Are government or non-governmental organization staff more effective in delivering frontline mental health services? Evidence from a cluster randomized controlled trial

Melissa Hidrobo, Negussie Deyessa, Daniel O. Gilligan, Adiam Hagos Hailemicheal, Tigabu Getahun, Jessica Leight, Michael Mulford[1]

Pre-analysis plan

Abstract: Large treatment gaps in mental health services exist in low- and middle-income countries, and psychotherapy interventions delivered by non-specialists have increasingly shown promise as a strategy to fill this gap. However, scaling up these interventions is challenging, and there is little evidence assessing whether delivering through existing large-scale government health structures is feasible or effective. Using a randomized controlled trial conducted in rural Ethiopia, this study assesses the effectiveness of group psychotherapy, group Problem Management plus (gPM+), delivered in separate arms by either stipended local facilitators engaged by a non-governmental organization or government-employed health extension workers. Primary outcomes of interest assessed one month and twelve months post-intervention include mental health (depression, stress, anxiety, and post-traumatic stress) and child development. In addition, to understand mechanisms for any differential effects across delivery platforms, we will collect detailed data on knowledge, attitudes, and productivity comparing across government and NGO workers.

[1] Hidrobo: International Food Policy Research Institute (IFPRI), m.hidrobo@cgiar.org; Deyessa: University of Rwanda, negdaysun@gmail.com; Gilligan: IFPRI, d.gilligan@cgiar.org; Hailemichael: World Bank, ahailemicheal@worldbank.org; Getahun: EconInsight, tigyget14@gmail.com; Leight: IFPRI, j.leight@cgiar.org; Mulford: World Vision, michael_mulford@wvi.org

1. Introduction

Mental health disorders are among the leading causes of the global burden of disease (World Health Organization, 2017), and the prevalence of these conditions — including but not limited to depression, anxiety, and post-traumatic stress disorder — is particularly high in conflict-affected populations (Charlson et al., 2019). Mental health challenges can also have meaningful economic implications, affecting investment in children, labor productivity, and individuals' self-perceptions (Angelucci & Bennett, 2024; Baranov et al., 2020; Bhat, De Quidt, Haushofer, Patel, Rao, Schilbach, & Vautrey, 2022). Despite this high prevalence, adequate care for mental illness is largely inaccessible in low- and middle-income countries due to an acute shortage of skilled staff, particularly in rural areas (Patel et al., 2018): the treatment gap (the percentage of individuals who require care but do not receive it) is estimated at 80% (Ridley et al., 2020).

To address this gap, simplified forms of psychotherapy delivered by non-specialist providers have been tested and shown to generate promising results, but the majority of these interventions have been relatively small-scale (Singla et al., 2017). Wider scaling of health interventions typically relies heavily on government health systems (Yamey, 2012). At the same time, a growing literature has generated evidence that across multiple sectors, government-delivered interventions achieve meaningfully smaller effects vis-à-vis interventions delivered by non-governmental organizations (NGOs), or fail to generate any positive effects at all (Bold et al. 2018, Cameron et al. 2019, Fitch-Fleischmann and Kresch 2021, Henderson and Lee 2015, Mo et al. 2020, Schwartz and Bhushan 2004). This raises the question of whether government health systems can effectively deliver frontline mental health services in rural low-income settings.

Using a randomized controlled trial (RCT) conducted in rural Ethiopia, this study seeks to assess the effectiveness of a low-cost psychotherapy intervention, group Problem Management plus (gPM+), delivered in separate arms by government-employed health extension workers (HEWs) or by stipended local facilitators (LFs) engaged and compensated by an NGO. gPM+ is a flexible, transdiagnostic intervention that has shown to be promising in addressing a range of mental health challenges in the short term: there is evidence of positive effects of gPM+ on mental health in other contexts (Jordans et al., 2021; Rahman et al., 2019) and in early findings from another trial conducted by the same research team in another region of Ethiopia (Alderman et al., 2022). The study will be conducted in Tigray, in the aftermath of a two-year war that acutely disrupted health services and generated devastating effects on food security, health, and livelihoods (Abay et al., 2022; Weldegiargis et al., 2023). Trauma and mental distress are extremely high in the region (as described in more detail below drawing on data from our baseline survey), rendering the need to find scalable solutions for the provision of psychosocial support even more acute.

More specifically, this RCT randomly assigned 222 villages across four districts in Tigray to three study arms. The first arm serves as a control arm and did not receive any targeted psychotherapy; the second arm received gPM+ delivered by HEWs; and the third arm received gPM+ delivered by LFs who are community members engaged on a part-time basis by World Vision. (While the timeline will be described in more detail further below, the baseline survey was concluded in June 2024 and the intervention of five weeks rolled out immediately thereafter, and thus the intervention is now effectively concluded; we drafted and prepared this registered report for submission immediately upon accessing the cleaned baseline data to facilitate a more detailed analysis of baseline characteristics and baseline balance.) We plan to assess the effectiveness and cost-effectiveness of gPM+ delivered by the two sets of facilitators

(HEWs and LFs) on the primary outcomes of mental health and child development measured at one month and 12 months post- program, as well as a range of secondary outcomes including economic outcomes (labor, savings, and food security), behavioral outcomes (alcohol abuse and intimate partner violence), cognitive and non-cognitive skills (forward digit span and General Self-Efficacy), social support, and childcare practices. In addition, to better understand mechanisms for any differential effects across delivery platforms, we have collected and will collect detailed data on knowledge, attitudes, and productivity of both HEWs and LFs.

gPM+ as delivered in this study targeted a sample of individuals who are all beneficiaries of Ethiopia's Productive Safety Net Program (PSNP) and are thus among the poorest 15% of households in rural Ethiopia; in addition, they were identified as eligible for therapy based on the prevalence of depressive symptoms at baseline measured using the PHQ-9. We focused on PSNP households for two reasons: first, the collaborating non-governmental organization (World Vision) has a mandate to support delivery of the PSNP and provide supplementary services to PSNP households. Second, extremely poor households are presumably among those most vulnerable to mental health-related challenges and also characterized by lowest access to health services (Peters et al., 2008), a challenge magnified in post-conflict settings, and thus exploring the effectiveness of various delivery platforms in serving this population is of particular interest.

This study will contribute to a growing body of literature analyzing the relative effectiveness of government vis-à-vis NGO delivery, a literature in which evidence around delivery of health services is very limited. A systematic analysis of impact estimates from 635 papers across 20 interventions found that the effects of government-implemented programs are generally lower vis-a-vis NGO-implemented programs (Vivalt, 2020), and this evidence is complemented by an increasing number of targeted trials comparing different modalities of intervention implementation. In Kenya and China, the scale-up of an education intervention had significantly larger effects when implemented by an NGO vis-à-vis government implementation (Bold et al., 2018; Mo et al., 2020), and similar findings were identified for a sanitation intervention in Indonesia (Cameron et al., 2019). The performance of NGOs has also been found to exceed government performance in the targeting and provision of aid following natural disasters (Fitch-Fleischmann & Kresch, 2021; Henderson & Lee, 2015).[2] There is almost no evidence from the health sector, however, other than one study conducted in Cambodia where the contracting of government-provided primary health services to NGOs was found to be more effective in enhancing access of services by the poor (Schwartz & Bhushan, 2004).

Our study will also contribute to the larger literature on the effectiveness of psychotherapy in LMICs. While this literature generally finds positive effects, the majority of this evidence is drawn from short-term studies, conducted within six months of the conclusion of the intervention (Bhat, De Quidt, Haushofer, Patel, Rao, Schilbach, Vautrey, et al., 2022; Singla et al., 2017). Only a handful of studies assess impacts one year or more after the program ends, and these findings are more mixed. For example, in Kenya individual PM+ did not have any significant effects on economic or mental health-related outcomes one year post-endline (Haushofer et al., 2020). Baranov et al. (2020b) find significant long-term effects on both mental health and economic outcomes of psychotherapy delivered to women experiencing post-partum depression in Pakistan. However, a similar intervention delivered by peers

[2] Other related evidence suggests that effective NGO implementation is primarily observed in contexts where the organization has prior engagement with the community (Usmani et al., 2022); and the provision of health care by an NGO can complement the government's provision of healthcare when labour is abundant but not when labor is scarce (Deserranno et al., 2024).

rather than community health workers had null effects (Maselko et al., 2020). Similarly, a long-term follow-up of two psychotherapy interventions delivered to women in India found that one trial delivered by counselors had positive long-term effects on both mental health and economic outcomes, while the other delivered by peers, had no significant effects, even in the short term (Bhat, De Quidt, Haushofer, Patel, Rao, Schilbach, Vautrey, et al., 2022).

The heterogeneous findings in the literature to date suggests that the delivery platform may meaningfully shape both the initial effects of psychotherapy as well as the sustainability of effects, but this literature as well as the broader literature on frontline health services has generated relatively little evidence directly comparing modalities of delivery. A wider scaling of psychotherapy interventions at the population level would require a better understanding of how to train, supervise, and compensate a large number of non-specialist providers (Bunn et al., 2021) either through existing government health systems or by coordinating with other actors such as NGOs, and our trial is well-positioned to contribute evidence to inform this question.

2. Methods

a) Study setting

Tigray is a region recovering from acute conflict, following the conclusion in November 2022 of two years of active war between the Tigray People's Liberation Front and forces allied with the government of Ethiopia; all evidence suggests a population facing an acute multidimensional crisis. The onset of conflict led to widespread and near-complete disruption of health services (Gesesew et al., 2021) and food systems, leading to high rates of food insecurity (Weldegiargis et al., 2023). Smallholder agriculture was also severely affected as four out of five households lost their crops; three-quarters lost their livestock; nearly half lost their farm tools; and one-third were internally displaced (Manaye et al., 2023). There has also been widespread direct exposure to violence and extremely high reported rates of war-related sexual and gender-based violence, with nearly half of women reported to be affected (Fisseha et al., 2023). Estimates of the direct mortality effects are also in the range of 100,000 deaths or more.[3] While the consequences of the violence in terms of mental health and psychosocial outcomes remain largely unmeasured, they are anticipated to be severe.

