INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
IFPRI

CGIAR FOOD FRONTIERS AND SECURITY

CGIAR POLICY INNOVATIONS

**IFPRI Discussion Paper 02450**

October 2026

# 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
Tigabu Getahun
Daniel O. Gilligan
Adiam Hagos Hailemicheal
Jessica Leight
Michael Mulford
Sarim Zafar

Poverty, Gender, and Inclusion Unit

## INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

The International Food Policy Research Institute (IFPRI), a CGIAR Research Center established in 1975, provides research-based policy solutions to sustainably reduce poverty and end hunger and malnutrition. IFPRI's strategic research aims to foster a climate-resilient and sustainable food supply; promote healthy diets and nutrition for all; build inclusive and efficient markets, trade systems, and food industries; transform agricultural and rural economies; and strengthen institutions and governance. Gender is integrated in all the Institute's work. Partnerships, communications, capacity strengthening, and data and knowledge management are essential components to translate IFPRI's research from action to impact. The Institute's regional and country programs play a critical role in responding to demand for food policy research and in delivering holistic support for country-led development. IFPRI collaborates with partners around the world.

## AUTHORS

Melissa Hidrobo (m.hidrobo@cgiar.org) is a Senior Research Fellow in the Poverty, Gender, and Inclusion (PGI) Unit of the International Food Policy Research Institute (IFPRI), Washington, DC.

Negussie Deyessa (negdaysun@gmail.com) is a Professor of Epidemiology at the University of Rwanda, Kigali, Rwanda.

Tigabu Getahun (tigyget14@gmail.com) is a Lead Researcher at the Policy Studies Institute, Addis Ababa, Ethiopia.

Daniel O. Gilligan (d.gilligan@cgiar.org) is the Director of IFPRI's PGI Unit, Washington, DC.

Adiam Hagos Hailemicheal (ahailemicheal@worldbankgroup.org) is an Economist with the Africa Region Gender Innovation Lab at the World Bank, Washington, DC.

Jessica Leight (j.leight@cgiar.org) is a Senior Research Fellow in IFPRI's PGI Unit, Washington, DC.

Michael Mulford (michael.mulford@kellogg.northwesern.edu) is the Director of the Global Poverty Research Lab at Northwestern University, Evanston, IL.

Sarim Zafar (s.zafar@cgiar.org) is a Research Analyst in IFPRI's PGI Unit, Washington, DC.

**Notices**

[1]IFPRI Discussion Papers contain preliminary material and research results and are circulated in order to stimulate discussion and critical comment. They have not been subject to a formal external review via IFPRI's Publications Review Committee. Any opinions stated herein are those of the author(s) and are not necessarily representative of or endorsed by IFPRI.

[2] The boundaries and names shown and the designations used on the map(s) herein do not imply official endorsement or acceptance by the International Food Policy Research Institute (IFPRI) or its partners and contributors.

[3]Copyright remains with the authors. The authors are free to proceed, without further IFPRI permission, to publish this paper, or any revised version of it, in outlets such as journals, books, and other publications.

**Abstract**

Large treatment gaps in mental health services exist in low- and middle-income countries. While low-intensity psychological interventions delivered by non-specialists show promise in closing this gap, evidence is limited regarding whether government health systems can scale these services effectively. Utilizing a randomized controlled trial across 222 villages in Tigray, Ethiopia, we compare the effectiveness of a psychological intervention, group Problem Management plus (gPM+), delivered in five sessions using two distinct platforms—government-employed Health Extension Workers (HEWs) and NGO-stipended local facilitators (LFs)—vis-a-vis a control arm that receives no intervention. At the one-month follow-up, we find that gPM+ significantly reduces adverse mental health symptoms (particularly depression and anxiety) across both delivery arms, and the intervention is more effective when delivered by LFs (0.22 SD) compared to HEWs (0.11 SD). However, at the 12-month follow-up, treatment effects largely attenuate to zero (mainly due to spontaneous recovery in the control arm) and we find no impacts on child development. The cost per month of depression averted is lower for LFs (PPP $120) vis-a-vis HEWs (PPP $145 to $162, depending on whether HEW time is costed). These findings show that low cost psychological interventions can effectively reduce adverse mental health symptoms in the short term when delivered through both government and NGO platforms, but NGO platforms are slightly more cost-effective.

**Keywords:** Mental health, randomized control trial, NGO vs. government delivery, Ethiopia

**Acknowledgments**

We gratefully acknowledge funding for the impact evaluation from USAID under Cooperative Agreement No AID-FFP-A-16-00008 and the World Bank's Umbrella Facility for Gender Equality. This work was also supported by the donors who fund the CGIAR's Science Program on Policies and Food Frontiers and Security through their contributions to the CGIAR Trust Fund. We would also like to thank Lucy Billings for excellent project management support, EconInsight for careful data collection, and the SPIR team for outstanding collaboration in trial design and implementation.

# 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). These challenges can have meaningful economic implications, affecting investments in children, labor productivity, and individuals' self-perceptions (Angelucci and Bennett, 2024; Baranov et al., 2020; Bhat et al., 2022). Despite this high prevalence, adequate care for mental illness is largely inaccessible in low- and middle-income countries (LMICs) due in part to an acute shortage of skilled staff, particularly in rural areas (Patel et al., 2018). The treatment gap, defined as the percentage of individuals who require care but do not receive it, is estimated at 80% (Ridley et al., 2020).

To reduce the treatment gap, low-intensity psychological interventions delivered by non-specialist providers have been developed and tested. While findings are promising, the majority of these interventions have been implemented at a limited scale (Singla et al., 2017). Wider scaling of health interventions typically relies on integration in government health systems (Yamey, 2012). Yet, a growing literature has generated evidence that across multiple sectors, government-delivered interventions delivered at scale achieve meaningfully smaller effects vis-'a-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 questions of whether government health systems can effectively deliver frontline mental health interventions in rural low-income settings and how government delivery compares with NGO-led delivery.

To answer these questions, we implemented a randomized controlled trial (RCT) in which 222 villages across four districts in Tigray, Ethiopia were randomly assigned to three study arms. The first arm serves as a control arm and did not receive any targeted intervention; the second arm received a psychological intervention, group Problem Management plus (gPM+), delivered by government-employed health extension workers (HEWs); and the third arm received gPM+ delivered by local facilitators (LFs) engaged and compensated by an NGO. We test these delivery models in the aftermath of a devastating two-year civil war that generated severe, adverse impacts on food security, health, and livelihoods in Tigray (Abay et al., 2022; Weldegiargis et al., 2023). Trauma and mental distress were extremely high in the region, rendering the need to find scalable solutions for the provision of psychological support even more acute.

Developed by the World Health Organization to address mental health challenges in populations affected by adversity (World Health Organization, 2018), gPM+ is a low-intensity intervention that consists of five sessions that teach participants effective

strategies to manage emotional and practical challenges (Bryant et al., 2017; Dawson et al., 2015). gPM+ is designed to be implemented by non-specialists, allowing it to be affordably scaled up in a context of limited resources (Akhtar et al., 2022). While a growing body of evidence suggests that gPM+ is promising for addressing a range of mental health challenges in the short term, prior studies did not deliver gPM+ at scale (Schaefer et al., 2023)[1].

The motivation for implementing gPM+ through HEWs is that it can be implemented at scale. HEWs are already present in all communities and could potentially absorb this new task without additional compensation. Whether HEWs can deliver gPM+ effectively is not obvious, however. HEWs are already responsible for a broad portfolio of health services and may lack the time to add a five-session group intervention, or may prioritize it below existing duties, compromising fidelity, session frequency, or participant attendance. At the same time, their standing in the community and their experience delivering health services may position them well to recruit participants and sustain engagement. Whether they can deliver it as effectively as NGO delivery is also unclear since the two sets of service providers differ along multiple dimensions including employment contracts (HEWs have a fixed contract while LFs are engaged for a short-term period and paid a stipend), supervision structures, prior work experience, and work responsibilities. On the one hand, LFs may be more effective at delivering gPM+, if short term contracts sharpen incentives to perform well to obtain new contracts and if fewer competing demands from work allow greater focus on gPM+. On the other hand, HEWs may be more skilled and experienced and they may have more convening power and established rapport with community members. HEWs 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.

We estimate the effects of gPM+ delivered by HEW (compared to a control), the effects of gPM+ delivered by LFs (compared to control) and the relative effectiveness of HEW and LF delivery. Because HEW delivery is expected to be less costly at the margin (given no additional compensation for the task) while LF delivery requires paid facilitators, differences in effectiveness alone do not determine which model is preferable; we therefore complement the effectiveness comparison with a cost-effectiveness analysis. Finally, because all HEWs in Ethiopia are women, while LFs could be women or male, we further randomized the gender of the LF at the village level. Accordingly, we will also assess how effects vary with provider gender and with the match between provider and participant gender.

The sample in this trial is constituted by individuals who were beneficiaries of Ethiopia's

[1]In the metanalysis of gPM+ conducted by (Schaefer et al., 2023), no studies enrolled more than 1000 participants. A more recent study conducted in Ethiopia enrolled 3,055 participants and was implemented through an NGO.

Productive Safety Net Program (PSNP) and were thus among the poorest 15% of households in rural Ethiopia; in addition, they were identified as eligible for therapy based on the prevalence of moderate to moderately severe depressive symptoms at baseline. The trial enrolled 3,477 men and women who were surveyed at three time points; before implementation of gPM+ (baseline), one month following the scheduled conclusion of gPM+ (one-month survey), and 12 months following scheduled conclusion of gPM+ (12-month survey).

The primary outcomes of interest, as specified in the pre-analysis plan (PAP), include mental health and child development. Given the flexible and transdiagnostic structure of gPM+ and the high rates of multiple co-occurring mental health disorders in post-conflict settings (Charlson et al., 2019), we measure depression, anxiety, stress, and post-traumatic stress. Child development is measured at 12 months using parental reports of child behavioral and emotional problems, and is included as a primary outcome because enhanced mental health is hypothesized to enhance the quality and the level of parent's investment in children, thereby improving their socioemotional development (Baranov et al., 2020; Angelucci and Bennett, 2024; Bryant et al., 2022a,b). In addition to the primary outcomes, we explore a range of secondary outcomes that correspond to both channels for potential improvements in mental health and benefits of improved mental health; these include psychosocial, behavioral, cognitive, economic, and childcare outcomes.

We find that in the short-run — one month post-treatment — gPM+ is highly effective in reducing adverse mental health symptoms across both the LF and HEW delivery platforms. However, the decline in mental health symptoms is around twice as large when gPM+ is delivered by LFs (0.22 standard deviations) compared to when it is delivered by HEWs (0.11 standard deviations), a difference that is statistically significant. In both subarms, the intervention is also more effective when delivered to women. The enhancement in mental health is primarily driven by a reduction in symptoms of depression and anxiety, with more modest effects on stress and post-traumatic stress. At the twelve month follow-up, however, the treatment effects observed in the short run have largely attenuated to zero across both subarms, due to a robust pattern of spontaneous recovery in the control arm: the reduced symptoms of anxiety, depression, and stress in the treatment arms persist, but individuals who never participated in gPM+ show similar reductions in these adverse symptoms by 12 months. This suggests that gPM+ accelerates recovery, rather than permanently shifting individuals' trajectories. We find little evidence of any effects in child development at 12 months, though there is some weak evidence of deteriorating scores in the LF arm.

For psychosocial outcomes, we see a broadly similar pattern of initial positive effects attenuating over time. At one month, respondents exposed to gPM+ show evidence of higher self-efficacy, better daily functioning (reductions in impairment linked to mental health challenges), better management of tension and coping with stress, and greater

social support; these effects are quite similar for gPM+ as delivered by both HEWs and LFs, but in both subarms, the positive effects fade over time. We observe no significant shifts in the short-or longer term for other outcomes including behavioral outcomes (alcohol use, intimate partner violence), cognitive skills, economic outcomes (the probability of engaging in any productive activity as well as food security and savings), or childcare outcomes (including the incidence of child neglect and violent discipline).

Despite the larger treatment effects achieved by LFs, we see relatively few differences in fidelity of implementation comparing across the two subarms: while female LFs do conduct slightly more sessions than HEWs, male LFs do not. Conditional on a session being offered, attendance is higher for the HEW arm, but unconditional attendance is consistent across subarms. Interestingly, male participants report lower satisfaction with gPM+ as delivered by HEWs, and both men and women are somewhat less likely to report prior interaction with HEWs than with local facilitators, a pattern that could be consistent with the reduced effects of HEW-provided therapy. We also present some suggestive evidence that the persistent engagement of facilitators in delivery of gPM+ over the long term may vary by modality: following the conclusion of the official trial period, HEWs largely ceased providing gPM+, while LFs show a continued and roughly consistent level of engagement.

Finally, we conduct a detailed cost analysis (all cost estimates are in 2021 PPP dollars), and find that the cost per beneficiary for gPM+ is relatively low using both delivery mechanisms (between \$90 and \$100 for the HEW arm, depending on whether HEW time is valued, and \$127 for the LF arm).[2] This compares favorably to the cost of therapy delivered to adolescents in Uganda at \$150 (Baird et al., 2025), gPM+ delivered in pilot form in Turkey at \$188 (Acarturk et al., 2022), and individual-level PM+ costing more than \$3,200 in Kenya (Haushofer et al., 2020), although individual therapy in India is less costly still at \$75 (Patel et al., 2017; Bhat et al., 2022). In terms of cost-effectiveness, the LFs achieve a lower cost per month of depression averted over the first year (at around \$120) relative to HEWs (\$145–\$162); this is meaningfully higher than the \$39 per month of depression averted implied by the Healthy Activity Program in India (Bhat et al., 2022), but well below the \$462 implied by group therapy for adolescents in Uganda (Baird et al., 2025).

This study contributes 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

[2]Another dimension of cost could be diversion of HEWs from their core responsibility of providing other forms of health services; however, there is no evidence that the probability that HEWs provide other services to sampled individuals declines in the villages assigned to receive HEW-provided gPM+.

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 compared to the government (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 and Kresch, 2021; Henderson and Lee, 2015). 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 and Bhushan, 2004).

