more about the social norm interventions planned as part of the NFWP. These qualitative insights informed our development of measurement instruments for the impact evaluation. RQAs will enable us to be flexible and iterative in our methodological approach throughout the course of the study.

Formative Assessment

The objectives of the formative assessment are as follows:

1. Gain a deeper understanding of the local policies, social norms, expectations, and other contextual dynamics that shape women's experiences and perceptions of women's groups, including WAGs.

2. Collect and analyze data that will help inform and refine the impact evaluations' theory of change, measurement instruments, and overall design.

We will collect data for the formative assessment during the baseline phase. We propose three main data collection methods for the formative assessment:

1. Focus group discussions (FGDs) with women's group members, former and non-women's group members, and spouses of women who are in women's groups. The FGDs will include a participatory livelihoods assessment (PPA).

2. In-depth interviews (IDIs) with women's group members and non-WAG women in the community, including a financial diaries exercise.
3. Key informant interviews (KIIs) with staff from the World Bank as well as federal, state, LGA-level, and ward level officials.

Sampling Approach

We will sample from all 6 states where the program is implemented—Abia, Edo, Kebbi, Niger, Ogun, and Taraba. Further, we will draw the qualitative sample from the treatment and comparison groups in the quantitative sample. Taking this approach will strengthen the mixed-methods design and enable researchers to triangulate data from the same pool of respondents. In each state, we will purposively select 2 LGAs and 1 ward per LGA. In total, we will collect qualitative data from **12 wards across the 6 states** (see Table 4 below)**.** The selection criteria will aim to maximize the variation of women's groups with different components. Another potential sampling criterion is levels of social organization in LGAs and wards. For instance, within a state, we may choose to select LGAs and wards in which women were already organized in a variety of groups before the start of the NFWP as well as LGAs and wards where women have fewer group organization options or where there are fewer existing women's

groups. We will determine the final selection of states, LGAs, and wards in consultation with the World Bank.

Table 4. FGD and IDI Sampling for the Formative Assessment

Respondents	Formative Research Sampling
Ward level (12 selected wards)	
Women's Group members	• 12 FGDs (1 FGD in each of the 12 wards) --PPA component
	• 12 IDIs (1 IDI in each of the 12 wards) --Financial diaries
Former and non-women's group women in the community	• 12 FGDs (1 FGD in each of the 12 wards) --PPA component
	• 12 IDIs (1 IDI in each of the 12 wards) --Financial diaries
Spouses of women's group members	• 12 FGDs (1 FGD in each of the 12 wards)
Total number of formative FGDs and IDIs	**36 FGDs and 24 IDIs**

Focus Group Discussions

We will conduct 36 FGDs in the 12 selected wards (3 FGDs per ward). Focus group research involves guiding a diverse group of participants through a discussion on a particular topic. This qualitative method is well suited for obtaining diverse perspectives on particular issues and offers the possibility of observing intra-group dynamics and norms during the discussion (Morgan, 1996). We believe this method is best suited for investigating women's perspectives because it allows us to go beyond individual experiences and gain understanding of the community-level norms that influence women's participation in WAGs as well as broader attitudes toward women's empowerment. Generally, focus groups include five to eight participants, and the participants are guided through various discussion topics by a trained facilitator. We will conduct 12 FGDs with women's group members, 12 FGDs with former and non-women's group women in the community (i.e., comparison group participants), and 12 FGDs with spouses of women's group members (see Table 2 above). FGDs with women's group members will shed light on members' experiences with women's groups before the start of the program, gendered social norms, and the livelihoods of women's group members. FGDs with other women in the community will help us gain knowledge broader knowledge of social norms pertaining to gender and local livelihoods as well as barriers to entry to women's groups. FGDs with the spouses of women's group members will shed light on these topics from men's perspectives as well as men's attitudes about women's group participation.