Our study is also conducted within the system of community health workers (deemed health extension workers or HEWs) in Ethiopia. The Health Extension Program (HEP) was launched in 2003 to address significant health challenges, especially in rural and underserved areas of the country (Assefa et al., 2019). HEWs are primarily women who are chosen from the communities they serve and receive a year-long training (Nsibande et al., 2018), and are subsequently employed as formal salaried government staff operating from health posts. In rural areas, each health post is staffed with two HEWs that serve a kebele, the subadministration unit below district, including approximately 3,000 to 5,000 people (World Health Organization, 2023). HEWs deliver health services at health posts and through home visits and community outreach under four themes: family health, disease prevention and control, hygiene and environmental sanitation, and health education and communication. In 2016, mental health was added as part of disease prevention and control (World Health Organization, 2023).

[3] See for example Taylor, Adam. "A historic rise in global conflict deaths suggests a violent new era." https://www.washingtonpost.com/world/2023/06/29/conflict-war-deaths-global-peace-rise-casualty/

Despite the large scale and penetration of HEWs, however, there remain meaningful challenges to their ability to effectively deliver health services. These challenges include lack of motivation, skills, knowledge, and productivity, in addition to systemic challenges related to frequent turnover, high workload, and lack of supervision or support (World Health Organization, 2023). There is widespread evidence that HEWs devote relatively little time to direct service provision (Tilahun et al., 2017), perform required tasks infrequently and and/or are inadequately prepared for those tasks (Desta et al., 2017); and achieve low rates of service utilization (Tiruneh et al., 2024). This raises important questions around the efficacy of using HEWs to scale frontline services for psychosocial support.

b) Intervention

Problem Management Plus (PM+) is an intervention developed by the World Health Organization in 2013 to address common mental health problems such as depression, stress, and anxiety. It encompasses problem management (PM) plus (+) selected behavioral strategies to address both psychological issues (e.g., stress, fear, feelings of helplessness) and, where possible, practical challenges (e.g., livelihood problems, conflict in the family) (World Health Organization, 2016). While PM+ was initially conceived as an individual counselling intervention, it has since been adapted to be delivered as a group-based intervention (denoted gPM+) with studies in Nepal and Pakistan demonstrating its effectiveness and feasibility (Jordans et al., 2021; Rahman et al., 2019). gPM+ is designed to be scalable and implemented by non-specialists following an intensive 10-day training, rendering it particularly appropriate for a low-income context with an acute shortage of health personnel with expertise in mental health and psychiatry.

Most recently, gPM+ has been adopted and implemented in the Amhara and Oromia regions of Ethiopia by World Vision, and preliminary findings from a randomized controlled trial conducted by this research team revealed positive effects at one-month post-treatment across a variety of standardized mental health outcomes, including significant reductions in depression, anxiety, and stress. The intervention was also characterized by a high fidelity of implementation, with the average respondent invited to participate attending 4.5 out of five sessions. While these initial results are promising, questions remain about how this intervention can feasibly be scaled given the requirement for a large number of trained, community-based staff.

Accordingly, this trial tests the effectiveness of two implementation models for delivering gPM+. The first model relies on HEWs, public employees already under government supervision who were trained and equipped by World Vision to deliver gPM+; the study design did not entail any supplemental compensation for providing this service. The second entails LFs recruited for a short-term assignment and compensated hourly (i.e., not on a full-time basis) by our partner NGO, World Vision.[4] HEWs were supervised by their government supervisors (also trained by World Vision), while LFs were supervised by NGO staff. The objective was to recruit LFs with similar educational and professional qualifications to HEWs (more information about the characteristics of the recruited sample is provided below). Both sets of workers and their supervisors were also provided with access to a phone equipped with a CommCare Connect application developed by Dimagi to facilitate and monitor sessions as well as allow supervisors to engage with facilitators.

All facilitators were invited to a 10-day training convened by World Vision and provided with a per diem during the trained period, with supervisors trained separately. Both supervisors and HEWs/LFs were also engaged in practical skill-building in which they facilitated sample gPM+ sessions outside the study area.

[4] Formally, the LFs are contracted by the woreda health office, but technical supervision and support is provided by World Vision MHPSS Officers and the stipend is provided by World Vision.

Following the training and practicum period, each facilitator was assigned two gPM+ groups including up to eight individuals each, in order to deliver five 90–120-minute group sessions weekly over five weeks. The number of groups was consistent for all facilitators (whether HEWs or LFs), and the total anticipated workload for each facilitator was thus around 6—8 hours weekly for a period of five weeks (three to four hours per group). As the provision of mental health support is a recently added responsibility for HEWs and thus within their scope of work, HEWs did not receive any additional compensation for engaging in this new activity. LFs received a monthly payment for three months,[5] compensating them for time committed to training, practice, preparation, group facilitation, and follow-up during this period.

c) Study design

The evaluation sample was drawn from four districts or woredas (Samre, Tsaeda Emba, Keletawelaelo, and Enderta); within this sample, randomization was conducted at the village level (the sub-administrative unit below the kebele), stratified by kebele. The sample comprises 222 villages across 61 kebeles, and the research team conducted randomization in Stata in February 2024, prior to any data collection or implementation, to allow for facilitator recruitment.

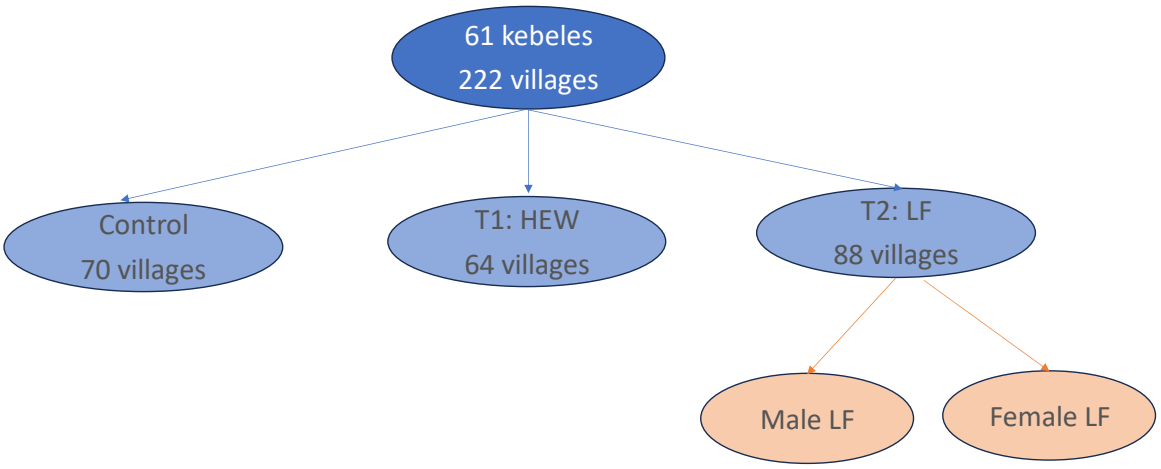
Within each kebele, villages were randomized to receive one of the three experimental arms described below; Figure 1 summarizes the study design.

- Control: no intervention
- Treatment 1: gPM+ delivered by HEWs
- Treatment 2: gPM+ delivered by LFs

Given that the number of villages per kebele differs across kebele, ranging from two to seven, the exact assignment per kebele varied (see Appendix Table 1). For example, in kebeles with only 2 villages, the assignment ratio entailed one control village to 1 HEW village to 0 LF villages. In kebeles with 3 villages, the ratio was 1:1:1. In kebeles with four villages, the ratio was 1:1:2, etc. Within T2, there was a further sub-randomization in which villages were randomly assigned to either a male or female LF (see figure 1).

Figure 1: Study design

[5] The stipend provided is 2,600 etb ($45) per month for facilitating two groups, or roughly one third of a monthly HEW salary.



Given that HEWs already have a prior division of labor across villages within their kebeles, in kebeles with only one village assigned to T1, the HEW who was already internally assigned to work in this village was assigned to deliver gPM+. In kebeles with two villages assigned to T1, two HEWs were assigned to deliver gPM+ (one HEW per village). Each HEW (who are all women) was assigned to facilitate one gPM+ group for men and one gPM+ group for women, and thus, eight men and eight women were recruited in each village served by a HEW.

World Vision also recruited both male and female local facilitators at the kebele level. Male LFs were assigned to facilitate two gender-matched (male) groups, while female LFs were assigned to facilitate one male and one female group.[6] Thus, in villages served by male LFs, 16 men were recruited to participate in gPM+, while in villages served by female LFs, eight women and eight men were recruited.

In both HEW and LF arms, individuals recruited to participated in gPM+ were PSNP clients (either direct support or public works) between 18 and 64 years old, who were heads of household or the spouse of the household head, and who exhibited signs of moderate to moderately severe depression (PHQ-9 score between 10 and 19, inclusive).

d) Hypotheses

Our primary objective is to assess the following null hypotheses:

1. gPM+ has no impact on mental health outcomes one month and one year post-intervention.
2. gPM+ has no impact on related indicators (including the primary outcome of child development and a range of secondary outcomes) at one month and one year post-intervention.
3. There are no differences in the effects of gPM+ on the primary outcomes of interest compared across HEW and LF delivery platforms.

In testing the first and second hypotheses, we build on limited evidence evaluating the effectiveness of gPM+ in post-conflict setting both in the short term and longer term (Akhtar et al., 2022). In testing the

[6] Our previous experience and discussions with stakeholders suggested that male facilitators leading group therapy sessions for women is unlikely to be culturally acceptable or effective.

third hypothesis, we provide novel evidence about the relative effectiveness of two different modalities for delivering frontline psychosocial services on a large scale and build on a very limited literature analyzing the "NGO effect" in the health sector.