Our study also contributes to the larger literature on the effectiveness of psychological interventions 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 et al., 2022; Singla et al., 2017; Rahman et al., 2019; Bryant et al., 2017). Only a handful of studies assess impacts one year or more post-program, 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). Two studies with similar timelines as this one show that initial short-term impacts of gPM+ on mental health outcomes fade by one year (Bryant et al., 2022a,b; Hidrobo et al., 2026).

The mixed findings in the literature also suggest that the delivery platform may meaningfully shape both the initial effects of psychological interventions as well as the sustainability of effects. For example, Baranov et al. (2020) 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 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 et al., 2022).[3] However, this literature as well as the broader literature on frontline health services has generated relatively little evidence directly comparing modalities of delivery in the same study.

Lastly, our paper speaks to the relatively under-explored topic of gender effects with mental health interventions. Studies often exclusively target women, who, in general,

[3]The difference across the two trials analyzed here may be closely linked to variation in the severity and persistence of depression: the Healthy Activity Program (HAP) intervention, showing a significant long-term treatment effect, was targeting a very depressed population where the control arm experienced limited recovery.

experience much higher lifetime rates of depression than men, including in LMICs (Kessler et al., 1993; Salk et al., 2017; Guerra et al., 2016; Bromet et al., 2011). For studies in LMICs that include men, few disaggregate impacts by gender of participants, and those that do generally find limited heterogeneity across gender (Haushofer et al., 2020; Hidrobo et al., 2026; Karyotaki et al., 2022). To our knowledge this study is the first to explore treatment heterogeneity across gender while experimentally varying the gender match between participants and facilitators (for men), whereas previous studies generally impose gender-matched facilitation by design. We find stronger impacts among women participants who exclusively received the intervention from female HEWs or LFs. Among men, for whom facilitator gender was experimentally varied, male LFs perform better than female LFs or HEWs. However, the gender gap in effects is not uniform across dimensions of mental health: reductions in moderate to severe depression are of comparable magnitude between men and women, while reductions in post-traumatic stress and anxiety are concentrated entirely among women (leading to larger effects on an overall mental health index among women).

A wider scaling of psychological interventions at the population level requires a better understanding of how to select, train, supervise, and compensate a large number of non-specialist providers (Bunn et al., 2021). Our trial informs this debate by providing evidence that suggests both NGO and government delivery can be effective in the short-term; however, NGO supported delivery is both more effective and more cost-effective.

# 2 Experimental design

## 2.1 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 a multidimensional crisis. The onset of conflict led to widespread and near-complete disruption of food systems, leading to high rates of food insecurity (Weldegiargis et al., 2023). Civilian populations experienced widespread direct exposure to violence and extremely high reported rates of war-related sexual and gender-based violence, estimated to have affected nearly half of all women in the region (Fisseha et al., 2023). While the consequences of the violence in terms of mental health and psychosocial outcomes remain largely unmeasured, they are anticipated to be severe. The conflict also led to widespread disruptions of health services (Gesesew et al., 2021), leading to a treatment gap that (while largely unmeasured) is presumed to be large.

Within this context of widespread conflict-related trauma, identifying strategies to provide psychosocial support at scale is particularly important. Ethiopia's National

Mental Health Strategic Plan advocates for the integration of mental health into the broader health system with a focus on primary health care and community-based approaches, including delivery by HEWs (Federal Ministry of Health, 2020). HEWs are government employees deployed in rural areas as part of a nationwide network to deliver health services[4]; they are primarily women who are recruited locally, receive a year-long training (Nsibande et al., 2018),[5] and then provide services at health posts and through home visits and community outreach. In rural areas, each health post is staffed with two HEWs that serve a subdistrict with a population of approximately 3,000 to 5,000 people (Haile Mariam et al., 2023)

Despite the large-scale penetration of HEWs, there remain meaningful challenges to their ability to effectively deliver 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 (Haile Mariam et al., 2023). Some evidence suggests that HEWs devote a substantial amount of time to administrative activities and waiting for clients, and relatively little time to direct service provision (Tilahun et al., 2017), perform required tasks infrequently and/or are inadequately prepared for those tasks (Desta et al., 2017); and achieve low rates of service utilization (Tiruneh et al., 2024). HEWs were also directly affected by the violence and trauma linked to the recent conflict (Gesesew et al., 2021), with many HEWs experiencing displacement (Tesema and Kinfu, 2021). This raises important questions around the efficacy of using HEWs to scale additional frontline health services, particularly in this post-conflict setting.

## 2.2 Intervention and delivery modalities

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, 2018).[6] 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

[4]The Health Extension Program (HEP) was launched in 2003 to address significant health challenges, especially in rural and underserved areas of Ethiopia (Assefa et al., 2019). They deliver health services under four themes: family health, disease prevention and control (including mental health), hygiene and environmental sanitation, and health education and communication.

[5]HEWs are generally from the same district, but not necessarily from the same community where they serve.

[6]gPM+ consists of five weekly sessions that include 1) understanding PM+ and managing stress; 2) managing problems; 3) get going, keep doing; 4) strengthening social support; and 5) staying well and looking forward.(World Health Organization, 2020)

effectiveness and feasibility (Jordans et al., 2021; Rahman et al., 2019). A previous trial implemented by the same team also demonstrates that gPM+ delivered in rural Ethiopia was effective in the short-term (Hidrobo et al., 2026).[7]

The specific objective of this trial is to test 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 an NGO, World Vision, to deliver gPM+, but did not receive any supplemental compensation for this assignment as it falls under their scope of work.[8] The second entails LFs recruited for a short-term assignment by World Vision. While all HEWs were women as is typical in Ethiopia, both male and female LFs were recruited. LFs received a monthly payment for four months, compensating them for their time devoted to training, practice, preparation, and group facilitation.[9] The stipend provided was about 2,600 ETB (Ethiopian birr; $45) per month, or roughly one third of a monthly HEW salary.

In terms of implementation, all facilitators (both HEWs and LFs) participated jointly in a single 10-day training. Following training, HEWs and LFs were supervised separately by their respective (government and NGO) supervisors, though all were also monitored via a CommCare Connect application developed by Dimagi to facilitate the implementation of the intervention. The implementation of gPM+ entailed five 90–120 minute group sessions delivered weekly over five weeks, and three additional monthly follow-up sessions. Each facilitator was assigned to two groups, including up to eight individuals each. Thus, the total anticipated workload for each facilitator was around 6—8 hours weekly for a period of five weeks (three to four hours per group).[10] Female facilitators (whether HEWs or LFs) facilitated one group of men and one group of women, while male LFs facilitated two groups of men.[11]

## 2.3 Sample and randomization

The evaluation sample was drawn from four districts in eastern Tigray (Samre, Tsaeda Emba, Keletawelaelo, and Enderta) and comprised 222 villages across 61 sub-districts; Figure A1 in the Appendix summarizes the geographic position of the sample districts within Tigray and Ethiopia at large. Randomization was conducted at the village level

[7]The previous study was implemented in the Amhara and Oromia regions and featured a distinct experimental design, research objectives, and target population (whom had lower rates of depression as the cutoff for inclusion to the study was lower). That study also documented that short-term treatment effects attenuated over time.

[8]Government policy generally prohibits any supplemental compensation for HEWs to undertake specific assignments.

[9]LFs received a fifth payment if they completed the three additional monthly follow-up sessions, and in practice, virtually all did complete the sessions and received the fifth payment; this is accordingly costed as part of the LF implementation cost, as described in more detail below.

[10]This was anticipated to include 1.5–2 hours for each session, in conjunction with travel and preparation time as well as time allocated to provide reminders to participants.

[11]Delivery of therapy by a male facilitator to female clients was considered inappropriate in this context.

(the sub-administrative unit below the sub-district), stratified by sub-district; within each sub-district, villages were assigned to receive no intervention or gPM+ delivered by HEWs or LFs.[12]. Within the LF arm, there was a further sub-randomization in which villages were randomly assigned to either a male or female LF as captured in Figure 1.

To be eligible for the study individuals had to be PSNP clients, head of the household or spouse, aged 18-64 years, and exhibit signs of moderate or moderately severe depression as measured by the Patient Health Questionnaire-9 (PHQ-9), corresponding to a score of between 10 and 19 on the scale.[13] Individuals experiencing severe depression or experiencing suicidal symptoms were referred directly to psychiatric care and were not screened into the intervention. Both men and women were screened for eligibility in all villages except those assigned to the male LF arm where only men were screened. The screening process entailed using a random listing of PSNP households (public works and direct support clients) and screening until the target number of individuals of each sex was met for each village[14]

Figure A3 summarizes the sample. At baseline, 6,208 PSNP clients were assessed for eligibility[15], and 3,477 (56%) were identified as eligible (PHQ-9 score between 10 and 19, inclusive) and entered the trial. This reflects a post-conflict population characterized by severe psychological distress; with depression levels substantially exceeding those documented in typical conflict-affected settings Charlson et al. (2019). This extreme prevalence aligns with the multidimensional crisis detailed in Section 2, where the compounding effects of widespread exposure to violence, catastrophic food system collapse, and the near-total disruption of health services have left a deeply vulnerable population facing a massive mental health treatment gap.

## 2.4 Data

Data was collected at baseline (March – May 2024); one month post-intervention (July 2024); and twelve months post-intervention (July 2025). (Figure A2 summarizes the study timeline.) Fieldwork was conducted by EconInsight using teams of trained, gender-matched enumerators. Key modules at both baseline and follow-up included household socioeconomic data as well as individual-level modules measuring depression, stress, anxiety, coping, self-efficacy, childcare and development, savings, gender based violence,

[12]Given that the number of villages per sub-district differs across sub-districts, the exact assignment ratio varied. See pre-analysis plan for more details

[13]The criteria requiring head of household or spouse was imposed given our interest in measuring economic effects of the intervention, as we hypothesized this channel was likely most relevant for economically active adults.

[14]We also stratified our sample by type of PSNP client to ensure 75 percent public works and 25 percent direct support were reached across all intervention arms.

[15]The sample listing included 16,872 PSNP households, of which 10,201 were successfully tracked and 6,208 were eligible (i.e, household head or spouse 18–64 years, of the right sex for the specific village), available, and consented to be screened for depression, and if eligible entered the study.

and time use/ labor participation; at baseline we also collected information on conflict exposure. All screened study subjects received information on locally available mental health services, and all women who were administered the gender based violence module received information on support services available in the area. As summarized in Figure A3, attrition was minimal in both follow-up surveys: 1% at one month and 2% at twelve months.

Surveys were also conducted with the facilitators (both LFs and HEWs) prior to the training, immediately post-training, and at the twelve month follow-up. These instruments included 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); and their engagement in gPM+ itself. We also draw on administrative data tracking facilitator activity (sessions conducted and attendance) that was recorded in the Commcare application, and worked with implementing partners to develop a cost data collection platform to track implementation costs.

All subjects provided informed consent for data collection. Ethical approval for the trial 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).

## 2.5 Primary and secondary outcomes

Primary outcomes were pre-specified in a detailed protocol registered on the American Economic Association registry (trial ID AEARCTR-0013922), and are described in Table A1. The primary 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 PTSD Checklist for DSM-5 (PCL-5) instrument, anxiety as measured by the Generalized Anxiety Disorder 7 (GAD-7), and child development as measured using the Strengths and Difficulties Questionnaire (SDQ). For mental health, we report the raw scores and create a summary index following Kling et al. (2007), calculating the simple unweighted mean across the four standardized mental health outcomes. For child development, we also report both the raw and standardized SDQ scores. In both cases we standardize using the mean and standard deviation (SD) observed in the control arm at baseline and higher values correspond to more adverse outcomes (more symptoms of poor mental health or behavioural challenges).

Secondary outcomes include a range of variables corresponding to both channels for potential improvements in mental health (such as enhanced coping skills or social support)

and benefits of improved mental health (such as improved economic activity, enhanced childcare practices, reduced prevalence of adverse behaviors, and better cognitive performance). We categorize secondary outcomes into five domains; economic, child care, behavioural, psychosocial, and cognitive, with more details provided on each domain in Table A2. The economic domain includes measures of savings, food insecurity, and engagement in economic activities; the childcare domain includes measures of childcare activities, child discipline, and child neglect; the behavioural domain includes measures of alcohol consumption and intimate partner violence; the psychosocial domain includes measures of self-efficacy (the General Self-Efficacy Scale, GSE), functional impairment (the World Health Organization Disability Assessment Schedule, WHODAS), tension reduction, coping strategies (the Brief COPE inventory), and perceived social support (the Multidimensional Scale of Perceived Social Support, MSPSS); and the cognitive domain is measured using a forward digit span task.[16] Similar to our primary outcomes, we report both the raw scores and create a summary index at the domain level whereby we calculate the simple unweighted mean across the standardized outcomes within the domain. Where needed, we reverse code raw scores so that higher scores indicate better outcomes.

Detailed power calculations were conducted using the primary outcomes and are reported in the pre-analysis plan. In general, the minimum detectable effects for primary outcomes of interest for pairwise comparisons across arms range between 0.15 and 0.2 SDs, suggesting the trial is adequately powered to detect relatively small effects of the interventions.

# 3 Empirical analysis

## 3.1 Specification of interest

We estimate the treatment effects for individual-level variables using analysis of covariance (ANCOVA) models capturing intent-to-treat effects. The primary equation of interest compares outcomes of individuals assigned to the HEW and LF arms to outcomes of individuals assigned to the control arm. Given that the male LF subarm is characterized by a male-only sample (in contrast to the other arms that all include both men and women), this specification includes a full set of interactions with gender of the

[16]Secondary outcomes were also pre-specified; however, the domain categorization was updated following data collection in some cases to better reflect the relevant underlying concepts. For example, cognitive and non-cognitive skills (including behavioral measures) were initially categorized as a single category; subsequently, we identified that many of the indicators captured under non-cognitive skills were direct channels through which gPM+ would improve mental health. These skills were part of the gPM+ curriculum, and are now labeled "psychosocial skills" and reported separately. Cognitive skills and behavioral outcomes are considered outcomes of improved mental health and thus are reported in two distinct domains.

participant in order to abstract from any bias introduced by the varying gender sample composition.[17]

$$
\begin{aligned}
Y_{ivk,t} = \beta_0 &+ \beta_1 \, \text{HEW}_{vk} \times \text{Female}_{ivk} + \beta_2 \, \text{HEW}_{vk} \times \text{Male}_{ivk} \\
&+ \beta_3 \, \text{FLF}_{vk} \times \text{Female}_{ivk} + \beta_4 \, \text{FLF}_{vk} \times \text{Male}_{ivk} \\
&+ \beta_5 \, \text{MLF}_{vk} \times \text{Male}_{ivk} + \beta_6 \, \text{Male}_{ivk} + \gamma \, Y_{ivk,t=0} \\
&+ \chi_k + \varepsilon_{ivk,t}
\end{aligned} \tag{1}
$$

Here, $Y_{ivk,t}$ denotes the outcome of interest for an individual $i$ in village $v$ in sub-district $k$ at time period $t$ (either 1 month or 12 month follow-up); $Y_{ivk,t=0}$ denotes the outcome at baseline; and $\chi_k$ denotes sub-district strata fixed effects. (For outcomes that were not measured at baseline, the relevant control is omitted; this point is denoted in the tables.) $\text{HEW}_{vk}$ is a binary variable equal to one if a village was randomized to the HEW arm; $\text{FLF}_{vk}$ is a binary variable equal to one if a village was randomized to the female LF arm; $\text{MLF}_{vk}$ is a binary variable equal to one if a village was randomized to the male LF arm. $\text{Male}_{ivk}$ and $\text{Female}_{ivk}$ refer to the gender of the participant. Standard errors are clustered as the village level.