As part of the FGDs, we will conduct **participatory livelihood assessments** using a social mapping exercise. Social mapping (Mikkelsen, 2005) is a participatory tool designed to involve community members— the subjects of the research—in the research process as active agents and stakeholders and not just respondents. Approaches such as participatory mapping offer unique insights by involving research subjects in the elaboration and definition of categories and

interpretations. We will use this approach to collect data on local perceptions of gender and poverty, access to services and resources, participants' perceptions of their economic and social situation (as well as that of their community), and available livelihood options. Social mapping at the formative stage will help contextualize FGD findings on social norms and experiences with women's groups and will allow us to explore local definitions of women's empowerment in more depth.

In-Depth Interviews

Due to the sensitive nature of topics such as gendered household dynamics, we will conduct one-on-one IDIs with 24 women participants (see Table 2 above). From the FGD sample, we will select 12 women's group members and 12 women who used to be or never have been women's group members to participate in the IDIs (24 total). These IDIs will help us gain in-depth knowledge about social norms related to gender, intra-household gender dynamics, and household decision-making in a private setting where women may feel more comfortable discussing these topics.

As part of the IDIs, we will conduct **a financial diary exercise**. The evaluation team will implement a streamlined financial diaries approach based on the technique developed by Collins, Morduch, Rutherford, and Ruthven (2010). The financial diaries method allows researchers to understand household-level income flows and expenditures over time using simplified income statements (see Table 5). Such diary studies can help researchers gain a better understanding of the processes that take place in the context of women's groups. Researchers, for example, may conduct analyses on the topics discussed during women's group meetings or on the formal and informal ways in which women and other household members use savings, credit, and livelihoods to mitigate shocks (Collins et al., 2010).

Table 5. Sample Household Income Statement

Fixed monthly income		**Fixed monthly expenditure**	
Non-agricultural employment income (men)	₦	Housing (e.g., rent, mortgage)	₦
Non-agricultural employment income (women)	₦	Children's education	₦
Agricultural income (if applicable)	₦	Interest payments on formal loans	₦
State grants	₦	Interest payments on informal loans	₦
Formal loans	₦	Savings	₦
Other (e.g., separate business accounts)	₦	Contributions to women's group	₦
Total:	₦	**Total:**	
Variable weekly income		**Variable weekly expenditure**	
Informal loans (e.g., borrowed from friends, relatives, money lenders, adashi, esusus, ajo, etc.)	₦	Food/groceries	₦
Gifts	₦	Business-related expenses	₦
Remittances received	₦	Health-related expenses	₦

	₦	Remittances sent	₦
	₦	Airtime/data related costs	₦
	₦	Gifts	₦
	₦	Transportation	₦
	₦	Entertainment (e.g., dining, alcohol, shows, etc.)	₦
Total:	₦	**Total:**	₦

At baseline, moderators will fill out household income statements together with IDI respondents. Respondents will then be trained to fill out one income statement per week for the following three weeks for a total duration of 4 weeks. This exercise will be repeated at midline. Moderators will contact participants weekly to help them fill out their weekly income statements and will provide airtime credits as an incentive as well as to accommodate follow-up sessions over the telephone. At the conclusion of the 1-month period, moderators will collect all weekly income statements in their physical form, if feasible. If not feasible, moderators will arrange to receive the income statement inputs over the telephone. In sum, a single respondent will be asked to fill out 8 weekly balance sheets (4 at baseline and 4 at midline).

Key Informant Interviews

We will conduct 48 KIIs with stakeholders involved in the design and implementation of the NFWP as well as local community leaders (see Table 6 below). For our purposes, a key informant is a person who possesses expert knowledge about the NFWP or about a region where the program is being implemented. One-on-one interviews with key informants provide an ideal forum for engaging with people who possess expert knowledge about a program, including service providers who can offer insights about how the program was designed and how it interacts with other services.