The theory of change around the overall effectiveness of gPM+ centers around the potential for the intervention to provide participating individuals with better strategies to manage challenges both psychological (e.g., stress, fear, feelings of helplessness) and practical (e.g., livelihood-related obstacles and intrahousehold conflict). These shifts are hypothesized to enhance mental health in both the short and medium-term. Given the more flexible and transdiagnostic structure of gPM+ vis-à-vis other psychotherapy interventions that more specifically target certain dimensions of mental health (Dawson et al., 2015), we additionally hypothesize that these effects will be manifest across multiple domains of mental health — a characteristic that is particularly important in this post-conflict setting where we anticipate high rates of multiple co-occurring mental health disorders (Charlson et al., 2019). The hypothesis of significant effects of gPM+ also rests on the assumption that the reductions in adverse mental health symptoms generated by the intervention will be larger and more persistent than any patterns of spontaneous remission that may be observed in the control arm; given that the eligible individuals were all characterized by at least moderate depression, this rate of remission may in any case be somewhat reduced.[7]

Given the hypothesis of enhanced mental health, our theory of change further envisions that enhanced mental health will shift behaviors, leading individuals to engage in more productive use of time and increase their level of investment in children (Baranov et al., 2020); shift their economic engagement and/or increase their economic productivity (Lund et al., 2022); reduce their use of counterproductive coping behaviors such as alcohol abuse or violence (Jane-Llopis & Matytsina, 2006); and strengthen their social ties and draw on more available social support (Wang et al., 2018). These shifts in behavior should in turn assist in sustaining the positive effects of psychotherapy in the medium-term.

Our third hypothesis evaluates the relative effectiveness of two delivery models in enhancing mental health. Here, the weight of previous evidence comparing government and NGO services suggests that NGO services often outperform their government counterparts (as summarized above in the literature review); however, there are differences between the two sets of service providers along multiple dimensions that could potentially shape effectiveness.

On the one hand, the effectiveness of gPM+ as delivered by LFs may be higher vis-à-vis the effectiveness of delivery through HEWs, given that the local facilitators may have stronger incentives for engagement (to obtain continued remuneration) and may face fewer competing demands on their time. On the other hand, the HEWs may be more skilled and experienced and thus better able to deliver the intervention, and they may have more convening power and established rapport with community members, increasing take-up of beneficiaries; they may also be better able to focus on their work given the relative stability of their government employment, while LFs may face more short-term stress if they are also managing, or seeking, other livelihoods activities. The two sets of workers will also be differently supervised (HEWs are supervised by public employees, LFs by NGO staff), and supervisors' incentives and constraints will differ. In general, both HEWs and HEW supervisors are expected to be multitasking across a wide range of health-related service responsibilities, and thus may be more likely to face trade-offs between effectively delivering gPM+ and effectively meeting other work-related responsibilities. Appendix Table

[7] Bhat et al. (2022) highlight in their analysis that rates of spontaneous remission are generally higher for less severe forms of depression.

2 summarizes the differences between HEWs and LFs across these multiple dimensions and provides an overview of how each type of service provider may be plausibly advantaged or disadvantaged along each dimension ex ante; in section 4 we present data from the baseline surveys conducted that allow us to further assess whether our priors around HEW and LF characteristics were correct.

In addition to testing these three core hypotheses, we will also conduct some exploratory analysis around the effects of facilitator gender, as our design allows us to investigate whether the effects of LF-provided psychotherapy are greater for men when the LF is male versus female. (We cannot conduct a similar exercise for the female sample as we do not have a group of male facilitators delivering gPM+ to females, a strategy that was deemed culturally unacceptable.) Here, the literature does not suggest any particularly clear hypothesis: there is some evidence that same-sex education can be more effective than mixed-gender education in some low- and middle-income contexts (Jackson, 2012, 2021), and some evidence of better-quality or more effective service provision when service providers are more socially proximate to beneficiaries (Karachiwalla, 2019; Raghunathan et al., 2023). However, to our knowledge, there is no evidence around the effectiveness of gender-matched health services. Given the absence of a clear theoretical framework — and given that this analysis may be underpowered, since it is restricted only to male respondents — we designate this line of inquiry as exploratory.

e) Primary and secondary outcomes

The study's primary outcomes (summarized in Table 1) include mental health and child development. The mental health-related outcomes include depressive symptoms as measured by the PHQ-9, stress as measured by the Perceived Stress Scale (PSS-10), post-traumatic stress as measured by the abbreviated PCL-C instrument, and anxiety as measured by the General Anxiety Disorder 7 (GAD-7).

Child development is measured using the Strengths and Difficulties Questionnaire, and is identified as a primary outcome for two primary reasons: previous literature has suggested the potential for psychotherapy to have significant and persistent effects on child development, with potentially meaningful intergenerational implications (Baranov et al., 2020), and our baseline data (described in more detail below) suggests a strikingly high prevalence of behavioral and developmental challenges among young children in this context. All mental health outcomes will be collected at baseline, one month, and one year post-program, while the SDQ will only be collected at baseline and one year post-program.

Table 1: Primary outcomes

Domain	Measure	Description	Timing of measurement
Mental health	PHQ-9	Continuous total score (standardized) and binary for any moderate to severe depression (PHQ-9>=10).	1 month post-program, 12 months post-program
	Perceived stress scale (PSS-10)	Continuous total score (standardized)	
	Post traumatic stress (abbreviated PCL-C)	Continuous total score (standardized)	
	General Anxiety Disorder 7 (GAD-7-)	Continuous total score (standardized)	

Child development	Emotional and behavioral challenges as measured by Strengths and Difficulties Questionnaire (SDQ)	Continuous total score for children 4-15 years old (standardized)	12 months post-program

In addition to the primary outcomes of interest, we are interested in a range of secondary outcomes summarized in Table 2. These include economic outcomes (labor, savings, and food security); behavioral outcomes (alcohol abuse and intimate partner violence); cognitive and non-cognitive skills (General Self-Efficacy Scale, tension reduction checklist, forward digit span, and the Brief Cope), social support (as measured by the Multi-dimensional Scale of Perceived Social Support); daily activities (including time use and WHODAS); and childcare (childcare activities and neglect). These outcomes generally correspond to both channels for potential improvements in mental health and independent outcomes with associated welfare implications; the majority of secondary outcomes will be measured at one month and one year follow-up.

Table 2: Secondary outcomes

Domain	Measure	Details
Economic outcomes	Labor participation and economic activities	Binary for any participation across categories and continuous for total time and income across categories
	Food security as measured by the Food Insecurity Experience Scale (FIES)	Continuous total score
	Savings	Binary for any savings across different savings categories.
Behavioral outcomes	Alcohol use	Binary for any use
	Intimate partner violence (as measured by the WHO Violence Against Women Instrument)	Binary for any emotional, physical, and sexual violence
Cognitive and non-cognitive skills	Eight-item General Self-Efficacy Scale	Continuous total score
	Tension reduction checklist	Continuous total score
	Forward digit span	Continuous total score
	Brief Cope (abbreviated version)	Continuous total score
Social support	Multi-dimensional Scale of Perceived Social Support (MSPSS)	Continuous total score
Daily activities	Time use in productive activities	Continuous total score in minutes for time devoted to productive activities
	WHO Disability Assessment Schedule 2.0, WHODAS 2.0	Continuous total score
Childcare	Childcare activities	Continuous total score
	Child neglect	Continuous total score

	Child discipline	Continuous total score for different types of discipline

To better understand what shapes the relative effectiveness of the two delivery modalities (hypothesis 3), we will also assess a range of variables capturing facilitator characteristics. Knowledge of gPM+ and mental health in general and intrinsic motivation to deliver services will be measured using both survey and experimental measures in questionnaires administered directly to the facilitators pre- and post-training. Productivity (including the number of sessions conducted, the average number of minutes devoted to gPM+ activities, and the average attendance of beneficiaries at sessions) will be measured using administrative data.

Table 3: Facilitator outcomes

gPM+ knowledge	Score on survey module capturing knowledge of mental health and gPM+	Immediately post-training
Intrinsic motivation	Score on survey questions capturing motivation; experimental measure of motivation (donation to mental health charity in the dictator game)	Immediately pre-training
Productivity	Number of sessions conducted; average number of minutes devoted to gPM+ activities weekly; beneficiaries at sessions	Administrative data collected during implementation

f) Power calculations

Power calculations were conducted primarily drawing on data from the baseline survey conducted for this trial and focusing on the primary outcome enumerated in Table 1: PHQ-9, GAD-7, PSS, post-traumatic stress, and the SDQ. We do not take attrition into account in power calculations given that recent trials conducted by the same research team using samples of PSNP beneficiaries have achieved minimal attrition (on average, 2-3%).[8]

We estimate the minimum detectable effect (MDE) using the total number of clusters in the trial (222) and the cluster size of 16, accounting for intracluster correlation; for SDQ, we set the cluster size as 12, as approximately 75 percent of the sample had a child in the SDQ age range (four to 15 years old). Power calculations were conducted in STATA using the "power" command, setting the desired statistical power as 80% and the significance level at 0.05. Given the slight differences in the number of villages per intervention arm, we conduct power calculations separately for each comparison.

Tables 4a through 4c report a MDE that ranges from 0.15 SD for PSS on the comparison of the LF to the control (Table 4b) to 0.22 for SDQ on the comparison of the HEW to control (Table 4a). The MDE for PHQ-9 ranges from 0.16-0.18 SD, which is similar to the impact of 0.18 SD found in the one-month post-

[8] Household mobility among PSNP beneficiary households is typically extremely low given that they are in receipt of benefits that are tied to their original location.

intervention evaluation in Amhara and Oromia and smaller than the impacts found in Nepal and Pakistan immediately after the intervention (Jordans et al., 2021; Rahman et al., 2019).