$\beta_1$ through $\beta_5$ capture gender-specific treatment effects for each arm. We also calculate the average treatment effect for HEWs (the mean of $\beta_1$ and $\beta_2$), and a weighted average treatment effect for LFs in which the gender composition is parallel to the HEW arm (i.e., half men and half women): we construct the latter weighted average effect as $0.5\beta_3 + 0.167\beta_4 + 0.333\beta_5$, with the weights on $\beta_4$ and $\beta_5$ capturing the percentage of men who receive gPM+ as delivered by female and male LFs, respectively.[18] We also report the simple average treatment effect for female LFs (the mean of $\beta_3$ and $\beta_4$), and report the p-values corresponding to tests of equality across the average HEW and LF effects.

For aggregate indices, we do not adjust for multiple hypothesis testing. However, when we analyze outcomes for each individual indicator included in the aggregate indices, we also report sharpened q-values that are adjusted for multiple hypothesis testing. For these individual indicators we correct for multiple tests of one coefficient across all variables within a domain and survey round and report the Anderson sharpened q-values.

[17] This specification is the fully interacted model and reflects equation (2) in the registered PAP. We also estimate a simpler specification where we pool the female and male LF sample (similar to equation (1) in the PAP), but focus on the fully interacted model given the varying sample composition in the male LF arm.

[18] Given that each male LF has two gPM+ groups including men and each female LF has only one gPM+ group including men (and one gPM+ group including women), two thirds of the men in the LF arm are served by male LFs.

## 3.2 Baseline characteristics and balance: Household survey

Appendix Table A3 reports summary statistics to characterize the 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. The penultimate column tests the joint significance across treatment-sample categories for each baseline characteristic. The final row of the table reports a p-value corresponding to the test of joint significance across baseline characteristics within each treatment sample category.

On average, respondents in the control group are 45 years old and the level of human capital attainment is low, with respondents reporting less than three years of education on average. Around 60% of respondents are married, and the average household size is around six members. Respondents in the control group have interacted with a HEW only twice in the past year, suggesting (presumably) a relatively low level of familiarity with the HEWs themselves.

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 consistent with the eligibility criteria requiring a PHQ-9 score between 10 and 19 to enter the sample), and there is also a non-trivial prevalence of anxiety, stress, and post-traumatic stress; the average child SDQ score is nearly 10. The observed intensity of mental health symptoms is also consistent with the high reported rates of recent conflict experiences: about a fifth of households in the control arm report that they were temporarily displaced for at least a month, and about 80% report an armed group was present in their village during the conflict. Most strikingly, 23% 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; the joint tests fail to reject the hypothesis of balance across all arms for every variable except two (years of education and presence of an armed group). As a complementary check, we also implement an omnibus balance test following Kerwin et al. (2024), regressing treatment assignment on the full set of baseline characteristics jointly and reporting a randomization-inference p-value (correcting for the over-rejection that sampling-based joint tests exhibit when many covariates are included, as reported in the notes to Appendix Table A3). This omnibus test does not reject the null of balance for the pooled treatment (p=0.11) or for the HEW arm (p=0.30), but it does indicate some joint imbalance in the LF arm (p=0.008); because our primary specifications control for the baseline value of each outcome, we assess that any influence of this imbalance on the estimated treatment effects is likely to be limited. As a further check, we will subsequently present estimates fro the primary findings augmenting the baseline controls

with covariates selected by the post-double-selection lasso (Belloni et al., 2014).

## 3.3 Facilitator characteristics

Table A4 summarizes characteristics of the facilitators, as reported in the pre-training survey. The first panel summarizes basic demographic characteristics; the second panel reports variables specific to the sample of LFs. 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) and have slightly less education (almost 13 years compared to around 14 years for LFs). Consistent with their older age and more established professional trajectory, HEWs are significantly more likely to be married than female LFs (though not male LFs), 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.[19] There is no evidence of any significant differences in health worker motivation or in choices in the dictator game: across all subarms, participants chose to retain approximately 40%–50% of their endowment and donate the balance. However, HEWs did score meaningfully higher on the Inclusion of Others in Self (IOS) scale, suggestive of greater pro-sociality.

In Panel B, we observe that the average LF has less than six months prior 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 30–41% reported that they were unemployed or out of the labor force prior to taking up this role. (Though not reported in the table, the average HEW has 12 years of work experience as a HEW, but less than a year in any other role.)

Overall, 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 limited evidence of any differential pro-social motivation, suggesting that experience in the health sector has not generated a cadre of HEWs with any noticeable difference in motivation compared to a set of young and almost entirely inexperienced LFs.

[19] We measure self-efficacy using the General Self Efficacy Scale and general knowledge of mental health using the Mental Health Knowledge Schedule (MAKS) as previously adapted in Ethiopia (Tesfaye et al., 2022).We measure health worker motivation using survey questions from a 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).

# 4 Results

## 4.1 Primary outcomes

Table 1 presents the findings for the primary outcomes: the mental health summary index at one month (column 1) and twelve months (column 2), and children's standardized SDQ score at twelve months only (column 3). (Note the indices are standardized relative to the control mean at baseline; we also report the contemporaneous control mean in the table, and the negative values for the contemporaneous control mean accordingly capture the fact that mental health is also improving, with adverse symptoms declining, in the control arm over time.) In the short-term, gPM+ as delivered by both HEWs and LFs leads to a decrease in reported symptoms of poor mental health: the average effect is around 0.11 SDs for HEWs and 0.2 SDs for LFs on average (when weighting to account for differential gender composition) and for female LFs. Thus LFs are around twice as effective, and these differences are statistically significant: $p = 0.009$ when comparing HEWs to LFs on average, and $p = 0.029$ when comparing HEWs to female LFs.

Examining the coefficients by subarm, we also observe meaningful heterogeneity with respect to participant gender: women show much larger decreases in symptoms of poor mental health (between 0.17 and 0.35 SDs), compared to men (between 0.05 and 0.1 SDs). Male beneficiaries show slightly larger effects in the male LF subarm (0.1 SD, statistically significant at the ten percent level), vis-a-vis the female LF subarm (0.07 SD, not statistically significant) or (always female) HEW subarm (0.05 SD. not statistically significant), suggesting that a better gender match enhances effects for male participants, but overall, the treatment effects for men are meaningfully smaller. This underlies the importance of comparing HEWs and LFs using samples characterized by parallel gender composition as in the weighted LF average constructed here, given that LFs (on average) serve more men than HEWs, and the treatment seems to be less effective for men.

Moving to the longer-term effects, Column (2) reveals that these promising treatment effects do not persist, as the coefficients of interest have generally declined to zero in the longer-term follow-up. Interestingly, for HEWs the coefficient is now slightly positive (suggestive of slightly worse mental health outcomes), while for LFs, there are persistent small effects of around 0.06 to 0.08 SDs; though none of these effects are significantly different from zero, the difference between the HEW and LF effects is statistically significant at the ten percent level, weakly consistent with the short-term evidence. There is similarly little evidence of meaningful effects on child development as measured by the SDQ score; if anything, the coefficients are weakly positive (suggestive of more behavioral challenges for children of individuals exposed to gPM+).[20] Table A5 in the Appendix reports the same findings in a simpler specification reporting the treatment

[20] Table A6 reports the treatment effects for each category of the SDQ score.

effects for HEWs and LFs (pooling the effect for female and male LFs serving men), as well as average effects for female and male respondents, where a test of equality confirms that the estimated short-term treatment effects are significantly larger for women than men.[21]

We also explore the effects for each indicator of mental health separately in Tables 2 and 3, for the one month and twelve month follow-up, respectively. Table 2 shows that the decline in mental health challenges in the short-term is driven by declining depressive symptoms (Column 1 and 2) and anxiety (Column 5). There are some noisily estimated reductions in stress, albeit significant in the weighted LF effects (Column 3), while the only shift in post-traumatic stress symptoms is observed for gPM+ delivered by LFs to women (Column 4). For depressive symptoms, the decline is statistically significant for gPM+ delivered to both men and women by both LFs and HEWs, but again the decline in moderate and severe depression achieved by LFs (13.5 percentage points) is nearly 60% larger than the corresponding decline achieved by HEWs (8.6 percentage points), and this difference is statistically significant. For anxiety, the declines in symptoms are observed for women only served by both LFs and HEWs.

The findings reported in Table 3 for the 12 month follow-up unpack in more detail the point already noted in the summary index: in general the short-term effects do not persist over time, but there is some weak evidence of persistent effects (albeit smaller) on depression. For depression, the findings in Columns (1) and (2) suggest a weakly persistent average effect for LFs (around a four percentage point decline in moderate to moderately severe depression), and the difference between the average HEW and average female LF effects is weakly significant in Column (1). In Columns (3) through (5), the coefficients are varied in sign and generally not significant, with some evidence of adverse effects on post traumatic stress (higher stress) in the HEW arm.[22] None of the estimated effects are significant using sharpened q-values, but the conjunction of weakly improved outcomes in the LF arm and weakly deteriorating effects in the HEW arm is consistent with the overall pattern observed in the summary index.

Examining the means in the control arm highlights the pattern of spontaneous remediation that is observed over time: while all respondents (including respondents in the control arm) were characterized by moderate to moderately severe depression at baseline when screened, in the short-term follow-up the incidence of moderate to moderately severe depression is only 44% in the control arm (more than half of these individuals are

[21] In the pre-analysis plan, this simpler specification was identified as primary, but the substantial difference in treatment effects observed for male and female beneficiaries ultimately led us to conclude that the fully saturated specification is more informative.

[22] This is consistent with some evidence in the literature that shorter, psychological debriefing sessions can have adverse effects on post-traumatic stress, though this evidence is entirely from high-income countries (Rose et al., 2002). While gPM+ was designed as a broader and more flexible intervention for populations facing adversity, the risk of adverse effects on a population characterized by elevated levels of trauma may not have been fully eliminated.

no longer identified as depressed), and by the twelve month follow-up, the incidence is only 32% in the control arm and similar across treatment and control. In other words, the vast majority of respondents have recovered from depression one year later, and the probability of recovery is no higher in the LF or HEW arms.

This pattern is also captured graphically in Figure 2, showing the longitudinal trends over time in the four continuous mental health outcomes of interest (depression, stress, post-traumatic stress, and anxiety). Between baseline and the first time period (the one month follow-up), symptoms decline much more rapidly in the arms offered gPM+ (both HEW and LFs, and both men and women) compared to the control arm, and somewhat more rapidly in the LF subarm. However, between the one month and twelve month follow-ups, the control arm continues to decline in symptoms while the treatment arms level off, generating convergence at twelve months. Overall, the level of recovery is notable: depressive symptoms have declined by just under 40% (from 12.95 to 8.13 in the control arm), stress symptoms have declined by 10–15%, post-traumatic stress symptoms have declined by about 10%, and anxiety symptoms have declined by about 30%; all of these shifts are consistent across the treatment and control arms.

## 4.2 Secondary outcomes

Tables 4 and 5 report the summary effects for five families of secondary outcomes: psychosocial, behavioral, cognitive, economic, and childcare. We observe that the primary significant effects generated by gPM+ are on enhanced psychosocial outcomes in the short-term, where positive and significant effects are observed symmetrically for gPM+ as delivered by both HEWs and LFs and for both male and female participants. However, similar to mental health outcomes, these impacts subsequently attenuate over time. For the other outcome families, we find no effects on cognitive, behavioral, and economic outcomes at one month or 12 months. For childcare, there is some weak evidence of positive effects on childcare in the HEW arm at one month, an effect that is significantly different from the corresponding effect generated by LFs. However, at 12 months, there seem to be some adverse effects on childcare, particularly in the LF arm.

For psychosocial outcomes, Appendix Table A7 unpacks the findings in more detail. In the one month follow-up, there is evidence of increased self-efficacy (Column 1), reduced impairment due to mental health challenges (Column 2), better management of tension (column 3), enhanced coping strategies (column 4), and enhanced social support as measured by the Multidimensional Scale of Perceived Social Support (MSPSS) (Column 5); the latter three variables capture concepts directly taught in the gPM+ curriculum. The relative magnitude of these effects are similar across HEW and LF arms. Across men and women, the shift in the World Health Organization Disability Assessment Schedule (WHODAS) and enhanced social support are observed primarily among women, while the

enhanced use of coping strategies (Column 4) is observed primarily among men. However, none of these effects persist in the twelve month follow-up survey, and the longitudinal patterns suggest this is due to reversion to baseline levels in the treatment arms.

Appendix Tables A8 through A10 then capture the experimental effects on individual indicators within the remaining family of outcomes (cognitive, behavioral, economic, and childcare outcomes). There is little evidence of any significant impacts in the short or longer term, consistent with the generally weak impacts in the summary index.