Table 6. Formative Assessment KII Sample

Respondents	Formative Phase
National level	
World Bank staff	2 KIIs
Federal Ministry of Women's Affairs officials	2 KIIs
Federal Project Coordinating Units officials (FPCUs)	2 KIIs
State level (6 states)	
State Project Coordinating Units officials (SPCUs)	6 KIIs
State-level Ministry of Women's Affairs officials	6 KIIs
Institutional Capacity Building Advisors (ICBAs)	6 KIIs
LGA level (12 LGAs)	
LGA Field Supervisors	12 KIIs
Ward level (12 wards)	
Ward facilitators	12 KIIs

Respondents	Formative Phase
Total:	48 KIIs

Process Evaluation

We propose a mixed-methods process evaluation to assess implementation fidelity and beneficiary experiences. Process evaluations are important to understanding the why and how of a program and help with interpretation of other evaluation results (Oakley, Strange, Bonell, Allen, & Stephenson, 2006). In conjunction with impact evaluations, process evaluations can help ascertain whether a program is ineffective because of its underlying theory or simply because its delivery was of low quality (Rychetnik, Frommer, Hawe, & Shiell, 2002). The process evaluation component will occur at midline and will focus on how the program was implemented, including to what extent program activities were implemented as intended and how beneficiaries experienced the program. Further, the process evaluation will explore external and contextual factors that influence program implementation. We will triangulate qualitative process evaluation findings with quantitative data on implementation fidelity (which we will obtain from data collected at the group level and the intervention monitoring) to shed light on the mechanisms that may influence program uptake. For the qualitative process evaluation, we will collect data in the same 12 wards that were selected for the formative assessment, as described above.

We propose the following methods for the process evaluation:

1. FGDs with WAG members, former and non-WAG members, and spouses of women who are in WAGs. The FGDs will include a participatory livelihoods assessment.

2. In-depth interviews (IDIs) with women's group members and former and non-WAG women in the community, including a financial diaries exercise.

3. Key informant interviews (KIIs) with staff from the World Bank as well as federal, state, LGA-level, and ward level officials.

Sampling Approach

We will follow the same sampling approach as we did at baseline, that is, we will sample from all 6 states where the program is implemented and randomly draw respondents from the treatment and comparison groups from the quantitative sample (See Table 7 below).

Table 7. Process Evaluation Sample

Respondents	Formative Research Sampling
Ward level (12 selected wards)	
WAG members	• 12 FGDs (1 FGD in each of the 12 wards) --PPA component
	• 12 IDIs (1 IDI in each of the 12 wards)

Respondents	Formative Research Sampling
	--Financial diaries
Former and non-WAG women in the community	• 12 FGDs (1 FGD in each of the 12 wards) --PPA component
	• 12 IDIs (1 IDI in each of the 12 wards) --Financial diaries
Spouses of WAG members	• 12 FGDs (1 FGD in each of the 12 wards)
Total number of process evaluation FGDs and IDIs	**36 FGDs and 24 IDIs**

Focus Group Discussions

For the process evaluation, we will conduct a second round of FGDs with WAG members (12 FGDs, 1 per ward), with former and non-WAG women in the same community (12 FGDs, 1 per ward) and spouses of WAG members (12 FGDs, 1 per ward). At the process evaluation stage, FGDs with WAG members will focus on women's experiences of program implementation, perceived challenges in women's group functioning, and perceptions of changes in women's livelihoods or empowerment in relation to women's groups participation. FGDs with spouses of WAG members will focus on potential changes in men's attitudes about women's participation in WAGs and perceived impacts of the program (see Table 5 above).

In-Depth Interviews

The process evaluation will include a second round of in-depth interviews with the same WAG members (12 IDIs, 1 per selected ward) and former and non-WAG members (12 IDIs, 1 per selected ward) we interviewed at baseline. IDIs with WAG members at this stage will focus on women's experiences of the program, intra-household gender dynamics, household decision-making and finances, and perceived changes in women's empowerment. IDIs with WAG members will also include a second round of financial diaries, with instruments that will be tailored based on findings from the formative assessment about how women think about and spend money in selected study areas. Financial diaries with WAG members at the process evaluation stage will provide a window into how women are using the funds acquired through women's group participation and how this interacts with household decision-making. IDIs with former and non-WAG members will investigate reasons for leaving women's groups, perceived barriers or challenges to retention of WAG members, and consequences for women's livelihoods (see Table 5 above).