Table 4a. Power calculations, comparison HEW to control

	MDE	ICC	K villages control	K villages HEW	N per cluster	Total sample for comparison
PHQ-9 total score	0.177	0.076	70	64	16	2,144
General Anxiety disorder score	0.187	0.092	70	64	16	2,144
Perceived stress scale score	0.163	0.054	70	64	16	2,144
Post traumatic stress score	0.212	0.138	70	64	16	2,144
SDQ score	0.217	0.128	70	64	12	1,608

Clusters are villages. MDE is minimum detectable effect reported in standard deviations. ICC is intracluster correlation

Table 4b. Power calculations, comparison LF to control

	MDE	ICC	K villages control	K villages LF	N per cluster	Total sample for comparison
PHQ-9 total score	0.164	0.076	70	88	16	2,528
General Anxiety disorder score	0.173	0.092	70	88	16	2,528
Perceived stress scale score	0.151	0.054	70	88	16	2,528
Post traumatic stress score	0.197	0.138	70	88	16	2,528
SDQ score	0.201	0.128	70	88	12	1,896

Clusters are villages. MDE is minimum detectable effect reported in standard deviations. ICC is intracluster correlation

Table 4c. Power calculations, comparison LF to HEW

	MDE	ICC	K villages HEW	K villages LF	N per cluster	Total sample for comparison
PHQ-9 total score	0.168	0.076	64	88	16	2,432
General Anxiety disorder score	0.177	0.092	64	88	16	2,432
Perceived stress scale score	0.155	0.054	64	88	16	2,432
Post traumatic stress score	0.202	0.138	64	88	16	2,432
SDQ score	0.206	0.128	64	88	12	1,824

Clusters are villages. MDE is minimum detectable effect reported in standard deviations. ICC is intracluster correlation

3. Data

a) Sampling and eligibility criteria

The sample includes sixty-one kebeles across four woredas; the kebeles were purposively sampled based on World Vision's operational plans to roll out services. Within each of the 61 selected kebeles, all 222 villages were included in the trial.

At the household level, the sampling frame was constituted by the beneficiary list of PSNP households within the sample kebeles. Within this sample, a screening exercise was conducted to identify individuals who met the additional eligibility criteria: working-age adults (between 18 and 64 years old) who were heads of household or the spouse of the household head, and who exhibited signs of depression (PHQ-9 score between 10 and 19, inclusive). More specifically, the beneficiary list was randomly ordered and

stratified by village and PSNP category (public works or direct support),[9] with households within each village and PSNP category visited in random order to screen the head and/or spouse in each household. In the villages randomly assigned to the male LF, only men were screened. In the villages randomly assigned to the HEW, LF female, or control arms, both men and women were screened. Screening continued until the target eligible sample of 16 was reached in each village. All eligible individuals screened into the study were then administered a full baseline survey (described in more detail below).

We restricted the eligible sample population to working-age adults given the broader objective of measuring economic outcomes related to mental health, such as childcare and economic activity. Individuals screened and identified to be experiencing severe depression (a PHQ-9 of 20 or above) or those reporting suicidal attempts or suicidal ideation were referred to health facilities for the provision of intensive psychosocial support services. They were not included in the study.

b) Data collection

An Ethiopian survey firm (EconInsight) conducted fieldwork in close collaboration with the study team, collecting the screening and baseline surveys in March – May 2024; the one month follow-up survey is scheduled to be launched in July 2024 and the one year follow-up survey in July 2025. The target timing for the short-term follow-up is one month following intervention conclusion. All surveys were and will be administered in Tigrinya using Computer Assisted Personalized Interviews (CAPI) implemented via SurveyCTO. Enumerators are male and female, with female enumerators surveying female respondents and male (or female) enumerators surveying male respondents. More details about data management are provided in Appendix B.

i) Baseline and endline surveys

The baseline and endline questionnaires were and will be questionnaires of approximately 90 minutes, composed of two parts. The first part included household-level modules such as a household roster and questions around housing, assets, food security, and household-level shocks. The second part included individual-level modules administered to the individual screened for the study. These modules encompassed instruments for measuring depression, stress, anxiety, coping, self-efficacy, childcare and development, savings, intimate partner violence of women, and time use/ labor participation. At baseline we also collected information on gender-based violence and conflict exposure. All screened study subjects received information on locally available mental health services, and all women who are administered the IPV module received information on support services (Ministry of Women and Social Affairs office) available in the area.

ii) Monitoring and facilitator data

Throughout the training and intervention period, the study team is collecting detailed data on the facilitators' performance as well as monitoring data on the group sessions facilitated by the use of a CommCare Connect application.

[9] Public works households are PSNP beneficiary households including a prime-age adult, who are required to contribute labor to public works in order to receive benefits. Direct support households are households that do not include a prime-age adult (typically, elderly households or households characterized by a disability) and thus have no labor requirement. Given the post-conflict context, direct support households are particularly common, and thus included in the sample. We stratified our sampling criteria by type of PSNP client to ensure equal proportions (75% public works and 25% direct support) were reached across all intervention arms.

First, we have administered or will administer several assessments to the facilitators themselves: a pre-test at the beginning of training, a post-test at the conclusion of training, and a debriefing questionnaire following the conclusion of the intervention. These instruments include or will include information on the facilitators' socioeconomic, educational, and professional background; their professional motivation; their interest in and knowledge of mental health-related challenges; their specific understanding of the principles and content of gPM+ and facilitation skills (in the post-training and post-intervention assessments); their exposure to conflict and recent shocks; and their mental health. The questionnaires for HEWs also particularly focused on their baseline allocation of effort and time across tasks. We obtained written consent for the facilitators and HEWs to participate at the initiation of the pre-training and post-training assessments.

Second, ongoing monitoring data collected in the CommCare application includes data on the date, time, and attendance at each session, the PSYCHLOPS scores of participating individuals, the facilitator's use of any ongoing learning tools, and the facilitator's allocation of time to preparation and mobilization for the session as well as travel to the session. Facilitators were informed upon recruitment that any data related to their use of the application is part of their ongoing professional duties and, thus, will be monitored by their supervisors and made available to the research team. Participants in gPM+ were also advised in the first gPM+ session that data on their evolving mental health outcomes will be collected to facilitate effective implementation of support and that this data would be available, if clinically required, to supervisors and available in anonymized form to the research team. The consent of participants was obtained and documented.

iii) Costing data

We are also compiling relevant cost data to assess the cost-effectiveness of the intervention as delivered by LFs and HEWs. The research team developed a cost data collection platform in close collaboration with the program implementation team at World Vision. The platform entails the compilation of cost data on expense items such as administration and staffing, beneficiary targeting, staff training, participants training, implementation (e.g., venue, stationery, transportation), and implementation supervision. The finance officers and project coordinators populated the cost data collection template by tracking actual expenses, measured in Ethiopian birr in the year incurred. Appendix Table 3 presents the different cost data collection categories.

4. Empirical analysis

a) Analytical strategy

Treatment effects for individual-level variables will be estimated using analysis of Covariance (ANCOVA) models capturing intent-to-treat effects. The primary analysis will pool both men and women and estimate equation (1), comparing outcomes in individuals assigned to the HEW (T1) and LF (T2) arms to outcomes in individuals assigned to the control arm. Given that the male LF subarm is characterized by a male-only sample only (in contrast to the other arms that all include both men and women), we estimate a specification including a full set of interactions with gender in order to abstract from any bias introduced by the varying gender sample composition.

$$Y_{ivk} = \beta_1 T1_{vk} \, X \, Female_{ivk} + \beta_2 T1_{vk} \, X \, Male_{ivk} + \beta_3 T2_{vk} \, X \, Female_{ivk} + \beta_4 T2_{vk} \, X \, Male_{ivk} + \beta_5 Male_{ivk} + \gamma Y_{vk,t=0} + X'_{ivk,t=0} \vartheta + \chi_k + \varepsilon_{ivk}, \quad (1)$$

The subscript ivk denotes an individual i in village (village) v in kebele k. $X'_{ivk,t=0}$ denotes any relevant baseline covariates, and χ_k denotes kebele strata fixed effects. Standard errors will be clustered as the village level.

We will also estimate the average effect of gPM+ as delivered by HEWs for both men and women (constructed as the mean of the coefficients β_1 and β_2) as well as the average effect of gPM+ as delivered by LFs (constructed as the mean of the coefficients $\beta_3 + \beta_4$). We will report the p-values corresponding to the hypothesis that the effects are equal across genders ($\beta_1 = \beta_2$ and $\beta_3 = \beta_4$); and the p-value corresponding to the hypothesis that these two mean effects are the same.

We also plan to separately estimate the effects of gPM+ as delivered by female local facilitators (T2F) and male local facilitator (T2M, in order to compare the effects of female local facilitator and HEWs (who are both serving mixed-gender samples including men and women). The coefficients of interest in this equation are those specific to the female LF (β_3 and β_4), and these coefficients will be compared to the parallel coefficients for HEWs (β_1 and β_2).

$$Y_{ivk} = \beta_1 T1_{vk} \text{ X } Female_{ivk} + \beta_2 T1_{vk} \text{ X } Male_{ivk} + \beta_3 T2F_{vk} \text{ X } Female_{ivk} + \beta_4 T2F_{vk} \text{ X } Male_{ivk} + \beta_5 T2M + \beta_6 Male_{ivk} + \gamma Y_{vk,t=0} + X'_{ivk,t=0}\vartheta + \chi_k + \varepsilon_{ivk}, \quad (2)$$

Lastly, we plan an exploratory analysis that will allow us to separately assess the effect of a gender-matched versus non-gender-matched LF (for male beneficiaries). Using equation)2_, we will test the hypothesis that the impact of a female LF on the male sample (β_4) is equal to that of a male LF (β_5).