## 4.3 Robustness checks

In addition to direct effects on mental health and other outcomes of interest, we recognize that there may be other benefits or costs associated with delivering gPM+ through HEWs or LFs. For example, HEWs are trained to deliver a large set of health interventions, and engagement in gPM+ may shift their ability to deliver these services (possibly by diverting their time and attention away from other services) or alternatively may make them more accessible by increasing exposure. To test this hypothesis, Table A11 reports estimates for a series of beneficiary-reported measures capturing access to and utilization of health services. We find almost no evidence of any negative effects in the HEW subarm (two coefficients are weakly negative, neither surviving correction for multiple testing, and some coefficients are positive), suggesting that any displacement of relevant services was extremely limited. There is also no evidence that individuals in the control arms differentially accessed HEW services for psychosocial support and thus recovered from symptoms of depression for this reason: use of HEW services is quite consistent across all arms.

Given the evidence previously presented of some imbalance across arms in baseline characteristics, we also estimate an alternate specification for each primary outcome, augmenting the baseline controls with covariates selected by the post-double-selection lasso (Belloni et al., 2014) from a pool of 37 candidate baseline characteristics (spanning demographics, mental health, psychosocial skills, child development, intimate partner violence, economic activity, and prior contact with HEWs). The treatment indicators, respondent gender, the baseline value of the outcome, and baseline-kebele fixed effects are always retained and are not subject to selection. The lasso selects only a small number of additional controls (between one and four per outcome), and the estimated treatment effects presented in Table A12 are essentially unchanged. The estimates at twelve months and for child development are similarly stable in magnitude and sign (the difference between the HEW and LF effects on the mental health index at twelve months is somewhat more precisely estimated, with $p = 0.048$ compared to $p = 0.073$), suggesting that the modest baseline imbalance in the LF arm does not meaningfully shift the estimated treatment effects.

## 4.4 Facilitator quality of implementation

To better understand what shapes the relative effectiveness of the HEW and LF delivery modalities, we also assess a range of variables capturing facilitator quality (see Table A13 for variable definitions). Panel A of Table 6 reports variables linked to fidelity of implementation, reported in administrative data recorded in the CommCare smartphone application used by all facilitators. We observe few meaningful differences in fidelity of implementation comparing across HEWs and LFs as both adhere closely to the gPM+ implementation protocol. Facilitators were mandated to deliver ten sessions (five sessions for each of the two therapy groups), and the mean for HEWs is 9.38 as observed in Column (1); this is higher for LFs, driven by a significant difference for female LFs who deliver virtually every session mandated. Conditional on a session being offered, the LFs achieve slightly lower attendance in percentage terms (driven by lower attendance for male LFs), but the unconditional difference in attendance, reported in Column (3), is insignificant. Overall, attendance is quite high: 85% of participants attend conditional on a session being offered, and the unconditional mean of attendance is 82%.

We also unpack other dimensions of (perceived) facilitator quality as reported by the participants in Panel B of the same table, restricting the sample to the treatment arm. We begin by exploring reported respondent satisfaction (Column 1): while around half of men in the male LF arm reported that they were highly satisfied with their facilitator, this is significantly lower (nine percentage points lower) among men served by HEWs, and at least weakly higher (seven percentage points) among women served by female LFs, consistent with the observed patterns of differential effectiveness in those subarms.

The findings in Columns (2) through (4) unpack the relevance of prior familiarity with the facilitator, and we see that relative to the omitted category (male respondents served by male LFs, for whom 69% report prior interaction), female respondents are less likely to report prior interaction with female local facilitators, and both male and female respondents are less likely to report prior interaction with HEWs. There is little difference in self-reported comfort interacting with the facilitator, as reported in Column (3); but both men and women are much less likely to report they were comfortable because they knew the HEWs-as-facilitators, relative to LFs (Column 4). This pattern, while tentative, could suggest that despite our hypothesis that HEWs would be better positioned to build a rapport with participants, in fact they were viewed as less familiar and less comfortable (particularly by men, consistent with the particularly weak effects of gPM+ when delivered by HEWs to men).

Following the conclusion of the cycle of group therapy that was evaluated as part of this trial, all facilitators were offered the opportunity to attend a refresher training and then continue to deliver gPM+, screening new beneficiaries and forming new groups in

the same village.[23] While evidence around this phase is only suggestive and was not part of our original experimental design, it provides some indicative evidence of how patterns of HEW and LF engagement might evolve over time, and we collected data on this in a facilitator survey at the 12-month follow-up.[24] In Column (1) of Table A14, we observe that attendance at a refresher training (offered following the completion of the first cycle of therapy groups) is consistently high at around 95% for both HEWs and LFs. However, following the conclusion of this training (for which training stipends were offered for attendance), the engagement of HEWs and LFs diverges almost completely: virtually no HEWs continue to form new therapy groups (only 2%), while virtually all LFs do, particularly female LFs (the treatment effects are 84 and 93 percentage points for male and female LFs, respectively). There is a similar gap between HEWs and LFs in column (3) in a motivational score around engagement in gPM+. Follow-up questions posed to HEWs around why they did not continue to engage in gPM+ highlights the absence of compensation for this activity; limited time; and limited support by supervisors.

Overall, these findings suggest a striking pattern: engaging government staff led to implementation of comparable fidelity (and comparable, though somewhat lower) effectiveness in the short-term, during the period in which the trial sample was actively tracked. The lower effectiveness achieved by the HEWs may reflect their weaker relationships with beneficiaries, particularly men. Following the initial round of services, however, HEWs' level of engagement quickly reverts to near-zero, while local facilitators show highly consistent engagement over time.

# 5 Cost and cost-effectiveness

To further deepen our understanding of the two delivery models for gPM+, we also conducted a detailed analysis assessing the cost of the intervention per beneficiary targeted, and the cost per month of depression averted.[25] In addition, for the HEW subarm we

[23]Since randomization was at the village level, new groups would have targeted only other individuals in the same, treated village; individuals in control villages would not have been targeted for the formation of new gPM+ groups.

[24]Only 136 out of the 151 facilitators were surveyed in the follow-up survey for attrition of around 10%.

[25]We use granular data on implementation costs provided by the implementing partner (World Vision), inclusive of staff and supervisory costs. All costs are measured in local currency and then converted to 2021 purchasing power parity-adjusted dollars. We use 2021 as a base year to facilitate comparison to other relevant trials. The conversion proceeds in three steps: costs are converted to birr at the 2024 average official exchange rate (82.60 birr per dollar), converted to international dollars at the 2024 purchasing power parity conversion factor for Ethiopia (27.10 birr per international dollar), and rebased to 2021 prices using the United States gross domestic product deflator; the first two series are drawn from the World Bank World Development Indicators. We note that the birr was floated in July 2024, and accordingly the 2024 average exchange rate is meaningfully higher than the official rate prevailing during most of the implementation period; using a pre-float rate would generate proportionately lower cost estimates.

consider two scenarios. In the first scenario, HEW time invested in the intervention is valued at an hourly rate equivalent to their imputed wage, consistent with a social planner perspective. In the second scenario, HEW time is valued at zero (given that they were not compensated explicitly for this task), mimicking the perspective – for example – of a government considering scale-up of additional health services within the context of an already funded HEW program.

We find that cost per beneficiary targeted is PPP $127.17 in the LF arm, and ranges between $90.22 and $100.39 in the HEW arm, depending on the valuation of HEWs' time. The LF arm was more costly because the hourly rate paid to the LFs was in fact somewhat higher (around 5% higher) than the implied hourly wage earned by a HEW, and was benchmarked deliberately to be slightly higher to facilitate recruitment. To estimate the cost per month of depression averted, we require additional assumptions regarding the time path of depressive symptoms between the short-term and long-term follow-up surveys. We assume a linear decline in symptoms in both the treatment and control arm (interpolating between the levels observed at one month and 12 months), and then calculate the cumulative months of depression averted assuming there are no further effects on depression from 12 months forward.[26] Using these assumptions, gPM+ as delivered by LFs averted around 1.06 months of depression on average, and as delivered by HEWs, around 0.62 months; this implies a cost per month of depression averted of PPP $120 for LFs and up to around $162 for HEWs, as summarized graphically in Figure A4 in the Appendix.

We then benchmark these costs relative to reported costs of other similar interventions in the literature. This includes PM+ implemented in Kenya and gPM+ implemented in pilot form in Turkey (Acarturk et al., 2022) and Colombia (Greene et al., 2024); the Healthy Activity Program (individual therapy) implemented in Goa, India (Bhat et al., 2022; Patel et al., 2017; Weobong et al., 2017); the WHO Thinking Healthy Programme, peer-delivered (THPP), as implemented in Pakistan (Vanobberghen et al., 2020; Gibbs et al., 2025) and delivered to pregnant women; and interpersonal therapy in groups (IPT-G) delivered to adolescents in Uganda (Baird et al., 2025).[27] The pilots do not provide data on effectiveness (only cost per beneficiary), and two studies find null effects on depression

[26]This is a cautious assumption given that there was some weak evidence of continued (small) effects on depression in the LF arm at 12 months, but we use it to generate a conservative estimate of cost-effectiveness. In particular, given the emphasis of gPM+ on building skills and providing strategies to facilitate coping with adversity, we might hypothesize that there could be significant treatment effects on reduced adverse mental health symptoms during future periods of stress, but we cannot substantiate this effect here.

[27]We do not benchmark against the Indian implementation of THPP, due to seemingly conflicting estimates of costs in the original paper (Fuhr et al., 2019) vis-a-vis the long-term follow-up (Bhat et al., 2022). Comparator costs are converted to 2021 purchasing power parity-adjusted dollars using the same three-step procedure described above, applied to the country and the year in which each cost was incurred; the exception is the Healthy Activity Program, for which Patel et al. (2017) report costs that are already denominated in 2015 international dollars, so that only the rebasing step is required.

at follow-up and thus no estimate of cost per case of depression averted is generated (Haushofer et al., 2020; Vanobberghen et al., 2020).[28] The inclusion criteria also vary widely across papers, including various cut-offs on the PHQ-9 and PHQ-8, as well as use of other indices to define depression (as summarized in Table A15).

The findings are summarized in Figure 3, with the studies sorted by cost per beneficiary. We observe that gPM+ as implemented in this context is among the least costly interventions per beneficiary observed in the literature, below the cost of gPM+ in Turkey ($188) and of group therapy in Uganda ($150), and roughly an order of magnitude below individual PM+ in Kenya ($3,282); the Healthy Activity Program in India is the one comparator that is less costly still, at $75 per beneficiary. Cost per month of depression averted is also relatively low, though within a small sample. HAP is the more cost-effective intervention on this measure, averting a month of depression in the short term at a cost of $39, compared to $120 for gPM+ delivered by LFs and between $145 and $162 when delivered by HEWs; both arms of gPM+, however, compare favorably to IPT-G targeting adolescents in Uganda, at $462. Overall, gPM+ performs well in terms of short-term cost-effectiveness relative to the alternatives for which comparable estimates are available, with the caveat that long-term effects are anticipated to be null (given null effects at one year); HAP, by contrast, had persistent effects even at five years, rendering the ultimate cost per month of depression averted over that longer time horizon approximately $8.

# 6 Conclusion

This paper presents novel evidence around the effectiveness of using two different strategies to scale a promising model of frontline psychological support (gPM+) in rural Ethiopia: deploying NGO-engaged facilitators and training government HEWs. In the short-run, gPM+ is effective at improving mental health (but not child development) when delivered by both sets of facilitators, though the effects are larger when gPM+ is delivered by NGO facilitators and when delivered to women. In the longer-run, these improvements attenuate to zero as individuals in the control arm recover from adverse mental health challenges reported at baseline, suggesting that gPM+ accelerates recovery rather than permanently shifting trajectories. Consistent with the short term impacts on mental health, we find short term improvements on self-efficacy, impairment, tension reduction and coping strategies, and social support, but these effects also attenuate in the longer-term. There is little evidence of any short term or longer term effects of gPM+ on other domains (including economic, cognitive, behavioral, and childcare outcomes).

[28] For the implementation of THPP in Pakistan, Gibbs et al. (2025) provides a cost estimate, and data on short-term effectiveness is provided in Vanobberghen et al. (2020) for a larger sample; the estimates of effectiveness for a Pakistan-specific sample in the latter paper as null.

Facilitator quality likely explains some of the differences observed across arms. While fidelity of implementation is high across HEWs and LFs, HEWs conduct slightly fewer sessions (although have higher attendance when sessions are conducted), and participants report higher satisfaction with LFs compared to HEWs. HEWs also show virtually zero continued engagement in delivering services following a first round of gPM+, while LFs continue to provide therapy to new clients over a full year (as they continued to be engaged with the providing organization). This constitutes suggestive evidence that while training government staff may be an effective strategy to deliver some frontline health services in the short-run, this engagement is unlikely to persist over time without continued reinforcement and support.

In terms of scaling up mental health interventions, two caveats are worth noting. First, by targeting PSNP clients, high take-up rates of gPM+ could be related to clients initial expectation that participation may lead to other program benefits. It may have also created important synergies across programs and outcomes (for example gPM+ may be more effective if households have less poverty related stress), which we are unable to measure in this study but have explored in other studies (Hidrobo et al., 2026). The second caveat is that even in the study's HEW delivery model, an NGO led the training and provided the materials. This model of NGOs providing support to government is common in Ethiopia and many other countries, and maybe a promising solution for scaling up mental health interventions in LMICs.

# References


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

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

**Acarturk, C, Ersin Uygun, Z Ilkkursun, Taylan Yurtbakan, Gülşah Kurt, Jais Adam-Troian, I Senay, R Bryant, P Cuijpers, N Kiselev et al.**, "Group problem management plus (PM+) to decrease psychological distress among Syrian refugees in Turkey: a pilot randomised controlled trial," *BMC psychiatry*, 2022, *22* (1), 8.

**Akhtar, A., P. Koyiet, A. Rahman, A. Schafer, S. U. Hamdani, P. Cuijpers, M. Sijbrandij, and R. A. Bryant**, "Residual posttraumatic stress disorder symptoms after provision of brief behavioral intervention in low- and middle-income countries," *Depression and Anxiety*, 2022, *39* (1), 71–82.

**Anderson, Michael L**, "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American statistical Association*, 2008, *103* (484), 1481–1495.

**Angelucci, M. and D. Bennett**, "The economic impact of depression treatment in India," *American Economic Review*, 2024, *114* (1), 169–198.

**Ashraf, N., O. Bandiera, E. Davenport, and S. S. Lee**, "Losing prosociality in the quest for talent?," *American Economic Review*, 2020, *110* (5), 1355–1394.