Key Informant Interviews

At the process evaluation stage, we will conduct a second round of 61 KIIs with NFWP stakeholders and community leaders (see Table 8 below). KIIs with NFWP stakeholders at the process evaluation stage will focus on how the program was implemented, to what extent it was implemented as intended, and how external or internal factors may have influenced

implementation. These KIIs will also explore perceived impacts of the program. KIIs with community leaders at the process evaluation stage will investigate perceived program impacts, including perceptions of change in social norms, women's livelihoods, and women's empowerment in selected wards.

Table 8. Process Evaluation KII Sampling

Respondents	Formative Phase
National level	
World Bank staff	2 KIIs
Federal Ministry of Women's Affairs officials	2 KIIs
FPCUs including PMIS staff	3 KIIs
State level (6 states)	
State Project Coordinating Units officials (SPCUs)	6 KIIs
State-level Ministry of Women's Affairs officials	6 KIIs
Institutional Capacity Building Advisors (ICBAs)	6 KIIs
LGA level (12 LGAs)	
LGA Field Supervisors	12 KIIs
Ward level (12 wards)	
Ward facilitators	12 KIIs
Barefoot Business Councilors (BBCs)	12 KIIs
Total:	61 KIIs

Analysis

We will code and analyze all data from KIIs and FGDs using the NVivo qualitative software program. We will create a preliminary coding structure based on the research questions, KII and FGD protocols, and memos of ideas that emerge during data collection. We will use this coding outline to organize and subsequently analyze the information gathered through KIIs and FGDs. The outline will be a living document and may be modified as new themes and findings emerge during data analysis. A list of definitions for the codes will accompany the outline, so that coders categorize data using the same standards. After inputting the raw data into NVivo, coders will select a sample of interviews to double-code to ensure interrater reliability. The team will then input the data into the thematic structure.

Data Collection

Field Staff Training. The integrity of any study rests on the accuracy and reliability of the data that are collected. Achieving this objective requires the recruitment of experienced field staff. We will supervise the data collection firm in recruiting people who have previously conducted household surveys with a focus on sensitive gender topics and have used tablets or other electronic devices for data collection. We will recruit a larger number of potential enumerators and moderators than required to enable the selection of field staff who perform better during

the training and in the final evaluation of performance during the training. We will also ensure that the data collection firm has access to a reserve of enumerators (for the quantitative data collection) and moderators (for the qualitative data collection) so we can select the most skilled for the actual data collection.

Next, it will be essential to thoroughly train field staff in applying the data collection materials and implementing quality control procedures for assessments conducted in the field. AIR will work with the data collection firm in providing training to the enumerators, supervisors, and other data collection staff, with support from our national consultants in Nigeria. The training will take place over two weeks and will consist of one week of GBV-specific training for female enumerators and supervisors only and one additional week for the full team. The GBV training will begin with a discussion of gender and GBV to sensitize the enumerators to administering surveys related to such topics. We will also provide training on best practices for administering GBV surveys, the referral process for linking respondents with support services, and enumerator self-care strategies as administering such surveys can be traumatizing for survey administrators as well as respondents. Lastly, the GBV-specific training will focus on reinforcing best practices and sensitivity through extensive role playing. The full team training will begin with a discussion of the theory underpinning the questionnaire, followed by a discussion of the questions in the questionnaire to ensure a complete understanding of each question's goal. During this phase of the training, we will also review survey protocols, roles and responsibilities, the use of electronic devices for data collection, and techniques to interview responders. The trainees will then practice what they have learned in role-play exercises, demonstrations, and other exercises that illustrate typical cases they may face in their field work. The classroom practice will be complemented with a field practice where trainees visit households and conduct the survey. After that, we will evaluate each trainee's performance in the field practice and their understanding of the materials and questions in order to assess whether more classroom or field practices are required, and to identify those who perform better and will be part of the field staff. We will make the training interactive by including various quizzes that will help to assess how well enumerators understand the survey materials.

Due to COVID-19, AIR staff will join the online, but trainings will happen in-person (but adapted to COVID-19) in Nigeria. The logistics of the training during COVID-19 requires working with smaller groups of enumerators, recording the sessions and sharing them with AIR staff to address potential connectivity issues (also allowing enumerators to review the sessions multiple times), and allocating more time to account for the possibility of connection issues of AIR staff or any additional steps that may be necessary.