For variables linked to facilitator qualifications and performance, we will estimate the following simple regression at the facilitator level using data from treatment villages only.[10]

$$Y_{vk} = \beta_2 T2_{vk} + \chi_k + \varepsilon_{vk},$$

There will be no imputation for missing data due to item non-response in the one month and twelve month follow-up surveys. Missing data on baseline variables will be set to zero and dummied out in the ANCOVA specifications.

We will address multiple hypothesis testing using two strategies. First, we will conduct indices within a family of outcomes following Kling (2007). Second, we will report p-values corrected for multiple hypothesis testing based on sharpened FDR (false discovery rate) q-values (Anderson, 2008). This correction will be implemented within the set of primary outcome variables; and within each family of secondary outcomes.

b) Cost-effectiveness analysis

The cost-effectiveness analysis will be conducted using an incremental approach to measure the cost per standard deviation reduction of the depression (PHQ-9 score) and perceived stress score.[11] The CEA will be used to compare the cost-effectiveness of gPM+ as delivered via two different modalities to other mental health interventions that have cost data available.

c) Heterogeneity

[10] Given that there is a one-to-one match between facilitators and villages, we continue to use village subscripts.

[11] All costs will be measured in birr and subsequently converted to USD using the average exchange rate of the year. The study will use a 3 percent discount rate, and overhead rates will be assumed to be 15 percent.

We also plan to conduct three analyses of heterogeneous effects, focusing on 1) subsamples of men and women; 2) subsamples characterized by different baseline levels of depression (moderate to moderately severe) and 3) subsamples characterized by different levels of exposure to conflict.

d) Baseline characteristics and balance: Household survey

Table 5 reports summary statistics to characterize the individual-level sample at baseline, as well as a series of balance tests evaluating whether there are significant differences across the male and female samples in the control and treatment arms.[12] The final row of the table reports a p-value corresponding to the test joint significance across baseline characteristics within each category; the penultimate column tests the joint significance across treatment-sample categories for each baseline characteristic.

On average, respondents in the control group are 47 years old and the overwhelming majority (88%) are married, living in households with more than six members. The level of human capital attainment and socioeconomic status is low, with respondents reporting less than three years of education on average; while over 60% of households have an improved roof, only 10% have an improved floor. In order to assess interaction with formal health services, we also asked respondents to report their frequency of interaction with HEWs, and this was not high: on average, respondents in the control group have interacted with a HEW only twice in the past year.

The baseline levels of mental health challenges in the sample are also notably high, consistent with both a setting of post-conflict recovery and the eligibility criteria for the trial. The average PHQ-9 score in the control group is nearly 13, corresponding to moderate depression, and almost half of respondents report moderate to severe anxiety (GAD of 10 or above); about one-third report moderate to severe post-traumatic stress. The average score on the strengths and difficulties questionnaire or SDQ (reported for children aged four to 15) is nine, and this corresponds to 17% of children characterized by a SDQ score over 14, a cutoff capturing the incidence of behavioral challenges according to a widely used classification.[13]

The observed intensity of mental health symptoms is also consistent with the high reported rates of recent conflict experiences: about a quarter of households in the control arm report that they were temporarily displaced for at least a month, and about 20 percent report an armed group was present in their village during the conflict. Most strikingly, about 18 percent report that they either directly experienced injury as a result of the conflict or lost a member of their household.

In general, these characteristics are balanced across the treatment and control arms. Most notably, the HEW treatment arm has smaller households than the control arm (both in the female and male subsample); in the male sample, HEW and LF have higher stress scores but are less likely to have moderate to severe anxiety compared to the control arm; and in the female LF arm (female sample), households are more likely to have had an armed group present in the village but less likely to have experienced an injury or death. However, the joint tests fail to reject the hypothesis of balance across all arms for every variable

[12] The regression is $Y = \beta_1 HEWFem + \beta_2 HEWMale + \beta_3 LFFFem + \beta_4 LFFMale + \beta_5 LFMMale + \beta_6 Male + \chi_k + \varepsilon$; where Y is each baseline variable.

[13] This is significantly higher than the corresponding incidence rate for behavioral challenges of around 7% in another recent sample of primary-age children in Ethiopia (Mekonnen et al., 2020), though other studies in other regions have shown a comparable incidence of behavioral challenges (Alenko et al., 2020).

except one (presence of an armed group). Moreover, within each treatment arm category, the joint test of significance across all baseline variables is never significant.

Table 5: Baseline Balance Tests

	HEW + M	HEW + F	LFF + M	LFF + F	LFM + M	Mean for Control	p-value for joint hyp on all T groups	N
Respondent's age (years)	-0.791	-0.649	0.015	1.018	0.629	46.77	0.125	3,475
	(0.677)	(0.706)	(0.800)	(0.776)	(0.633)			
Respondent is married	-0.016	-0.004	0.006	-0.003	0.008	0.88	0.912	3,475
	(0.022)	(0.033)	(0.024)	(0.035)	(0.019)			
Household size	-0.318**	-0.258**	-0.043	0.037	-0.156	6.37	0.072	3,475
	(0.149)	(0.131)	(0.170)	(0.163)	(0.145)			
Number of children in household	-0.208*	-0.163	-0.102	-0.032	-0.123	3.48	0.325	3,475
	(0.113)	(0.108)	(0.127)	(0.127)	(0.111)			
Years of education (respondent)	0.051	0.364*	-0.136	-0.325	-0.183	2.74	0.073	3,475
	(0.231)	(0.198)	(0.236)	(0.230)	(0.200)			
Binary variable: improved roof	0.016	0.009	-0.005	-0.040	-0.022	0.63	0.824	3,475
	(0.035)	(0.033)	(0.039)	(0.044)	(0.035)			
Binary variable: improved floor	0.012	-0.018	0.027	-0.023	0.025	0.10	0.856	3,475
	(0.026)	(0.028)	(0.029)	(0.042)	(0.027)			
Past-year interactions with HEW	0.003	-0.013	0.201	-0.357*	0.007	2.05	0.395	3,474
	(0.189)	(0.187)	(0.230)	(0.199)	(0.177)			
PHQ-9 total score	-0.067	-0.200	0.054	0.130	-0.080	12.79	0.468	3,475
	(0.141)	(0.148)	(0.156)	(0.155)	(0.136)			
Anxiety (GAD-7) score	-0.198	-0.384	-0.080	0.032	-0.280	8.95	0.623	3,474
	(0.260)	(0.269)	(0.304)	(0.292)	(0.266)			
Moderate / severe anxiety	-0.077**	-0.017	-0.076*	0.022	-0.089**	0.46	0.153	3,474
	(0.036)	(0.033)	(0.043)	(0.037)	(0.036)			
Stress (PSS) score	0.674**	-0.225	0.578*	0.007	0.529*	20.17	0.285	3,475
	(0.314)	(0.377)	(0.325)	(0.428)	(0.306)			
PTSD (PCL-5) score	0.741*	-0.282	-0.195	0.188	0.349	11.62	0.479	3,475
	(0.434)	(0.496)	(0.580)	(0.516)	(0.436)			
Moderate/severe PTSD	0.046	-0.021	-0.021	0.029	0.035	0.34	0.489	3,475
	(0.034)	(0.037)	(0.043)	(0.044)	(0.033)			
Child's SDQ score	0.098	-0.435	0.130	-0.910**	0.252	9.38	0.438	2,635
	(0.367)	(0.407)	(0.375)	(0.446)	(0.352)			
Displaced during conflict	0.062*	0.015	0.007	-0.027	0.040	0.23	0.372	3,475
	(0.031)	(0.031)	(0.034)	(0.040)	(0.029)			
Armed group present in village during conflict	0.007	-0.019	0.086*	0.082**	0.046	0.80	0.040	3,469
	(0.035)	(0.036)	(0.044)	(0.037)	(0.040)			

	HEW + M	HEW + F	LFF + M	LFF + F	LFM + M	Mean for Control	p-value for joint hyp on all T groups	N
Injury/death due to conflict	0.039	-0.024	-0.008	-0.075**	0.019	0.18	0.117	3,474
	(0.026)	(0.030)	(0.032)	(0.037)	(0.027)			
p-val for joint hyp on treatment group effect	0.010	0.243	0.129	0.238	0.038			

* p < 0.10, ** p < 0.05, *** p < 0.01
Controlling for respondent gender
Standard errors clustered at the village-level, Kebele fixed effects included
2 observations where LF is Male, and the respondent is female are dropped from the specification
LFM = local facilitator is male, LFF = local facilitator is female, HEW = Health Extension Worker, '+ M' = respondent is male, '+ F' = respondent is female

e) Baseline characteristics and balance: facilitator survey

Table 6 then summarizes characteristics of the facilitators, as reported in the pre-training survey. The first panel summarizes basic demographic characteristics; the second and third panels report variables specific to the samples of HEWs and LFs, respectively. Focusing first on panel A: it is evident that HEWs are significantly older (at 35 years) than both female and male LFs (25 and 30 years, respectively); but somewhat contrary to our ex-ante hypothesis, there is no evidence of any significant difference in education. Female LFs are in fact somewhat more educated (fourteen years of education), while both HEWs and male LFs report around thirteen years of education (marginally higher for male LFs). Male LFs are more likely to have prior paid professional work experience (for HEWs, this is defined as prior to their tenure as a HEW), at 20%; while less than 10% of HEWs and female LFs have prior experience. Consistent with their older age and more established professional trajectory, HEWs are significantly more likely to be married than local facilitators, and have roughly double the number of assets using a simple count index.