**Assefa, Y., Y. A. Gelaw, P. S. Hill, B. W. Taye, and W. Van Damme**, "Community health extension program of Ethiopia, 2003–2018," *Globalization and Health*, 2019, *15*, 1–11.

**Baird, Sarah, Berk Özler, Chiara Dell'Aira, Luca Parisotto, and Danish Us-Salam**, "Therapy, mental health, and human capital accumulation among adolescent girls in Uganda," *Journal of Development Economics*, 2025, *176*, 103473.

**Baranov, V., S. Bhalotra, P. Biroli, and J. Maselko**, "Maternal Depression, Women's Empowerment, and Parental Investment," *American Economic Review*, 2020, *110* (3), 824–859.

**Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen**, "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 2014, *81* (2), 608–650.

**Bhat, Bhargav, Jonathan De Quidt, Johannes Haushofer, Vikram H Patel, Gautam Rao, Frank Schilbach, and Pierre-Luc P Vautrey**, "The long-run effects of psychotherapy on depression, beliefs, and economic outcomes," Technical Report, National Bureau of Economic Research 2022.

**Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, and J. Sandefur**, "Experimental evidence on scaling up education reforms in Kenya," *Journal of Public Economics*, 2018, *168*, 1–20.

**Bromet, Evelyn, Laura Helena Andrade, Irving Hwang, Nancy A. Sampson, Jordi Alonso, Giovanni de Girolamo et al.**, "Cross-National Epidemiology of DSM-IV Major Depressive Episode," *BMC Medicine*, 2011, *9*, 90.

**Bryant, Richard A, Ahmad Bawaneh, Manar Awwad, Hadeel Al-Hayek, Luana Giardinelli, Claire Whitney, Mark JD Jordans, Pim Cuijpers, Marit Sijbrandij, Peter Ventevogel et al.**, "Effectiveness of a brief group behavioral intervention for common mental disorders in Syrian refugees in Jordan: A randomized controlled trial," *PLoS Medicine*, 2022, *19* (3), e1003949.

**_ , _ , _ , _ , _ , _ , _ , _ , _ , _ , _ , and _** , "Twelve-month follow-up of a randomised clinical trial of a brief group psychological intervention for common mental disorders in Syrian refugees in Jordan," *Epidemiology and Psychiatric Sciences*, 2022, *31*, e81.

**_ , Alison Schafer, Katie S Dawson, Dorothy Anjuri, Caroline Mulili, Lincoln Ndogoni, Phiona Koyiet, Marit Sijbrandij, Jeannette Ulate, Melissa Harper Shehadeh et al.**, "Effectiveness of a brief behavioural intervention on psychological distress among women with a history of gender-based violence in urban Kenya: a randomised clinical trial," *PLoS medicine*, 2017, *14* (8), e1002371.

**Bunn, Mary, Nicole Gonzalez, Idan Falek, Stevan Weine, and Mary Acri**, "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*, 2021, *19* (2), 155–179.

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

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

**Dawson, Katie S, Richard A Bryant, Melissa Harper, Alvin Kuowei Tay, Atif Rahman, Alison Schafer, and Mark Van Ommeren**, "Problem Management Plus (PM+): a WHO transdiagnostic psychological intervention for common mental health problems," *World Psychiatry*, 2015, *14* (3), 354.

**Deserranno, E.**, "Financial incentives as signals," *American Economic Journal: Applied Economics*, 2019, *11* (1), 277–317.

**Desta, Firew Ayalew, Girma Temam Shifa, Damtew WoldeMariam Dagoye, Catherine Carr, Jos Van Roosmalen, Jelle Stekelenburg, Assefa Bulcha Nedi, Adrienne Kols, and Young Mi Kim**, "Identifying gaps in the practices of rural health extension workers in Ethiopia: a task analysis study," *BMC health services research*, 2017, *17* (1), 839.

**Federal Ministry of Health**, "National Mental Health Strategic Plan 2020–2025 (2013–2017 EFY)," Technical Report, Diseases Prevention and Control Directorate, Addis Ababa, Ethiopia 2020.

**Fisseha, Girmatsion, Tesfay Gebregzabher Gebrehiwot, Mengistu Welday Gebremichael, Shishay Wahdey, Gebrekiros Gebremichael Meles, Kebede Embaye Gezae, Awol Yemane Legesse, Akeza Awealom Asgedom, Mache Tsadik, Abraha Woldemichael et al.**, "War-related sexual and gender-based violence in Tigray, Northern Ethiopia: a community-based study," *BMJ global health*, 2023, *8* (7), e010270.

**Fitch-Fleischmann, Benjamin and Evan Plous Kresch**, "Story of the hurricane: Government, NGOs, and the difference in disaster relief targeting," *Journal of Development Economics*, 2021, *152*, 102702.

**Fuhr, Daniela C, Benedict Weobong, Anisha Lazarus, Fiona Vanobberghen, Helen A Weiss, Daisy R Singla, Hanani Tabana, Ejma Afonso, Aveena De Sa, Ethel D'Souza et al.**, "Delivering the Thinking Healthy Programme for perinatal depression through peers: an individually randomised controlled trial in India," *The Lancet Psychiatry*, 2019, *6* (2), 115–127.

**Gesesew, Hailay, Kiros Berhane, Elias S Siraj, Dawd Siraj, Mulugeta Gebregziabher, Yemane Gebremariam Gebre, Samuel Aregay Gebreslassie, Fasika Amdeslassie, Azeb Gebresilassie Tesema, Amir Siraj et al.**, "The impact of war on the health system of the Tigray region in Ethiopia: an assessment," *BMJ Global Health*, 2021, *6* (11).

**Gibbs, Naomi Kate, Tao Chen, Abid Malik, Huma Nazir, Anum Nisar, Ahmed Waqas, Najia Atif, Duolao Wang, Atif Rahman, Siham Sikander**

**et al.**, "Cost-effectiveness of a technology-assisted peer-delivered perinatal mental health intervention in Pakistan: an economic evaluation using trial evidence," *BMJ global health*, 2025, *10* (11).

**Greene, M Claire, Diany Castellar, Manaswi Sangraula, Natalia Camargo, Jennifer Diaz, Valeria Meriño, Lucy Miller-Suchet, Ana Maria Chamorro Coneo, Marcela Venegas, Maria Cristobal et al.**, "Comparing implementation strategies for training and supervising nonspecialists in Group Problem Management Plus: A hybrid effectiveness-implementation trial in Colombia," *Cambridge Prisms: Global Mental Health*, 2024, *11*, e90.

**Guerra, Mariella, A. Matthew Prina, Cleusa P. Ferri, Daisy Acosta, Sara Gallardo, Yueqin Huang et al.**, "A Comparative Cross-Cultural Study of the Prevalence of Late Life Depression in Low and Middle Income Countries," *Journal of Affective Disorders*, 2016, *190*, 1019–1026.

**Haushofer, Johannes, Robert Mudida, and Jeremy P Shapiro**, "The comparative impact of cash transfers and a psychotherapy program on psychological and economic well-being," Technical Report, National Bureau of Economic Research 2020.

**Henderson, J Vernon and Yong Suk Lee**, "Organization of disaster aid delivery: spending your donations," *Economic Development and Cultural Change*, 2015, *63* (4), 617–664.

**Hidrobo, Melissa, Harold Alderman, Negussie Deyessa, Daniel O Gilligan, Parthu Kalva, Jessica Leight, Michael Mulford, and Heleene Tambet**, "The effects of cash and group therapy in the context of conflict: Evidence from a randomized evaluation in Ethiopia," *Journal of Development Economics*, 2026, p. 103724.

**Jordans, Mark JD, Brandon A Kohrt, Manaswi Sangraula, Elizabeth L Turner, Xueqi Wang, Pragya Shrestha, Renasha Ghimire, Edith van't Hof, Richard A Bryant, Katie S Dawson et al.**, "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*, 2021, *18* (6), e1003621.

**Karyotaki, Eirini, Ricardo Araya, Ronald C Kessler, Ahmed Waqas, Arvin Bhana, Atif Rahman, Camila T Matsuzaka, Clara Miguel, Crick Lund, Emily C Garman et al.**, "Association of task-shared psychological interventions with depression outcomes in low-and middle-income countries: a systematic review and individual patient data meta-analysis," *JAMA psychiatry*, 2022, *79* (5), 430–443.

**Kerwin, Jason, Nada Rostom, and Olivier Sterck**, "Striking the Right Balance: Why Standard Balance Tests Over-Reject the Null, and How to Fix It," IZA Discussion Paper 17217, Institute of Labor Economics (IZA) 2024.

**Kessler, Ronald C., Katherine A. McGonagle, Marvin Swartz, Dan G. Blazer, and Christopher B. Nelson**, "Sex and Depression in the National Comorbidity Survey. I: Lifetime Prevalence, Chronicity and Recurrence," *Journal of Affective Disorders*, 1993, *29* (2–3), 85–96.

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

**Mariam, Damon Haile, Abiy Estifanos, Dawit Wondimagegn, Bryony Simmons, and Lesong Conteh**, "Optimizing the Ethiopian Health Extension Programme: Strategies to address workforce challenges," Technical Report, WHO Regional Office for Africa, Brazzaville 2023.

**Maselko, Joanna, Siham Sikander, Elizabeth L Turner, Lisa M Bates, Ikhlaq Ahmad, Najia Atif, Victoria Baranov, Sonia Bhalotra, Amina Bibi, Tayyaba Bibi et al.**, "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*, 2020, *7* (9), 775–787.

**Mo, Di, Yu Bai, Yaojiang Shi, Cody Abbey, Linxiu Zhang, Scott Rozelle, and Prashant Loyalka**, "Institutions, implementation, and program effectiveness: Evidence from a randomized evaluation of computer-assisted learning in rural China," *Journal of Development Economics*, 2020, *146*, 102487.

**Nsibande, Duduzile, Marian Loveday, Karen Daniels, David Sanders, Tanya Doherty, and Wanga Zembe**, "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*, 2018, *18* (1), 188–197.

**Patel, Vikram, Benedict Weobong, Helen A Weiss, Arpita Anand, Bhargav Bhat, Basavraj Katti, Sona Dimidjian, Ricardo Araya, Steve D Hollon, Michael King et al.**, "The Healthy Activity Program (HAP), a lay counsellor-delivered brief psychological treatment for severe depression, in primary care in India: a randomised controlled trial," *The Lancet*, 2017, *389* (10065), 176–185.

**_ , Shekhar Saxena, Crick Lund, Graham Thornicroft, Florence Baingana, Paul Bolton, Dan Chisholm, Pamela Y Collins, Janice L Cooper, Julian**

**Eaton et al.**, "The Lancet Commission on global mental health and sustainable development," *The lancet*, 2018, *392* (10157), 1553–1598.

**Rahman, Atif, Muhammad Naseem Khan, Syed Usman Hamdani, Anna Chiumento, Parveen Akhtar, Huma Nazir, Anum Nisar, Aqsa Masood, Iftikhar Ud Din, Nasir Ali Khan et al.**, "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*, 2019, *393* (10182), 1733–1744.

**Ridley, M., G. Rao, F. Schilbach, and V. Patel**, "Poverty, depression, and anxiety," *Science*, 2020, *370* (6522), eaay0214.

**Rose, Suzanna C., Jonathan Bisson, Rachel Churchill, and Simon Wessely**, "Psychological debriefing for preventing post traumatic stress disorder (PTSD)," *Cochrane Database of Systematic Reviews*, 2002, (2), CD000560.

**Salk, Rachel H., Janet S. Hyde, and Lyn Y. Abramson**, "Gender Differences in Depression in Representative National Samples: Meta-Analyses of Diagnoses and Symptoms," *Psychological Bulletin*, 2017, *143* (8), 783–822.

**Schaefer, Sarah K, Lea M Thomas, Saskia Lindner, and Klaus Lieb**, "World Health Organization's low-intensity psychosocial interventions: a systematic review and meta-analysis of the effects of Problem Management Plus and Step-by-Step," *World Psychiatry*, 2023, *22* (3), 449–462.

**Schwartz, J Brad and Indu Bhushan**, *Cambodia: using contracting to reduce inequity in primary health care delivery*, World Bank, Washington, DC, 2004.

**Singla, Daisy R, Brandon A Kohrt, Laura K Murray, Arpita Anand, Bruce F Chorpita, and Vikram Patel**, "Psychological treatments for the world: lessons from low-and middle-income countries," *Annual review of clinical psychology*, 2017, *13* (1), 149–181.

**Tesema, Azeb Gebresilassie and Yohannes Kinfu**, "Reorienting and rebuilding the health system in war-torn Tigray, Ethiopia," *BMJ global health*, 2021, *6* (8).

**Tesfaye, Yonas, Zewdie Birhanu, Liyew Agenagnew, Susan Anand, Kiddus Yitbarek, Gutema Ahmed, Masrie Getnet, and Gudina Terefe Tucho**, "Knowledge and attitude of health extension workers regarding mental health problems in Jimma Zone, Ethiopia: a cross-sectional study," *BMJ open*, 2022, *12* (2), e048381.

**Tilahun, Hibret, Binyam Fekadu, Habtamu Abdisa, Maureen Canavan, Erika Linnander, Elizabeth H Bradley, and Peter Berman**, "Ethiopia's health

extension workers use of work time on duty: time and motion study," *Health policy and planning*, 2017, *32* (3), 320–328.

**Tiruneh, Misganaw Guadie, Eneyew Talie Fenta, Destaw Endeshaw, Amare Mebrat Delie, Ousman Adal, Abiyu Abadi Tareke, Eyob Ketema Bogale, and Tadele Fentabel Anagaw**, "Health extension service utilization in Ethiopia: systematic review and meta-analysis," *BMC health services research*, 2024, *24* (1), 537.

**Vanobberghen, Fiona, Helen A Weiss, Daniela C Fuhr, Siham Sikander, Ejma Afonso, Ikhlaq Ahmad, Najia Atif, Amina Bibi, Tayyaba Bibi, Samina Bilal et al.**, "Effectiveness of the Thinking Healthy Programme for perinatal depression delivered through peers: Pooled analysis of two randomized controlled trials in India and Pakistan," *Journal of Affective Disorders*, 2020, *265*, 660–668.