AIR will use a training-of-trainers approach such that AIR will remotely train the consultants and senior staff from the data collection firm, who will then train the enumerators in person and who will oversee a pilot of the survey instruments before the start of the baseline survey. Lessons from the training and the pilot survey can then be used to modify the survey instruments and achieve reliability and fidelity of implementation during data collection. The consultants will oversee the implementation of the data collection training by senior staff from the data collection firm, with the remote presence of AIR staff.

Data Collection Oversight. National consultants will provide quality control and ensure the technical soundness of data collection and entry, with oversight by AIR staff. We will generate several protocols to achieve this. For example, data collection supervisors should ensure that all surveys and materials have been completed correctly before leaving each evaluation site. We will use SurveyCTO®, an Open Data Kit–based platform that allows for data encryption and a high level of security. Data can be downloaded from the server daily and deleted from the tablets used for data collection.

For qualitative data, AIR will use a systematic and efficient process for organizing and analyzing qualitative data. This process includes audio recording all interviews in the local language, transcribing them in the local language, and then translating them into English. All analysis will be performed with deidentified data. AIR will ensure that data are handled according to procedures and protocols dictated by the AIR IRB.

Communication and Dissemination Plan

We recognize that although the NFWP is focused on six pilot areas within Nigeria, the project is ultimately intended to scale nationally, and its impacts are informative for global women's collective programming. We understand that it is crucial to involve key stakeholders, such as the federal and state ministries of women's affairs and social development, to ensure the evaluation's findings have maximum impact on policy development. To this end, we will work with the World Bank and the NFWP team to develop a framework for engaging key stakeholders throughout the evaluation process—first in cocreation, then in implementation, and finally in dissemination. Using this framework from the beginning of the impact evaluation will allow us to coordinate with stakeholders comprehensively and ensure that engagement occurs at the right time and in the right way, and that it is focused on producing the right results. During the inception phase, as part of the co-creative element of refining the evaluation design, we will revise the engagement and communication plan using a stakeholder map.

Throughout the evaluation of the NFWP, we will work with the NFWP team, including the WB, FPCU, and SPCU, to engage with program implementers and key stakeholders to continuously

disseminate results. We will create a dissemination plan that includes at least two policy briefs, two blogs, and an article for peer-reviewed journals. Further, we will present our findings to the World Bank and BMGF teams, as well as at academic and policy conferences and/or webinars.

Last, under the ECWG, AIR has developed a website that serves as the global repository for evidence on women's groups for the BMGF. Leveraging our position on the ECWG, we will regularly disseminate findings from the evaluation through the ECWG website on www.womensgroupevidence.org. This website creates opportunities for the joint research team to generate blogs and policy briefs about the evaluation findings and disseminate those resources to a wide audience through the ECWG newsletter.

Work Plan

AIR developed a work plan based on the terms of reference and the team's experience conducting similar studies. Work on managing the impact evaluation will commence in January 2021 and will continue through September 2023. We propose a timeline that incorporates refining the theory of change, the evaluation design, instruments, methods and the workplan, training, assistance with data collection, and report writing for both the baseline, midline, and endline evaluation rounds. Annex 3 shows the Gantt chart with the full work plan, and Table 10 presents the planned deliverables and due dates resulting from the work plan.

The work has three main phases: (1) inception, (2) evaluation, and (3) dissemination. The inception phase focuses on refining tools and instruments and communicating with the data collection firm about the design. The second phase focuses on the implementation of the impact evaluation, including the three rounds of data collection and analysis. As discussed above, we may also include additional rounds of data collection to ensure sufficient statistical power for detecting effects of the social norms intervention. The final phase focuses on disseminating the results of the impact evaluation, both within Nigeria and to the broader community of researchers, implementers, and policymakers focused on women's groups. A detailed description of each phase after the inception phase is provided in the remainder of this section.