However, on a range of other characteristics measured (self-efficacy, general knowledge of mental health, health worker motivation, pro-sociality, and their own current symptoms of depression) there is little evidence of any difference between HEWs and LFs. We measure self-efficacy using the General Self Efficacy Scale and general knowledge of mental health using the MAKS scale as previously adapted in Ethiopia (Tesfaye et al., 2022). We measure health worker motivation using survey questions from an context-specific scale adapted from Abate et al., (2022); and we measure pro-sociality using the Inclusion of Others in Self (IOS) scale as in Ashraf et al., (2020); as well as a simple experimental dictator game in which respondents can choose how much of an endowment (50 birr) they will direct to a charity related to mental health, as in Deserranno (2019). There is no evidence of any significant differences in health worker motivation or in choices in the dictator game, where the mean in each subarm indicates participants choose to retain approximately 40% of the endowment while donating 60%; however, health workers do score meaningfully higher on the IOS scale, suggestive of greater pro-sociality or altruism.

In Panels B and C, we observe that the average HEW has been engaged in this position for more than 12 years – consistent with the older age – but with minimal prior work experience. The average LF has less than a half-year of work experience; between 30% and 50% of LFs were most recently self-employed (many of them in agriculture, i.e. farming their own plot); a minority reported they had other wage employment, and 30-40% reported that they were unemployed or out of the labor force prior to taking up this role. Male LFs were more likely to report self-employment, and female LFs were more likely to report unemployment.

Returning to our conceptual framework, it is clear that HEWs enjoy a significant advantage in terms of health sector experience; they are generally mid-career professionals while LFs are generally young and characterized by minimal exposure to the labor market. That being said, HEWs do not show any evidence of greater education or any greater knowledge around mental health-related concepts prior to training, though presumably their broader health knowledge is stronger. Perhaps most importantly, there is no evidence of any differential pro-social motivation, suggesting that experience in the health sector has not generated a cadre of health workers with any noticeable difference in motivation compared to a set of young and almost entirely inexperienced local facilitators.

Table 6: Facilitator characteristics

	Sample size N	Mean and standard deviation			p-Value of difference	
Panel A: Comparing HEWs and LFs						
	Total	HEWs	Female LFs	Male LFs	HEWs vs Female LFs	HEWs vs Male LFs
Age of respondent	150	34.89 (7.60)	25.07 (4.13)	30.41 (5.95)	0.00	0.00
Years of education	150	12.73 (2.43)	14.27 (3.18)	13.67 (3.52)	0.01	0.10
Any prior work experience?	150	0.06 (0.25)	0.07 (0.26)	0.20 (0.40)	0.85	0.04
Currently married	150	0.60 (0.49)	0.24 (0.43)	0.50 (0.51)	0.00	0.29
Count index of assets owned	150	9.84 (6.80)	5.80 (3.48)	5.96 (4.86)	0.00	0.00
Self-efficacy	150	52.25 (5.08)	51.32 (6.10)	53.15 (7.83)	0.40	0.47
Index: Knowledge of mental illness	150	11.40 (1.84)	10.80 (1.81)	11.22 (1.56)	0.11	0.58
Index: pro-social motivation	150	51.65 (6.57)	52.73 (4.74)	53.26 (5.04)	0.37	0.17
Inclusion of others in self (IOS) scale	150	6.24 (1.19)	5.27 (1.45)	5.35 (1.39)	0.00	0.00
Dictator game: Amount retained	150	21.03 (16.07)	23.90 (14.30)	21.20 (15.75)	0.37	0.96
PHQ-9 total score	150	4.89 (4.33)	4.68 (4.57)	5.00 (4.08)	0.79	0.89
Panel B: HEW characteristics						
Years of experience as HEW	63	12.46 (5.31)				
Years of past work experience	63	0.17 (1.04)				
Panel C: LF characteristics						

Years of past work experience	87		0.19	0.41
			(1.13)	(1.06)
Currently self-employed	87		0.34	0.48
			(0.48)	(0.51)
Currently engaged in wage labor	87		0.20	0.20
			(0.40)	(0.40)
Currently unemployed	87		0.41	0.30
			(0.50)	(0.47)
Currently student	87		0.05	0.02
			(0.22)	(0.15)

Notes: Standard deviations are in parentheses. The p-value is from a regression of the variable of interest on a binary variable for HEW conditional on strata fixed effects.

5. Limitations and Challenges

Though this trial has a number of strengths, it also has some significant limitations. In terms of successfully concluding the trial, the primary risk is the re-emergence of the conflict that unfolded in Tigray in 2020-2022. The main warring parties signed a peace accord in November 2022 and the situation has remained calm since then. A possible re-emergence of the conflict would affect any further data collection, but as of now, we continue to actively monitor security conditions and no risks of further active conflict have been reported.

In terms of limitations in interpretation, we emphasize that the broader interpretation of the treatment effects in our trial may be somewhat specific to post-conflict settings that have experienced disruption in health services. In addition, the comparison across government and NGO staff in the delivery of psychotherapy services (a type of service that HEWs rarely engage in) may have limited broader implications for the delivery of other forms of health services where HEWs are active and may have more task-specific human capital. That being said, we have aimed to design the evaluation to allow for some generalizable insights.

Ethics

Ethics approval for the study was granted by the Ethiopian Public Health Association (EPHA) on March 8, 2024 (approval number EPHA/OG/195/24), and by the Institutional Review Board of the International Food Policy Research Institute (IFPRI) on February 6, 2024 (approval number PGI-24-0205). Appendix C provides more detail on ethical procedures followed during the baseline and will be followed in subsequent rounds.

Funding and Declaration of Potential Conflicts of Interest

The United States Agency for International development (USAID) Bureau for Humanitarian Assistance has funded the study. It is part of the five-year Resilience Food Security Activity (RFSA) (2021-2026) to Strengthen PSNP Institutions and Resilience (SPIR II) in Ethiopia. World Vision leads a consortium of implementing partners including CARE and ORDA Ethiopia. The International Food Policy Research Institute (IFPRI) leads the SPIR II learning agenda. This research is part of the learning agenda and intends to improve PSNP client's mental health and economic wellbeing using evidence-based action.

The World Bank Africa Gender Innovation Lab has also funded the study and is part of the research team.

All investigators declare the absence of any potential Conflicts of Interest.

References

Abate, M., Mulissa, Z., Magge, H., Bitewulign, B., Kiflie, A., Biadgo, A., Alemu, H., Seman, Y., Woldesenbet, D., Estifanos, A. S., Parry, G., & Quaife, M. (2022). Key factors influencing motivation among health extension workers and health care professionals in four regions of Ethiopia: A cross-sectional study. *PLOS ONE*, *17*(9), e0272551-. https://doi.org/10.1371/journal.pone.0272551

Abay, K. A., Abay, M. H., Berhane, G., Chamberlin, J., Croke, K., & Tafere, K. (2022). Access to health services, food, and water during an active conflict: Evidence from Ethiopia. *PLOS Global Public Health*, *2*(11), e0001015.

Akhtar, A., Koyiet, P., Rahman, A., Schafer, A., Hamdani, S. U., Cuijpers, P., Sijbrandij, M., & Bryant, R. A. (2022). Residual posttraumatic stress disorder symptoms after provision of brief behavioral intervention in low- and middle-income countries: An individual-patient data meta-analysis. *Depression and Anxiety*, *39*(1), 71–82. https://doi.org/10.1002/da.23221

Alderman, H., Melissa Hidrobo, Daniel O. Gilligan, Negussie Deyessa, & Michael Mulford. (2022). *Evaluation of the Group Problem Management Plus (Group PM+) pilot study in Oromia and Amhara Regional States, Ethiopia*. AEA RCT Registry.

Alenko, A., Girma, S., Abera, M., & Workicho, A. (2020). Children emotional and behavioural problems and its association with maternal depression in Jimma town, southwest Ethiopia. *General Psychiatry*, *33*(4).

Angelucci, M., & Bennett, D. (2024). The economic impact of depression treatment in india: Evidence from community-based provision of pharmacotherapy. *American Economic Review*, *114*(1), 169–198.

Ashraf, N., Bandiera, O., Davenport, E., & Lee, S. S. (2020). Losing prosociality in the quest for talent? Sorting, selection, and productivity in the delivery of public services. *American Economic Review*, *110*(5), 1355–1394.

Assefa, Y., Gelaw, Y. A., Hill, P. S., Taye, B. W., & Van Damme, W. (2019). Community health extension program of Ethiopia, 2003–2018: successes and challenges toward universal coverage for primary healthcare services. *Globalization and Health*, *15*, 1–11.

Baranov, V., Bhalotra, S., Biroli, P., & Maselko, J. (2020a). Maternal Depression, Women's Empowerment, and Parental Investment: Evidence from a Randomized Controlled Trial. *American Economic Review*, *110*(3), 824–859. http://www.aeaweb.org/aer/

Baranov, V., Bhalotra, S., Biroli, P., & Maselko, J. (2020b). Maternal Depression, Women's Empowerment, and Parental Investment: Evidence from a Randomized Controlled Trial. *AMERICAN ECONOMIC REVIEW*, *110*(3), 824–859.

Bhat, B., De Quidt, J., Haushofer, J., Patel, V. H., Rao, G., Schilbach, F., & Vautrey, P.-L. P. (2022). *The long-run effects of psychotherapy on depression, beliefs, and economic outcomes*. National Bureau of Economic Research.