**Vivalt, E.**, "How Much Can We Generalize From Impact Evaluations?," 2020, *18* (6), 3045–3089.

**Weldegiargis, Aregawi Weldegebreal, Haftom Temesgen Abebe, Hiluf Ebuy Abraha, Meron Micheale Abrha, Tsegay Berihu Tesfay, Rieye Esayas Belay, Alemnesh Abraha Araya, Mengish Bahresilassie Gebregziabher, Hagos Godefay, and Afework Mulugeta**, "Armed conflict and household food insecurity: evidence from war-torn Tigray, Ethiopia," *Conflict and Health*, 2023, *17* (1), 22.

**Weobong, Benedict, Helen A Weiss, David McDaid, Daisy R Singla, Steven D Hollon, Abhijit Nadkarni, A-La Park, Bhargav Bhat, Basavraj Katti, Arpita Anand et al.**, "Sustained effectiveness and cost-effectiveness of the Healthy Activity Programme, a brief psychological treatment for depression delivered by lay counsellors in primary care: 12-month follow-up of a randomised controlled trial," *PLoS medicine*, 2017, *14* (9), e1002385.

**World Health Organization**, *Depression and other common mental disorders: Global Health Estimates*, World Health Organization, 2017.

_ , *Problem Management Plus (PM+): Individual psychological help for adults impaired by distress in communities exposed to adversity*, World Health Organization, 2018.

_ , *Group Problem Management Plus (Group PM+): group psychological help for adults impaired by distress in communities exposed to adversity (generic field-trial version 1.0)*, World Health Organization, 2020.

**Yamey, Gavin**, "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*, 2012, *8* (1), 11.

Figure 1: Trial design

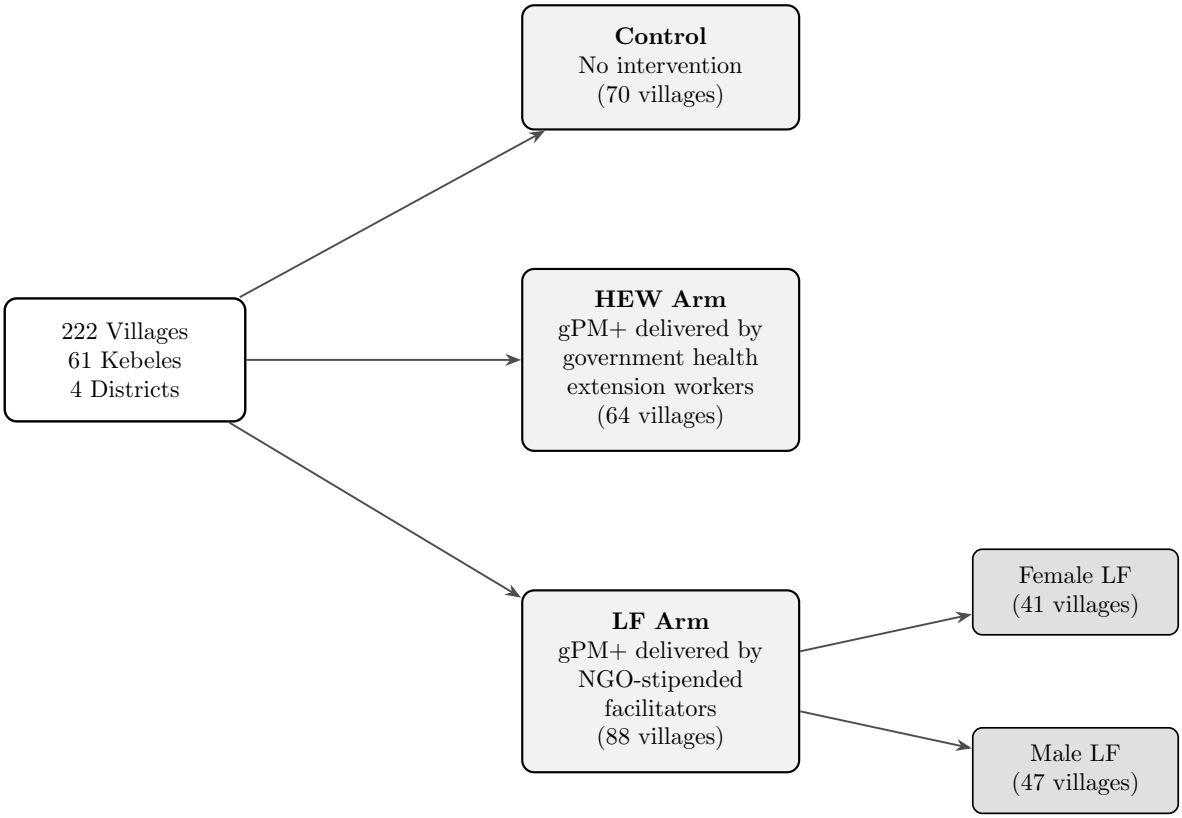
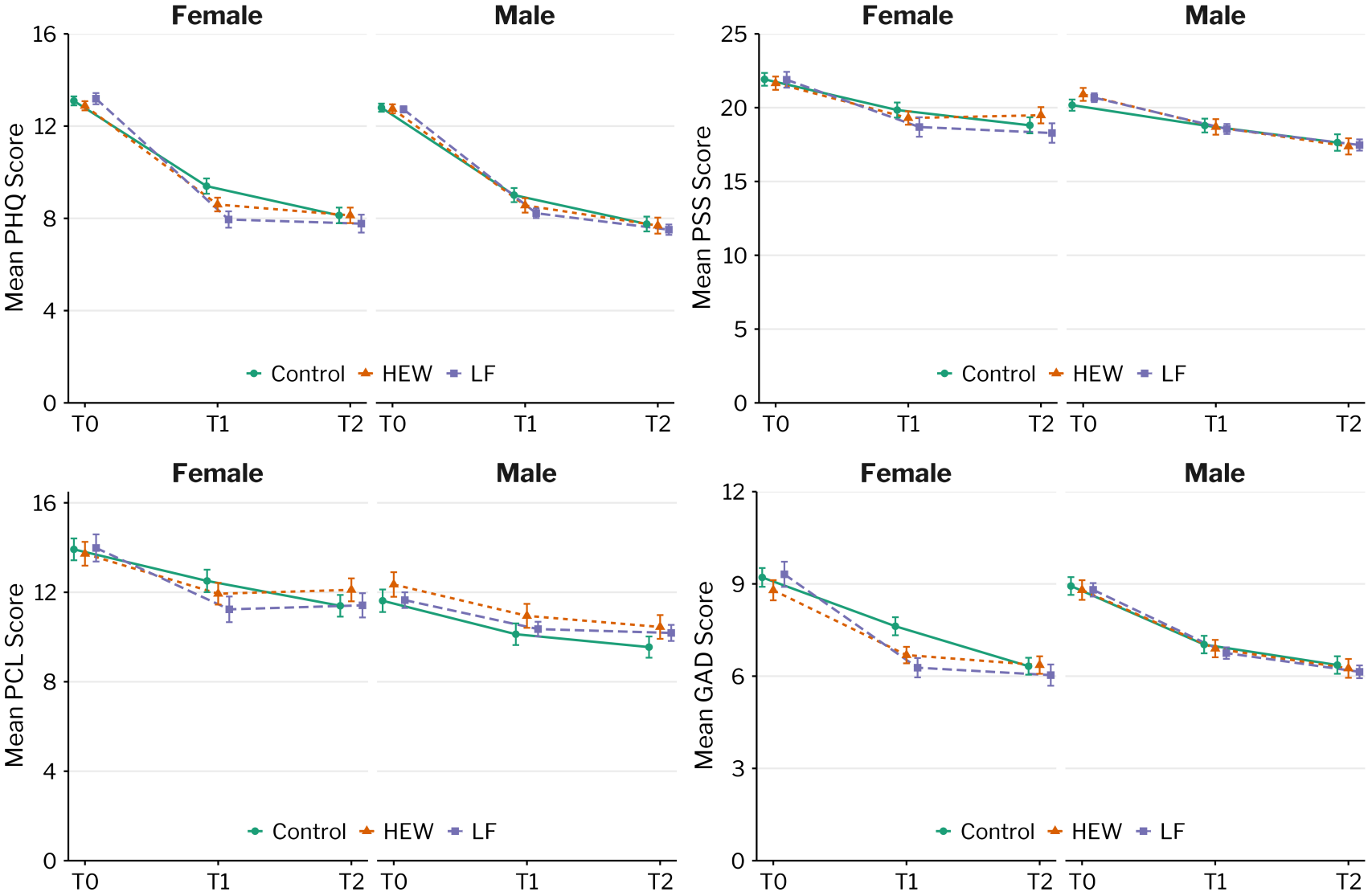
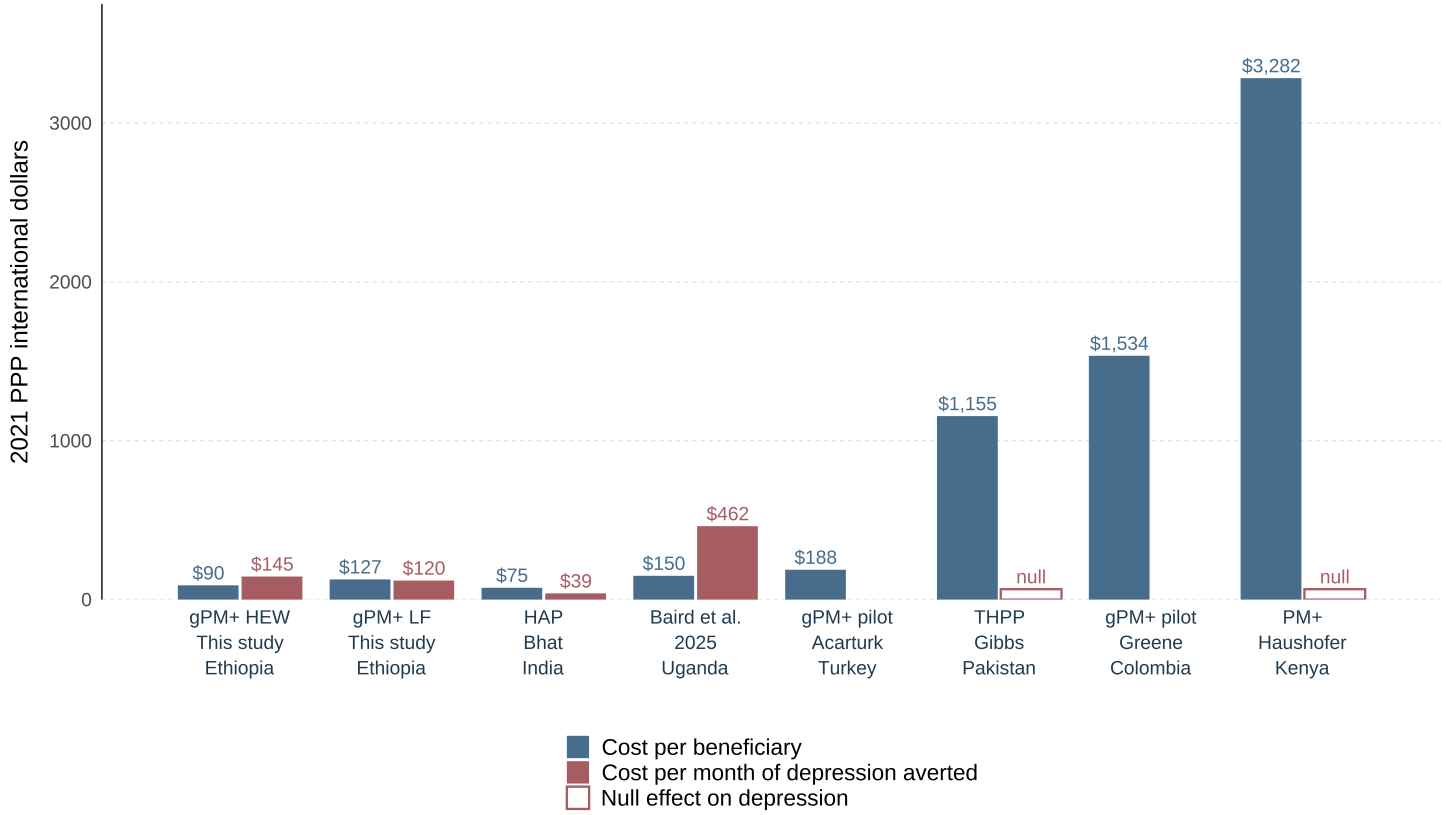


Figure 2: Longitudinal trends in mental health outcomes over time




Notes: This figure captures trends by arm in the key mental health outcomes over time. T0 denotes baseline, T1 denotes the one month follow-up, and T2 denotes the 12 month follow-up. The confidence intervals included are descriptive and do not account for clustering.

Figure 3: Cost effectiveness




Pilot studies, effectiveness on depression not measured: gPM+ pilot (Acarturk, Turkey); gPM+ pilot (Greene, Colombia).

Table 1: Primary Indices

| | (1) | (2) | (3) |
|---|---|---|---|
| | Mental health index | | Child development score |
| | 1-month follow-up | 12-month follow-up | 12-month follow-up |
| HEW x Female | -0.166** | 0.092 | 0.025 |
| | (0.065) | (0.071) | (0.092) |
| HEW x Male | -0.049 | -0.025 | 0.056 |
| | (0.056) | (0.071) | (0.072) |
| Female LF x Female | -0.347*** | -0.094 | 0.155 |
| | (0.079) | (0.096) | (0.096) |
| Female LF x Male | -0.071 | -0.072 | 0.040 |
| | (0.077) | (0.085) | (0.081) |
| Male LF x Male | -0.096* | -0.009 | 0.065 |
| | (0.055) | (0.066) | (0.067) |
| Average HEW effect | -0.108*** | 0.034 | 0.040 |
| | (0.041) | (0.046) | (0.054) |
| Weighted LF effect | -0.218*** | -0.062 | 0.106** |
| | (0.045) | (0.053) | (0.053) |
| Average Female LF effect | -0.209*** | -0.083 | 0.098 |
| | (0.049) | (0.058) | (0.061) |
| p-value (Avg HEW = Avg. Weighted LF) | 0.009 | 0.073 | 0.251 |
| p-value (Avg HEW = Avg Female LF) | 0.029 | 0.048 | 0.373 |
| Control mean | -0.57 | -0.90 | 0.10 |
| *N* | 3,429 | 3,391 | 2,702 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the follow-up outcome on HEW, female LF, and male LF treatment indicators interacted with respondent gender where applicable, respondent gender, the baseline value of the outcome, and baseline-sub-district fixed effects. Missing baseline values are imputed to zero with a missingness indicator. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Reported average effects use realized composition shares. The weighted LF effect weights the female-LF female-respondent, female-LF male-respondent, and male-LF male-respondent coefficients by 0.50, 0.167, and 0.333 respectively, reflecting that female respondents are served only by female LFs and male respondents roughly one-third by female and two-thirds by male LFs. Outcome definitions are provided in Table A1 and Table A2.