Phase 2: Evaluation

Baseline. The first phase of the evaluation will focus on baseline data collection and analysis. We will work in collaboration with the data collection firm and key stakeholders to refine indicators, improve the measurement plan, revise the mixed-methods plan, develop a sampling approach, and verify power calculations in case new data are available. After receiving IRB approval and the World Bank's approval of the inception report, the data collection firm will conduct a short census (the listing survey) in each of the selected wards to identify all women's groups and households with women who are at least 18 years old. Next, we can select wards in

comparison LGAs that are similar in observable characteristics to the wards in treatment LGAs using a matching approach. Following the selection of wards, AIR researchers—together with the two national consultants—will train enumerators from the data collection firm contracted by the World Bank on all quantitative and qualitative data collection instruments. As COVID-19 prevents international travel, the AIR team will join the training virtually, closely collaborating with the national consultants. Fully trained teams of data collectors will then be deployed to evaluation states to commence data collection. The national consultants will also play a role in overseeing data collection efforts. Data entry and transcription will take place concurrently with data collection, when possible. During this time, AIR researchers will conduct periodic backchecks of data quality to ensure the rigor and validity of the data being collected. Upon completion of the baseline data collection, the AIR team will conduct its analysis of both quantitative and qualitative data. The team will submit the draft baseline report and share the initial findings at a meeting with stakeholders in August 2021. Shortly after, the team will submit the revised final report incorporating feedback.

Midline and Monitoring Process. A year after the program's baseline data collection in April 2022, we will conduct the midline data collection, including the collection of monitoring evaluation data to examine the fidelity of NFWP implementation and identify any facilitators or inhibitors of the realization of the program's intended outcomes. The AIR team will develop tools for the monitoring evaluation; refine survey tools from baseline; train local enumerators; and oversee data collection, data analysis, and report writing, similar to the baseline data collection. The team will submit a draft midline report and share initial findings with stakeholders in July 2022, submitting final documents shortly thereafter. The monitoring data findings will inform course correction for program implementers and will be used to refine instruments for the endline evaluation.

Endline. Endline data collection for the final evaluation will take place at the beginning of 2023. This activity will involve revising and updating evaluation instruments and refining any tools to collect cost data as required. The subsequent analysis and reporting will require a similar series of steps and amount of time as the baseline and midline segments, culminating in the finalization of the endline report and several outputs by July 2023. The draft final report will be submitted by May 2023, using dissemination meetings to gather feedback.

Phase 3: Dissemination

AIR will engage with stakeholders throughout the duration of the study. Following each main data collection and reporting activity, the team will develop policy briefs and blog posts to post on social media as well as the ECWG website. AIR will share the evaluation results with key stakeholders in Nigeria and abroad by presenting in Nigeria and at international conferences, such as the What Works Global Summit and the Association for Public Policy and Management's annual conference. Further, AIR will host webinars for the World Bank and

BMGF teams to provide additional dissemination activities. Dissemination efforts will continue during 2023.

Timeline of Deliverables

Table 10. Timeline of Tasks and Deliverables

Deliverable	Due date
Contract signing	Jan 2021
Finalizing design and draft inception report	**19 Apr 2021**
Baseline data collection	May 2021
Baseline analysis	Jun-Jul 2021
Baseline report	**15 Aug 2021**
Midline data collection	May-July 2022
Midline analysis	Aug-Sept 2022
Monitoring process analysis	Aug-Sept 2022
Midline report	**15 Oct 2022**
Endline data collection	May-July 2023
Cost data analysis	Aug-Sept 2023
Endline analysis	Aug-Sept 2023
Endline report	**15 Oct 2023**
Dissemination (incl. additional products, e.g., presentation, blogs)	**Oct-Dec 2023**

Note. Deliverables are indicated in **bold**.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*, *72*(1), 1–19.

Alkire, S., Meinzen-Dick, R., Peterman, A., Quisumbing, A., Seymour, G., & Vaz, A. (2013). The women's empowerment in agriculture index. *World Development, 52*, 71–91.