Bhat, B., De Quidt, J., Haushofer, J., Patel, V. H., Rao, G., Schilbach, F., Vautrey, P.-L. P., Lawton, I., Xia, H., Ma, Z., Ye, H., Vira, K., & Aggarwal, S. (2022). *The Long-Run Effects of Psychotherapy on Depression, Beliefs, and Economic Outcomes*. http://www.nber.org/papers/w30011

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2018). Experimental evidence on scaling up education reforms in Kenya. *Journal of Public Economics*, *168*, 1–20. https://doi.org/10.1016/j.jpubeco.2018.08.007

Bunn, M., Gonzalez, N., Falek, I., Weine, S., & Acri, M. (2021). Supporting and Sustaining Nonspecialists to Deliver Mental Health Interventions in Low- and Middle-Income Countries: An Umbrella Review. *Intervention Journal of Mental Health and Psychosocial Support in Conflict Affected Areas*, *19*(2). https://journals.lww.com/invn/fulltext/2021/19020/supporting_and_sustaining_nonspecialists_to.3.aspx

Cameron, L., Olivia, S., & Shah, M. (2019). Scaling up sanitation: evidence from an RCT in Indonesia. *Journal of Development Economics*, *138*, 1–16.

Charlson, F., van Ommeren, M., Flaxman, A., Cornett, J., Whiteford, H., & Saxena, S. (2019). New WHO prevalence estimates of mental disorders in conflict settings: a systematic review and meta-analysis. *The Lancet*, *394*(10194), 240–248.

Dawson, K. S., Bryant, R. A., Harper, M., Tay, A. K., Rahman, A., Schafer, A., & Van Ommeren, M. (2015). Problem Management Plus (PM+): a WHO transdiagnostic psychological intervention for common mental health problems. *World Psychiatry*, *14*(3), 354.

Deserranno, E. (2019). Financial incentives as signals: experimental evidence from the recruitment of village promoters in Uganda. *American Economic Journal: Applied Economics*, *11*(1), 277–317.

Deserranno, E., Nansamba, A., & Qian, N. (2024). The Impact of NGO-Provided AID on Government Capacity: Evidence from Uganda. *Journal of the European Economic Association*, jvae029.

Desta, F. A., Shifa, G. T., Dagoye, D. W., Carr, C., Van Roosmalen, J., Stekelenburg, J., Nedi, A. B., Kols, A., & Kim, Y. M. (2017). Identifying gaps in the practices of rural health extension workers in Ethiopia: a task analysis study. *BMC Health Services Research*, *17*, 1–9.

Fisseha, G., Gebrehiwot, T. G., Gebremichael, M. W., Wahdey, S., Meles, G. G., Gezae, K. E., Legesse, A. Y., Asgedom, A. A., Tsadik, M., Woldemichael, A., Gebreyesus, A., Abebe, H. T., Haile, Y. A., Gezahegn, S., Aregawi, M., Berhane, K. T., Godefay, H., & Mulugeta, A. (2023). War-related sexual and gender-based violence in Tigray, Northern Ethiopia: a community-based study. *BMJ Global Health*, *8*(7), e010270. https://doi.org/10.1136/bmjgh-2022-010270

Fitch-Fleischmann, B., & Kresch, E. P. (2021). Story of the hurricane: Government, NGOs, and the difference in disaster relief targeting. *Journal of Development Economics*, *152*, 102702.

Gesesew, H., Berhane, K., Siraj, E. S., Siraj, D., Gebregziabher, M., Gebre, Y. G., Gebreslassie, S. A., Amdeslassie, F., Tesema, A. G., Siraj, A., Aregawi, M., Gezahegn, S., & Tesfay, F. H. (2021). The impact of war on the health system of the Tigray region in Ethiopia: an assessment. *BMJ Global Health*, *6*(11), e007328. https://doi.org/10.1136/bmjgh-2021-007328

Haushofer, J., Mudida, R., & Shapiro, J. P. (2020). The Comparative Impact of Cash Transfers and a Psychotherapy Program on Psychological and Economic Well-being. In *NBER Working Paper Series*. https://doi.org/10.2139/ssrn.3759722

Henderson, J. V., & Lee, Y. S. (2015). Organization of disaster aid delivery: spending your donations. *Economic Development and Cultural Change*, *63*(4), 617–664.

Jackson, C. K. (2012). Single-sex schools, student achievement, and course selection: Evidence from rule-based student assignments in Trinidad and Tobago. *Journal of Public Economics*, *96*(1–2), 173–187.

Jackson, C. K. (2021). Can introducing single-sex education into low-performing schools improve academics, arrests, and teen motherhood? *Journal of Human Resources*, *56*(1), 1–39.

Jane-Llopis, E. V. A., & Matytsina, I. (2006). Mental health and alcohol, drugs and tobacco: a review of the comorbidity between mental disorders and the use of alcohol, tobacco and illicit drugs. *Drug and Alcohol Review*, *25*(6), 515–536.

Jordans, M. J. D., Kohrt, B. A., Sangraula, M., Turner, E. L., Wang, X., Shrestha, P., Ghimire, R., Van't Hof, E., Bryant, R. A., Dawson, K. S., Marahatta, K., Luitel, N. P., & Van OmmerenI, M. (2021). Effectiveness of Group Problem Management Plus, a brief psychological intervention for adults affected by humanitarian disasters in Nepal: A cluster randomized controlled trial. *PLoS Medicine*, *18*(6), 1–22. https://doi.org/10.1371/journal.pmed.1003621

Karachiwalla, N. (2019). A teacher unlike me: Social distance, learning, and intergenerational mobility in developing countries. *Economic Development and Cultural Change*, *67*(2), 225–271.

Kling, J., Liebman, J., & Katz, L. (2007). Experimental Analysis of Neighborhood Effects. *Econometrica*, *75*(1), 83–119.

Lund, C., Orkin, K., Witte, M., Davies, T., Walker, J., Haushofer, J., Murray, S., Bass, J., Murray, L., & Patel, V. (2022). Treating mental health conditions improves labor market and other economic outcomes in low and middle-income countries. *University of Oxford, Working Paper*.

Manaye, A., Afewerk, A., Manjur, B., & Solomon, N. (2023). The Effect of the war on smallholder agriculture in Tigray, Northern Ethiopia. *Cogent Food & Agriculture*, *9*(1), 2247696.

Maselko, J., Sikander, S., Turner, E. L., Bates, L. M., Ahmad, I., Atif, N., Baranov, V., Bhalotra, S., Bibi, A., Bibi, T., Bilal, S., Biroli, P., Chung, E., Gallis, J. A., Hagaman, A., Jamil, A., LeMasters, K., O'Donnell, K., Scherer, E., … Rahman, A. (2020). Effectiveness of a peer-delivered, psychosocial intervention on maternal depression and child development at 3 years postnatal: a cluster randomised trial in Pakistan. *The Lancet Psychiatry*, *7*(9), 775–787. https://doi.org/10.1016/S2215-0366(20)30258-3

Mekonnen, H., Medhin, G., Tomlinson, M., Alem, A., Prince, M., & Hanlon, C. (2020). Impact of child emotional and behavioural difficulties on educational outcomes of primary school children in Ethiopia: a population-based cohort study. *Child and Adolescent Psychiatry and Mental Health*, *14*, 1–10.

Mo, D., Bai, Y., Shi, Y., Abbey, C., Zhang, L., Rozelle, S., & Loyalka, P. (2020). Institutions, implementation, and program effectiveness: Evidence from a randomized evaluation of computer-assisted learning in rural China. *Journal of Development Economics*, *146*, 102487.

Nsibande, D., Loveday, M., Daniels, K., Sanders, D., Doherty, T., & Zembe, W. (2018). Approaches and strategies used in the training and supervision of Health Extension Workers (HEWs) delivering integrated community case management (iCCM) of childhood illness in Ethiopia: a qualitative rapid appraisal. *African Health Sciences*, *18*(1), 188–197.

Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chisholm, D., Collins, P. Y., Cooper, J. L., Eaton, J., Herrman, H., Herzallah, M. M., Huang, Y., Jordans, M. J. D., Kleinman, A., Medina-Mora, M. E., Morgan, E., Niaz, U., Omigbodun, O., … UnÜtzer, Jü. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet*, *392*(10157), 1553–1598. https://doi.org/https://doi.org/10.1016/S0140-6736(18)31612-X

Peters, D. H., Garg, A., Bloom, G., Walker, D. G., Brieger, W. R., & Hafizur Rahman, M. (2008). Poverty and access to health care in developing countries. *Annals of the New York Academy of Sciences*, *1136*(1), 161–171.

Raghunathan, K., Alvi, M., & Sehgal, M. (2023). Ethnicity, information and cooperation: Evidence from a group-based nutrition intervention. *Food Policy*, *120*, 102478.

Rahman, A., Khan, M. N., Hamdani, S. U., Chiumento, A., Akhtar, P., Nazir, H., Nisar, A., Masood, A., Din, I. U., Khan, N. A., Bryant, R. A., Dawson, K. S., Sijbrandij, M., Wang, D., & van Ommeren, M. (2019). Effectiveness of a brief group psychological intervention for women in a post-conflict setting in Pakistan: a single-blind, cluster, randomised controlled trial. *The Lancet*, *393*(10182), 1733–1744. https://doi.org/10.1016/S0140-6736(18)32343-2

Ridley, M., Rao, G., Schilbach, F., & Patel, V. (2020). Poverty, depression, and anxiety: Causal evidence and mechanisms. *Science*, *370*(6522), eaay0214.

Schwartz, J. B., & Bhushan, I. (2004). *Cambodia: using contracting to reduce inequity in primary health care delivery*. World Bank, Washington, DC.

Singla, D. R., Kohrt, B. A., Murray, L. K., Anand, A., Chorpita, B. F., & Patel, V. (2017). Psychological Treatments for the World: Lessons from Low- and Middle-Income Countries. *Annual Review of Clinical Psychology*, *13*(Volume 13, 2017), 149–181. https://doi.org/https://doi.org/10.1146/annurev-clinpsy-032816-045217

Tesfaye, Y., Birhanu, Z., Agenagnew, L., Anand, S., Yitbarek, K., Ahmed, G., Getnet, M., & Tucho, G. T. (2022). Knowledge and attitude of health extension workers regarding mental health problems in Jimma Zone, Ethiopia: A cross-sectional study. *BMJ Open*, *12*(2), e048381.