Table 2: Mental Health Raw Scores (1 Month Follow-up)

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | PHQ-9 score | Mod-severe depression | PSS-10 score | PCL-5 score | GAD-7 score |
| HEW x Female | -0.663** | -0.067* | -0.506 | -0.538 | -0.830*** |
| | (0.268) | (0.036) | (0.435) | (0.447) | (0.239) |
| | [0.032] | [0.091] | [0.242] | [0.242] | [0.003] |
| HEW x Male | -0.378* | -0.107*** | -0.360 | 0.581 | -0.083 |
| | (0.227) | (0.034) | (0.418) | (0.384) | (0.195) |
| | [0.129] | [0.006] | [0.261] | [0.143] | [0.385] |
| Female LF x Female | -1.385*** | -0.147*** | -0.977 | -1.210*** | -1.307*** |
| | (0.290) | (0.039) | (0.609) | (0.463) | (0.282) |
| | [0.001] | [0.002] | [0.133] | [0.024] | [0.001] |
| Female LF x Male | -0.609** | -0.088** | -0.492 | 0.662 | -0.052 |
| | (0.308) | (0.040) | (0.501) | (0.423) | (0.255) |
| | [0.074] | [0.048] | [0.257] | [0.136] | [0.506] |
| Male LF x Male | -0.749*** | -0.141*** | -0.218 | 0.324 | -0.199 |
| | (0.226) | (0.031) | (0.365) | (0.359) | (0.200) |
| | [0.004] | [0.001] | [0.310] | [0.261] | [0.257] |
| Average HEW effect | -0.523*** | -0.086*** | -0.434 | 0.014 | -0.462*** |
| | (0.164) | (0.022) | (0.299) | (0.259) | (0.156) |
| Weighted LF effect | -1.043*** | -0.135*** | -0.643* | -0.387 | -0.729*** |
| | (0.172) | (0.022) | (0.334) | (0.250) | (0.163) |
| Average Female LF effect | -0.997*** | -0.117*** | -0.735** | -0.274 | -0.679*** |
| | (0.199) | (0.024) | (0.352) | (0.268) | (0.170) |
| p-value (Avg HEW = Avg. Weighted LF) | 0.002 | 0.023 | 0.515 | 0.138 | 0.090 |
| p-value (Avg HEW = Avg Female LF) | 0.015 | 0.181 | 0.377 | 0.305 | 0.193 |
| Control mean | 9.40 | 0.44 | 19.84 | 12.51 | 7.62 |
| *N* | 3,429 | 3,429 | 3,429 | 3,429 | 3,429 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the follow-up outcome on HEW, female LF, and male LF treatment indicators interacted with respondent gender where applicable, respondent gender, the baseline value of the outcome, and baseline-sub-district fixed effects. Missing baseline values are imputed to zero with a missingness indicator. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Reported average effects use realized composition shares. The weighted LF effect weights the female-LF female-respondent, female-LF male-respondent, and male-LF male-respondent coefficients by 0.50, 0.167, and 0.333 respectively, reflecting that female respondents are served only by female LFs and male respondents roughly one-third by female and two-thirds by male LFs. Sharpened two-stage q-values as described in (Anderson, 2008) are reported in square brackets. Outcome definitions are provided in Table A1 and Table A2.

Table 3: Mental Health Raw Scores (12 Month Follow-up)

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | PHQ-9 score | Mod-severe depression | PSS-10 score | PCL-5 score | GAD-7 score |
| HEW x Female | 0.096 | -0.018 | 0.666 | 0.699 | 0.089 |
| | (0.271) | (0.030) | (0.465) | (0.448) | (0.213) |
| | [1.000] | [1.000] | [1.000] | [1.000] | [1.000] |
| HEW x Male | -0.047 | -0.019 | -0.573 | 0.622 | -0.117 |
| | (0.260) | (0.030) | (0.470) | (0.423) | (0.237) |
| | [1.000] | [1.000] | [1.000] | [1.000] | [1.000] |
| Female LF x Female | -0.391 | -0.056 | -0.637 | 0.095 | -0.275 |
| | (0.359) | (0.039) | (0.643) | (0.431) | (0.294) |
| | [1.000] | [1.000] | [1.000] | [1.000] | [1.000] |
| Female LF x Male | -0.310 | -0.020 | -0.842 | 0.408 | -0.200 |
| | (0.269) | (0.033) | (0.585) | (0.501) | (0.282) |
| | [1.000] | [1.000] | [1.000] | [1.000] | [1.000] |
| Male LF x Male | -0.136 | -0.042 | -0.160 | 0.798** | -0.225 |
| | (0.243) | (0.029) | (0.449) | (0.371) | (0.228) |
| | [1.000] | [1.000] | [1.000] | [0.893] | [1.000] |
| Average HEW effect | 0.026 | -0.019 | 0.055 | 0.661*** | -0.012 |
| | (0.167) | (0.020) | (0.311) | (0.254) | (0.162) |
| Weighted LF effect | -0.293 | -0.045** | -0.512 | 0.382 | -0.246 |
| | (0.190) | (0.022) | (0.360) | (0.249) | (0.180) |
| Average Female LF effect | -0.351* | -0.038 | -0.739* | 0.252 | -0.237 |
| | (0.196) | (0.024) | (0.395) | (0.310) | (0.202) |
| p-value (Avg HEW = Avg. Weighted LF) | 0.106 | 0.241 | 0.119 | 0.312 | 0.200 |
| p-value (Avg HEW = Avg Female LF) | 0.066 | 0.435 | 0.048 | 0.216 | 0.264 |
| Control mean | 8.13 | 0.32 | 18.80 | 11.40 | 6.32 |
| $N$ | 3,391 | 3,391 | 3,391 | 3,391 | 3,391 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the follow-up outcome on HEW, female LF, and male LF treatment indicators interacted with respondent gender where applicable, respondent gender, the baseline value of the outcome, and baseline-sub-district fixed effects. Missing baseline values are imputed to zero with a missingness indicator. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Reported average effects use realized composition shares. The weighted LF effect weights the female-LF female-respondent, female-LF male-respondent, and male-LF male-respondent coefficients by 0.50, 0.167, and 0.333 respectively, reflecting that female respondents are served only by female LFs and male respondents roughly one-third by female and two-thirds by male LFs. Sharpened two-stage q-values as described in (Anderson, 2008) are reported in square brackets. Outcome definitions are provided in Table A1 and Table A2.

Table 4: Secondary Outcomes Indices (1 Month Follow-up)

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Psychosocial | Behavioral | Cognitive | Economic | Childcare |
| HEW x Female | 0.176*** | 0.051 | 0.081 | -0.026 | 0.084* |
| | (0.040) | (0.036) | (0.054) | (0.033) | (0.045) |
| HEW x Male | 0.089** | -0.058 | -0.003 | 0.008 | 0.042 |
| | (0.039) | (0.064) | (0.064) | (0.026) | (0.045) |
| Female LF x Female | 0.191*** | 0.038 | -0.007 | -0.034 | -0.017 |
| | (0.055) | (0.038) | (0.067) | (0.038) | (0.048) |
| Female LF x Male | 0.114** | -0.059 | -0.040 | 0.034 | -0.011 |
| | (0.051) | (0.059) | (0.073) | (0.037) | (0.045) |
| Male LF x Male | 0.120*** | -0.005 | -0.010 | -0.017 | -0.007 |
| | (0.038) | (0.057) | (0.059) | (0.027) | (0.037) |
| Average HEW effect | 0.133*** | -0.003 | 0.039 | -0.009 | 0.064** |
| | (0.031) | (0.037) | (0.041) | (0.021) | (0.032) |
| Weighted LF effect | 0.154*** | 0.007 | -0.014 | -0.017 | -0.013 |
| | (0.034) | (0.033) | (0.043) | (0.023) | (0.030) |
| Average Female LF effect | 0.153*** | -0.011 | -0.024 | 0.000 | -0.014 |
| | (0.040) | (0.037) | (0.048) | (0.027) | (0.033) |
| Control mean | -0.02 | 0.25 | -0.12 | 0.01 | -0.17 |
| $N$ | 3,429 | 3,429 | 3,429 | 3,429 | 2,819 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the follow-up outcome on HEW, female LF, and male LF treatment indicators interacted with respondent gender where applicable, respondent gender, the baseline value of the outcome (except for column 3), and baseline-sub-district fixed effects. Missing baseline values are imputed to zero with a missingness indicator. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Reported average effects use realized composition shares. The weighted LF effect weights the female-LF female-respondent, female-LF male-respondent, and male-LF male-respondent coefficients by 0.50, 0.167, and 0.333 respectively, reflecting that female respondents are served only by female LFs and male respondents roughly one-third by female and two-thirds by male LFs. Outcome definitions are provided in Table A1 and Table A2.

Table 5: Secondary Outcomes Indices (12 Month Follow-up)

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Psychosocial | Behavioral | Cognitive | Economic | Childcare |
| HEW x Female | 0.045 | 0.017 | -0.005 | -0.041 | 0.034 |
| | (0.051) | (0.035) | (0.055) | (0.040) | (0.044) |
| HEW x Male | 0.020 | 0.005 | -0.084* | 0.008 | -0.057 |
| | (0.041) | (0.059) | (0.050) | (0.043) | (0.046) |
| Female LF x Female | 0.079 | 0.014 | -0.024 | -0.014 | -0.086* |
| | (0.059) | (0.050) | (0.069) | (0.044) | (0.049) |
| Female LF x Male | 0.043 | 0.026 | -0.039 | 0.057 | -0.029 |
| | (0.051) | (0.058) | (0.065) | (0.051) | (0.050) |
| Male LF x Male | -0.020 | 0.087* | -0.015 | 0.001 | -0.099** |
| | (0.042) | (0.047) | (0.053) | (0.041) | (0.042) |
| Average HEW effect | 0.032 | 0.011 | -0.044 | -0.017 | -0.011 |
| | (0.033) | (0.032) | (0.035) | (0.030) | (0.030) |
| Weighted LF effect | 0.040 | 0.040 | -0.023 | 0.003 | -0.081** |
| | (0.035) | (0.033) | (0.040) | (0.027) | (0.032) |
| Average Female LF effect | 0.061 | 0.020 | -0.031 | 0.021 | -0.057 |
| | (0.038) | (0.039) | (0.045) | (0.032) | (0.036) |
| Control mean | 0.07 | 0.25 | -0.01 | 0.35 | -0.06 |
| $N$ | 3,391 | 3,391 | 3,391 | 3,391 | 2,776 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the follow-up outcome on HEW, female LF, and male LF treatment indicators interacted with respondent gender where applicable, respondent gender, the baseline value of the outcome (except for column 3), and baseline-sub-district fixed effects. Missing baseline values are imputed to zero with a missingness indicator. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Reported average effects use realized composition shares. The weighted LF effect weights the female-LF female-respondent, female-LF male-respondent, and male-LF male-respondent coefficients by 0.50, 0.167, and 0.333 respectively, reflecting that female respondents are served only by female LFs and male respondents roughly one-third by female and two-thirds by male LFs. Outcome definitions are provided in Table A1 and Table A2.

Table 6: Facilitator quality of implementation

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| **Panel A: Administrative data** | | | | |
| | Number of sessions conducted | Mean session attendance (conditional) | Mean session attendance (unconditional) | |
| Male LF | 0.194 | -7.265*** | -5.652 | |
| | (0.396) | (2.771) | (4.293) | |
| | [0.399] | [0.062] | [0.235] | |
| Female LF | 0.619** | -4.471 | 0.607 | |
| | (0.262) | (2.708) | (3.546) | |
| | [0.062] | [0.156] | [0.405] | |
| Average LF effect | 0.392 | -5.963** | -2.736 | |
| | (0.306) | (2.339) | (3.505) | |
| p-value (Male LF = Female LF) | 0.155 | 0.332 | 0.093 | |
| Control mean | 9.38 | 88.71 | 83.63 | |
| $N$ | 151 | 147 | 151 | |
| **Panel B: Facilitator quality** | | | | |
| | Highly satisfied with provider | Prior interaction with facilitator | Comfortable speaking with facilitator | Comfortable because they knew facilitator |
| HEW x Female | -0.013 | -0.101** | -0.019 | -0.091** |
| | (0.042) | (0.041) | (0.020) | (0.037) |
| | [0.441] | [0.042] | [0.235] | [0.042] |
| HEW x Male | -0.092** | -0.117*** | -0.033* | -0.157*** |
| | (0.042) | (0.044) | (0.017) | (0.036) |
| | [0.061] | [0.039] | [0.090] | [0.001] |
| Female LF x Female | 0.068* | -0.168*** | 0.009 | -0.081* |
| | (0.041) | (0.048) | (0.017) | (0.043) |
| | [0.123] | [0.005] | [0.355] | [0.090] |
| Female LF x Male | 0.012 | 0.030 | -0.019 | 0.068 |
| | (0.054) | (0.048) | (0.017) | (0.046) |
| | [0.452] | [0.327] | [0.212] | [0.146] |
| Male LF mean | 0.538 | 0.687 | 0.941 | 0.354 |
| $N$ | 2,026 | 2,002 | 2,002 | 2,000 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Panel A uses Dimagi admin data and reports OLS regressions of each outcome on male and female LF indicators, with HEWs as the reference group and heteroskedasticity-robust standard errors in parentheses. The control mean in Panel A is the mean among HEWs. Panel B uses household follow-up data and reports OLS regressions of each outcome on HEW, female LF, and male LF treatment indicators interacted with respondent gender where applicable, respondent gender, and baseline-sub-district fixed effects, with standard errors clustered at the baseline-village level in parentheses. Panel B column 1 measures satisfaction at the 1-month follow-up; columns 2–4 use the 12-month follow-up. Panel B includes only treated households with observed outcomes; the Male LF group is the omitted category with Male LF mean being the mean among male respondents assigned to male LFs. No baseline outcome controls are included in either panel. The average LF effect in Panel A weights the male- and female-LF coefficients by their shares among LFs in the relevant source dataset. Sharpened two-stage q-values as described in (Anderson, 2008) are reported in square brackets. Outcome definitions are provided in Table A13.