Alvi, M., Gupta, S., Meinzen-Dick, R., & Ringler, C. (2020). *Phone surveys to understand gendered impacts of COVID-19: A cautionary note.* Retrieved from https://pim.cgiar.org/2020/07/14/phone-surveys-to-understand-gendered-impacts-of-covid-19-a-cautionary-note/

Anderson, C., de Hoop, T., Desai, S., Siwach, G., Meysonnat, A., Gupta, R., Singh, R. S. (2019). *Investing in women's groups: A portfolio evaluation of the Bill & Melinda Gates Foundation's investments in South Asia and Africa* (Research Brief). Retrieved from http://www.womensgroupevidence.org

Barooah, B., Chinoy, S. L., Dubey, P., Sarkar, R., Bagai, A., & Rathinam, F. (2019). *Improving and sustaining livelihoods through group-based interventions: Mapping the evidence* (Evidence Gap Map Report 13). New Delhi, India: The International Initiative for Impact Evaluation (3ie).

Beegle, K., De Weerdt, J., Friedman, J., & Gibson, J. (2012). Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics, 98*, 3–18.

Benebo, F. O., Schumann, B., & Vaezghasemi, M. (2018). Intimate partner violence against women in Nigeria: A multilevel study investigating the effect of women's status and community norms. *BMC Women's Health, 18*(136). Retrieved from https://doi.org/10.1186/s12905-018-0628-7

Brody, C., de Hoop, T., Vojtkova, M., Warnock, R., Dunbar, M., Murthy, P., & Dworkin, S. L. (2015). Economic self-help group programs for improving women's empowerment: A systematic review. *Campbell Systematic Reviews, 11*(1), 1–182.

Calvi, R. (2020). Why are older women missing in India? The age profile of bargaining power and poverty. *Journal of Political Economy, 128*(7), 2453–2501.

Collins, D., Morduch, J., Rutherford, S., & Ruthven, O. (2010). *Portfolios of the poor: How the world's poor live on $2 a day.* Princeton, NJ: Princeton University Press.

de Hoop, T., Peterman, A., & Anderson, L. (2019). *Guide for measuring women's empowerment and economic outcomes in impact evaluations of women's groups.* Retrieved from http:///www.womensgroupevidence.org

de Hoop, T., Mulyampati, T., Namisango, E., Ojambo, K., Otike, C., Muhidini, N., Okiria, E., Kalyango, R., Wangi, R., Ndagire, R., Obuku, E., & White, H. (2020). The evidence base and gaps related to women's groups in Uganda. (Working Paper).

Demirguc-Kunt, A., Klapper, L., Singer, D., & Van Oudheusden, P. (2015). *The global Findex database 2014: Measuring financial inclusion around the world*. Washington, DC: World Bank.

Desai, P., Speed, M., & MacLean, L. (2018). *Nigeria for Women Project, social analysis—final report.* London, UK: Social Development Direct.

Desai, S., Misra, M., Das, A., Singh, R., Sehgal, M., Gram, L., … Prost, A. (2020). *Community interventions with women's groups to improve women's and children's health in India: A mixed-methods systematic review of effects, enablers and barriers*. BMJ Global Health, 5(12), e003304

Desai, S., de Hoop, T., Anderson, L., Darmstadt, G., & Siwach, G. (2019). Learning agenda on women's groups. Retrieved from https://womensgroupevidence.org/sites/default/files/Learning-Agenda-on-Women's-Groups.pdf

Dhaliwal, I., Duflo, E., Glennerster, R., & Tulloch, C. (2013). Comparative cost-effectiveness analysis to inform policy in developing countries: A general framework with applications for education. In P. Glewwe (Ed.), *Education policy in developing countries* (pp. 285–338). Chicago, IL: University of Chicago Press.

Dìaz-Martin, L., Gopalan, A., Guarnieri, E., & Jayachandran, S. (2020). *Greater than the sum of the parts? Evidence on mechanisms operating in women's groups*. Retrieved from https://faculty.wcas.northwestern.edu/~sjv340/womens_groups.pdf

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, *4*, 3895–3962.

Evans, D. K., & Popova, A. (2014). *Cost-effectiveness measurement in development: Accounting for local costs and noisy impacts*. Retrieved from https://documents.worldbank.org/en/publication/documents-reports/documentdetail/969291468340210399/cost-effectiveness-measurement-in-development-accounting-for-local-costs-and-noisy-impacts

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2016). *Impact evaluation in practice*. The World Bank.