Tilahun, H., Fekadu, B., Abdisa, H., Canavan, M., Linnander, E., Bradley, E. H., & Berman, P. (2017). Ethiopia's health extension workers use of work time on duty: time and motion study. *Health Policy and Planning*, *32*(3), 320–328.

Tiruneh, M. G., Fenta, E. T., Endeshaw, D., Delie, A. M., Adal, O., Tareke, A. A., Bogale, E. K., & Anagaw, T. F. (2024). Health extension service utilization in Ethiopia: systematic review and meta-analysis. *BMC Health Services Research*, *24*(1), 537.

Usmani, F., Jeuland, M., & Pattanayak, S. K. (2022). NGOs and the Effectiveness of Interventions. *The Review of Economics and Statistics* , 1–45.

Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association*, *18*(6), 3045–3089.

Wang, J., Mann, F., Lloyd-Evans, B., Ma, R., & Johnson, S. (2018). Associations between loneliness and perceived social support and outcomes of mental health problems: a systematic review. *BMC Psychiatry*, *18*, 1–16.

Weldegiargis, A. W., Abebe, H. T., Abraha, H. E., Abrha, M. M., Tesfay, T. B., Belay, R. E., Araya, A. A., Gebregziabher, M. B., Godefay, H., & Mulugeta, A. (2023). Armed conflict and household food insecurity: evidence from war-torn Tigray, Ethiopia. *Conflict and Health*, *17*(1), 22. https://doi.org/10.1186/s13031-023-00520-1

World Health Organization. (2016). *Problem Management Plus (PM+): Individual psychological help for adults impaired by distress in communities exposed to adversity*. World Health Organization.

World Health Organization. (2017). Depression and other common mental disorders. *Global Health Estimates*.

World Health Organization. (2023). *Optimizing the Ethiopian Health Extension Programme: Strategies to address workforce challenges*.

Yamey, G. (2012). What are the barriers to scaling up health interventions in low and middle income countries? A qualitative study of academic leaders in implementation science. *Globalization and Health*, *8*, 1–11.

Appendix

Appendix Section A: Tables

Appendix Table 1: Intervention assignment of villages

Category	Number of kebeles in the category	Total villages	Number of villages per arm					
			Control per kebele	Total Control	T1: HEWs per kebele	Total T1	T2: LFs per kebele	Total T2
2 villages per kebele	5	10	1	5	1	5	0	0
3 villages per kebele	24	72	1	24	1	24	1	24
4 villages per kebele	24	96	1	24	1	24	2	48
5 villages per kebele	5	25	2	10	1	5	2	10
6 villages per kebele	2	12	2	4	2	4	2	4
7 villages per kebele	1	7	3	3	2	2	2	2
Total	**61**	**222**		**70**		**64**		**88**

Appendix Table 2: Summarizing differences between HEWs and LFs

Characteristic	Differences between HEWs and LFs	What type of service provider is plausibly better qualified or positioned along this dimension?
Education level	HEWs generally have around 14 years of education; LFs expected to have < 12 years	HEWs
Community reputation	HEWs are plausibly well-established in their location of service; LFs may be relatively unknown	HEWs
Health sector experience	HEWs are plausibly more experienced in the health sector than LFs	HEWs
Career concerns	HEWs are unlikely to face any explicit rewards or punishments based on their performance in this task; LFs may be more motivated to perform well in the hope of further NGO employment	LFs
Supervision	HEWs are supervised by government supervisors, who may face a range of competing priorities; LFs are supervised by NGO supervisors, who are plausibly better able to prioritize this intervention	LFs
Hour workload (for gPM+)	Consistent	No difference
Compensation level	HEWs have a higher and more stable salary, including benefits; though LFs are compensated commensurately for the contracted hours	No clear hypothesis
Multitasking	HEWs may face many competing requirements within their position and may not prioritize gPM+; LFs may be more able to prioritize gPM+, though they may be engaged in other employment or livelihoods activities they may be engaged in to generate additional income	No clear hypothesis

Appendix Table 3: Cost data collection

	Cost items	Data collection strategy
1	Program implementation activities	Involves expenses on screening and recruiting participating individuals, training HEWs and LFs, and delivering the gPM+ sessions. Personnel costs of data collectors involved in the participant screening and are measured by person-day. Hourly rates measure the personnel costs of HEW and LF for delivering the intervention. The communication costs for participant recruitment, gPM+ session coordination, and related project supervision are accounted for. The team also gathers data on costs incurred for transportation, accommodation, venues, printing materials, and per diem of training participants.
2	Intervention-specific management personnel costs	The personnel costs of intervention-specific management staff such as program managers, monitoring and evaluation officers, admin and finance officers, counselors, HR officers and drivers are measured as the product of their monthly salary, the number of months spent in supporting the activity and the proportion of time spent in supporting the activity per month.
3	Overall program management and monitoring	Other personnel costs of implementing partners' staff and researchers' time in supporting the design and implementation of the program will be measured as the product of their monthly salary, the number of months spent in supporting the activity, and the monthly proportion of time spent in supporting the activity.
4	Capital costs	the capital costs will include the proportion of office buildings and vehicles used for the project period measured at their annual rental rates.
5	Opportunity costs	The average hourly earnings will be extracted from the baseline survey data to measure program participants' opportunity cost of attending the program screening and the gPM+ sessions. The average hourly earnings will be multiplied by the number of hours spent attending the program.

Appendix Section B. Data management

All survey data has been and will be collected with electronic tablets utilizing SurveyCTO, a standard data collection application that allows for secure data storage and options for monitoring data quality. At the end of data collection each day, the entered data is moved to a secure, password-protected server, accessible only by a small number of authorized data managers from IFPRI and the research firm. All information allowing the identification of an individual or household (such as names and addresses) will be stripped from the electronic datasets created for analysis. A separate dataset linking identifiers with information allowing identification of individuals or families (names, addresses, etc.) will be securely retained by the Principal Investigators to be used solely to enable investigators to track and revisit respondents in the end-line surveys.

Facilitator data will be drawn from two surveys: surveys administered directly to the facilitators (pre- and post-training and post-implementation) and monitoring data in the CommCare app, as described above. Surveys administered to the facilitators will follow a protocol parallel to the household surveys. Monitoring data will be sent to IFPRI via password secure datasets. All information allowing the identification of an individual or household (such as names and addresses) will be stripped from the electronic datasets created for analysis.

Appendix Section C: Ethical procedures

This section provides a brief overview of the ethical procedures employed as part of the trial.

a. **Permission from local authorities**: Upon the ethics committee's approval of the protocol, the survey form informed local and regional health authorities and community representatives and requested their approval for all planned survey activities.

b. **Voluntary and informed participation**: The study's objectives and procedures were and will be explained to all participants in their local language (Tigrinya) before commencing any survey. All informants reserve the right to refuse to participate in the study. Each enumerator is asked to read the consent statement slowly in full form and the local language to the participant. The enumerator then asks if the consent statement was understood and if there are any questions. Then, the enumerator asks the participants if they would like to participate in the surveys. At this time, the participant will be allowed to refuse, and it will be clarified that they able to refuse to participate at any time during the interview/discussion without repercussions to them or their family. Verbal informed consent will be obtained after understanding each procedure related to their involvement.

c. **Confidentiality:** Project staff, enumerators, and implementers have been and will be trained to maintain confidentiality while conducting interviews and refrain from any information with any individual other than the project field coordinator. Confidentiality of the data is protected by recording survey interview responses using Computer Assisted Personal Interviewing (CAPI), and thus no hard copy versions of survey questionnaires are available. All files containing raw data are securely stored in password-protected databases. Once all data has been uploaded onto the server, data will be deleted from enumerators' tablets. Access to the complete data is restricted to the authorized data managers from IFPRI, World Bank, and survey firms.

d. **Anonymity:** Care will be taken to ensure the anonymity of participants during data collection, data management, data analysis, and result dissemination. Datasets will be made anonymous by using a unique household ID. The name and geographic location of the respondent will be kept in a separate data file to which only the research team will have access. Anonymized versions of the data set that exclude these personal identifiers will be made available for public access. Personal identifiers will be deleted one year following the closure of the evaluation in 2025.

e. **Suicide and Severe Depression Referral:** All study subjects screened in the study will be provided information on local mental health services. Women and men who show signs of severe depression (PHQ-9 of 20 or above) or suicidal thoughts during the screening, baseline, and end-line surveys will be referred to a hospital for psychiatric treatment. During the screening survey, a safety protocol will assess suicide risk and will result in immediate referral for any subjects considered to be moderate to high risk. Referrals are confirmed through subsequent communication with these households by visit or phone call.

f. **IPV and GBV module and referral:** As per the WHO guidelines for researching violence against women, only one woman per household will be included in IPV and GBV modules.

Female enumerators will interview these women. After the interview, all women will receive information on support services available (World Health Organization, 2005, 2016). De-identified information cards will be offered to all women regardless of disclosure of IPV. These information cards will contain phone numbers for social services available in the woreda, which include the contact information for the Women's and Children's Affairs Office. Enumerators will carefully explain to participants what types of assistance these organizations offer and what they can expect if they contact one of the services. In addition, enumerators will be trained on what to do if a woman approaches them in distress.

g. **Risks and serious adverse events:** The level of risk anticipated for participants in the evaluation is low. There may be some risks of stigma or discrimination for individauls receiving psychosocial support; however, previous experience of the same organizations in implementing similar interventions suggests this risk is low, and WV staff are well-trained to manage these challenges. Given that the intervention also has the potential for significant benefits for individuals experiencing meaningful mental health challenges, the benefits are assessed to outweigh the risks. In-kind compensation (valued at 75 birr) will be provided to all respondents to thank them for the time devoted to the survey itself.