Hoffmann, V., Rao, V., Surendra, V., & Datta, U. (2020). Relief from usury: Impact of a self-help group lending program in rural India. *Journal of Development Economics*, 102567.

Karlan, D., Savonitto, B., Thuysbaert, B., & Udry, C. (2017). Impact of savings groups on the lives of the poor. *Proceedings of the National Academy of Sciences*, *114*(12), 3079–3084.

Kopper, S., & Sautmann, A. (2020). Best practices *for conducting phone surveys.* Retrieved from https://www.povertyactionlab.org/blog/3-20-20/best-practices-conducting-phone-surveys

Lau, C., Lombaard, A., Baker, M., Eyerman, J., & Thalji, L. (2019). How representative are SMS surveys in Africa? Experimental evidence from four countries. *International Journal of Public Opinion Research*, *31*(2), 309–330. https://doi.org/10.1093/ijpor/edy008

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, *99*(2), 210–221.

Mikkelsen, B. (2005). *Methods for development work and research: A new guide for practitioners*. New Delhi, India: SAGE Publications India. doi:10.4135/9788132108566

Morgan, D. L. (1996). Focus groups. *Annual review of sociology*, 22(1), 129-152.

Namisango, E., de Hoop, T., Holla, C., Siwach, G., Chidiac, S., Jayaram, S., Grzeslo, J., Sulaiman, M., Janoch, E., Majara, G., Adegbite, O., Anderson, L., Walcott, R., Jafa, K., Desai, S., Dirisu, O., Mulyampiti, T., Hakspiel, J., & Panetta, D. (2021). *Women's Groups and COVID-19: Effects on and challenges of savings groups* Retrieved from http:///www.womensgroupevidence.org

Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. (2006). Process evaluation in randomised controlled trials of complex interventions. *BMJ*, *332*(7538), 413–416.

Organisation for Economic Co-operation and Development (OECD). (2019). *SIGI global report 2019: Transforming challenges into opportunities*. Retrieved from https://www.oecd-ilibrary.org/development/sigi-2019-global-report_bc56d212-en

Rychetnik, L., Frommer, M., Hawe, P., & Shiell, A. (2002). Criteria for evaluating evidence on public health interventions. *Journal of Epidemiology and Community Health*, *56*(2), 119–127.

Siwach, G., de Hoop, T., Ferrari, G., & Belyakova, Y. (2019). *Guidelines on estimating cost-effectiveness of women's groups in international development.* Retrieved from http:///www.womensgroupevidence.org

Siwach, G., Paul, S., & de Hoop, T. (2021). *Economies of scale of large-scale international development interventions: Evidence from self-help groups in India.* Retrieved from http:///www.womensgroupevidence.org

Walcott, R., Schmidt, C., Kaminsky, M., Singh, R., Anderson, L., Desai, S., & de Hoop, T. (2021). *Women's groups, covariate shocks, and resilience: An evidence synthesis of past shocks to inform a future response.* Retrieved from http:///www.womensgroupevidence.org

World Health Organization (WHO). (2019). *Maternal health in Nigeria: Generating information for action*. Retrieved from https://www.who.int/reproductivehealth/maternal-health-nigeria/en/

AIR®

AMERICAN INSTITUTES FOR RESEARCH®

Established in 1946, the American Institutes for Research® (AIR®) is a nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally in the areas of education, health, and the workforce. AIR's work is driven by its mission to generate and use rigorous evidence that contributes to a better, more equitable world. With headquarters in Arlington, Virginia, AIR has offices across the U.S. and abroad. For more information, visit www.air.org.

MAKING RESEARCH **RELEVANT**

AMERICAN INSTITUTES FOR RESEARCH

1400 Crystal Drive, 10th Floor

Arlington, VA 22202-3289 | 202.403.5000

www.air.org

LOCATIONS

Domestic: Arlington, VA (HQ) | Sacramento and San Mateo, CA | Chicago, IL | Indianapolis, IN | Waltham, MA | Rockville, MD | Chapel Hill, NC | Austin, TX

International: Ethiopia | Haiti

XXXXX_02/21