# Appendix

Figure A1: Tigray and study districts



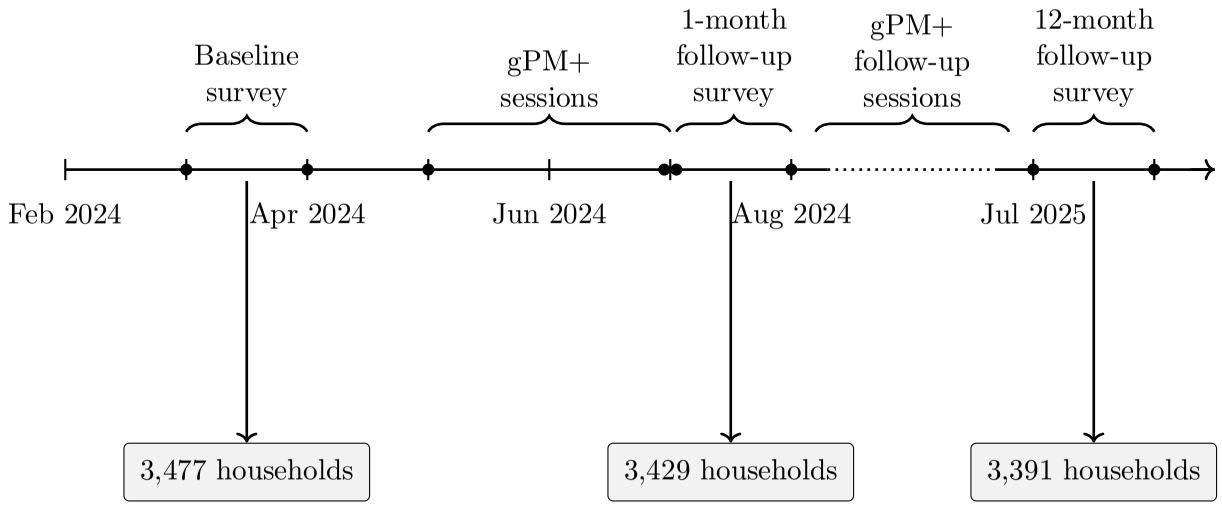



Figure A2: Project timeline

Figure A3: Flow chart

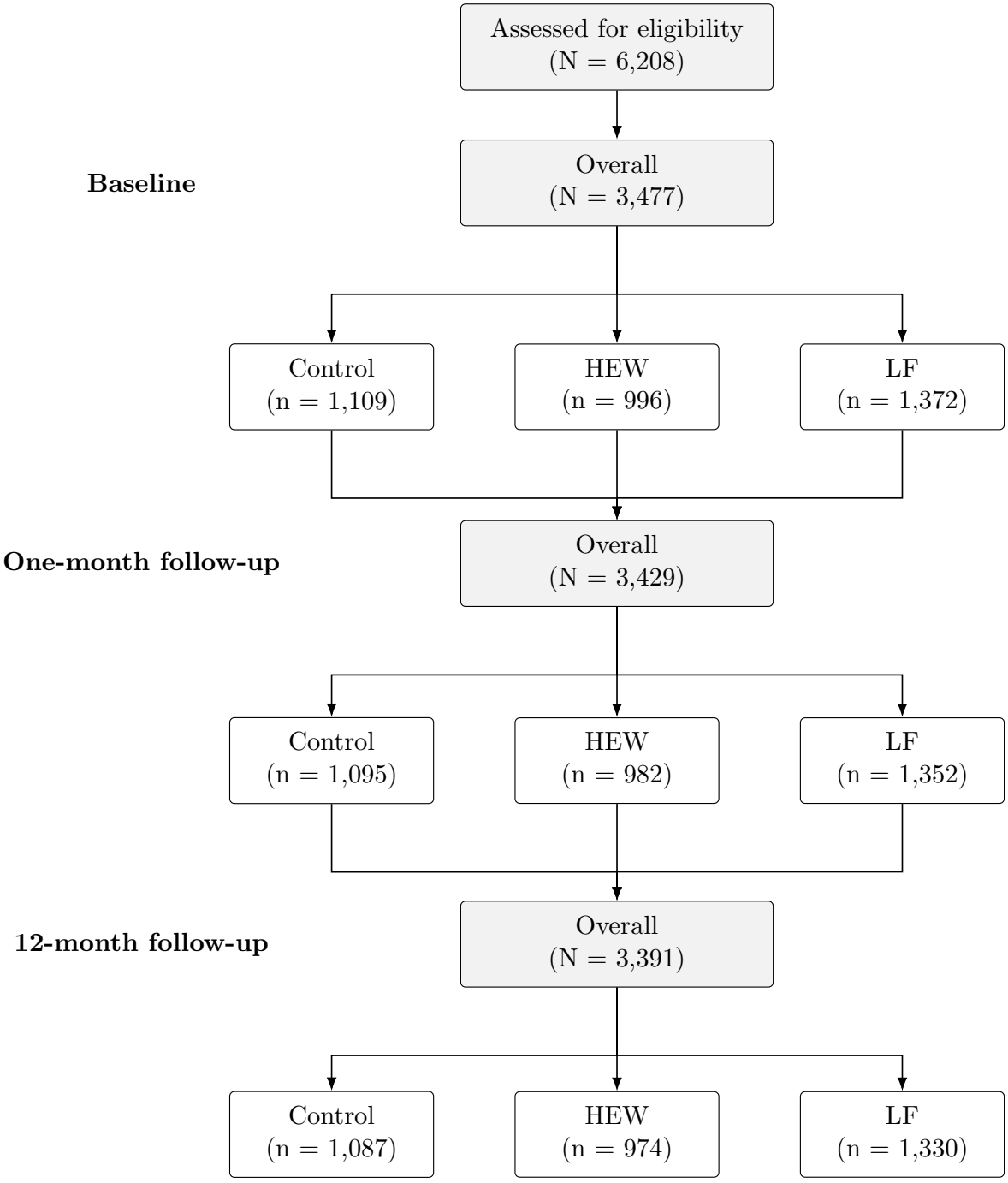
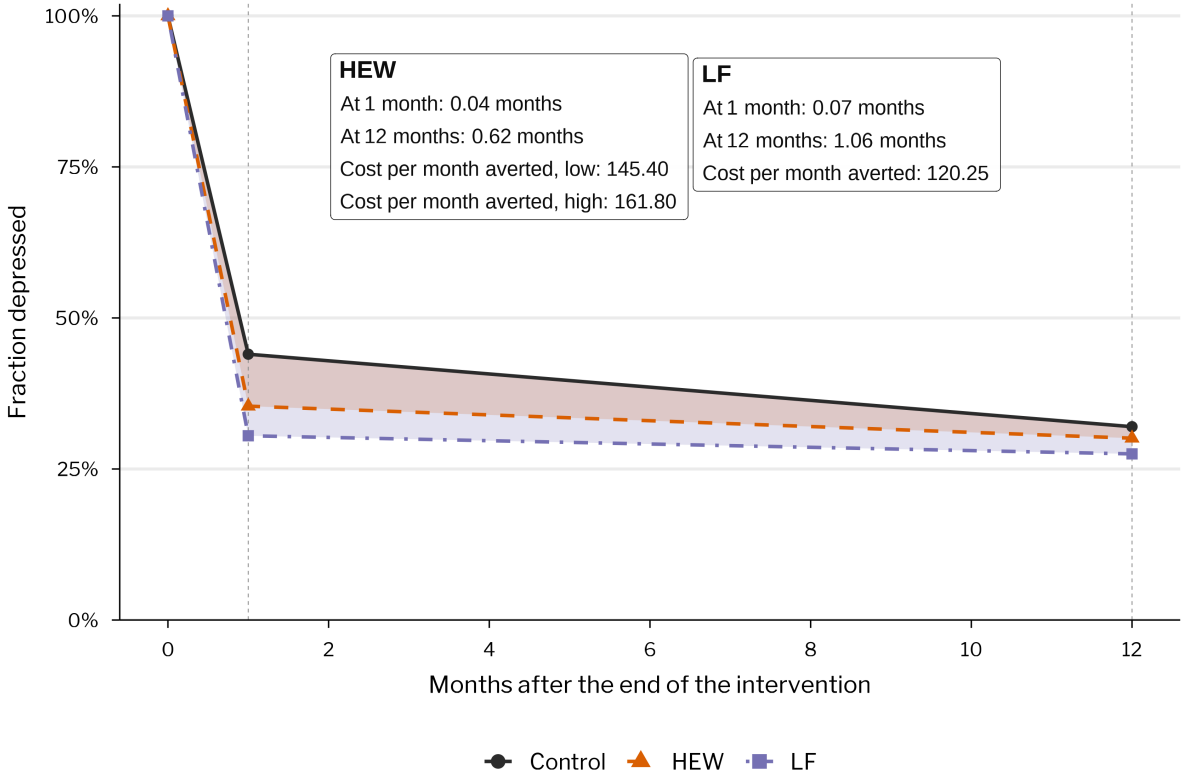


Figure A4: Months of depression averted




Note: Cost is reported in 2021 PPP-adjusted USD, and denotes the cost per month of depression averted. For the HEW arm, the low and high estimates value HEW time at zero and at their imputed wage, respectively.

Table A1: Definition of primary outcomes

| Outcome | Definition | Range |
|---|---|---|
| **Primary mental health index** | | |
| Mental health index | Average of the standardized PHQ-9, PSS-10, abbreviated PCL, and GAD-7 scores. Each component is standardized using the baseline control-group mean and standard deviation; higher values denote worse mental health. | |
| **Adult mental health symptoms** | | |
| Patient Health Questionnaire (PHQ-9) score | Total score from 9 PHQ items. Each item is coded from 0 = not at all to 3 = nearly every day, and the score sums all 9 items. | [0, 27] |
| Moderate to severe depression | Binary indicator equal to 1 when the PHQ-9 score is at least 10, and 0 otherwise. | [0, 1] |
| Perceived Stress Scale (PSS-10) score | Total score from 10 PSS items. Items are coded from 0 = never to 4 = very often; positively phrased items are reverse coded before summing. | [0, 40] |
| Post-traumatic stress score | Total score from 8 abbreviated PCL items. Items are coded from 0 = not at all to 4 = extremely, and the score sums all 8 items. | [0, 32] |
| Generalized Anxiety Disorder (GAD-7) score | Total score from 7 GAD items. Each item uses the same 0 to 3 response scale as the PHQ-9, and the score sums all 7 items. | [0, 21] |
| **Child mental health** | | |
| SDQ subscale scores | Emotional problems, conduct problems, hyperactivity problems, and peer problems subscales. Each subscale sums 5 SDQ items coded from 0 to 2 after applying the SDQ reverse-coding rules where relevant. | [0, 10] |
| Total SDQ score | Total Strengths and Difficulties Questionnaire score for children aged 4–15 years, calculated as the sum of the emotional, conduct, hyperactivity, and peer-problems subscales. | [0, 40] |

Table A2: Definition of secondary and exploratory outcomes

| Outcome | Definition | Range |
|---|---|---|
| **Secondary outcome indices** | | |
| Psychosocial index | Average of standardized self-efficacy, reverse-coded WHODAS percentage score, tension reduction, Brief COPE, and MSPSS scores. Components are standardized using the baseline control group where available; higher values denote better outcomes. | |
| Behavioral index | Average of standardized reverse-coded alcohol use at least twice per week and reverse-coded IPV indicators, so higher values denote less risky behavior or exposure. | |
| Economic index | Average of standardized productive-activity participation, reverse-coded moderate or severe food insecurity, any savings, and winsorized total savings. Higher values denote better economic outcomes. | |
| Childcare index | Average of standardized childcare activities for children aged 0–5, childcare activities for children aged 6–15, reverse-coded child neglect, and reverse-coded child discipline. Higher values denote better childcare outcomes. | |
| **Psychosocial outcomes** | | |
| General Self-Efficacy Scale | Average response across 8 self-efficacy items, each coded from 1 = strongly disagree to 5 = strongly agree. | [1, 5] |
| WHODAS score (%) | WHO Disability Assessment Schedule 2.0 score from 12 items coded from 1 = none to 5 = extreme or cannot do, summed and expressed as a percentage of the maximum possible score. Higher values denote greater functional impairment. | [20, 100] |
| Tension reduction | Total score from 9 tension-reduction items coded from 0 = not at all to 4 = all the time. | [0, 36] |
| Brief COPE score | Total score from 10 Brief COPE items covering problem-focused, emotion-focused, and avoidant coping; avoidant items are reverse coded before summing. | [10, 40] |
| MSPSS | Total Multidimensional Scale of Perceived Social Support score from 12 items coded from 1 = strongly disagree to 5 = strongly agree. | [12, 60] |
| **Behavioral and cognitive outcomes** | | |
| At least 2 drinks/week | Binary indicator equal to 1 if the respondent reports drinking alcohol at least twice per week, and 0 otherwise. | [0, 1] |
| Any IPV | Binary indicator equal to 1 if any reported intimate partner violence item in Section K2 is endorsed, and 0 otherwise. | [0, 1] |
| Forward digit span | Highest digit-span item answered correctly, with respondents who miss the first item coded as 0. | [0, 10] |

Definition of secondary and exploratory outcomes (continued)

| Outcome | Definition | Range |
|---|---|---|
| **Economic outcomes** | | |
| Any productive activity | Binary indicator equal to 1 if the respondent participated in any productive income-generating activity in the last 30 days, including business, livestock, farming, regular wage work, or casual wage work. | [0, 1] |
| Moderate or severe food insecurity | Binary indicator equal to 1 if the household is classified as moderately or severely food insecure based on the Food Insecurity Experience Scale items. | [0, 1] |
| Have any savings | Binary indicator equal to 1 if the respondent reports any positive savings in RuSACCO, VSLA, bank, or home savings. | [0, 1] |
| Savings (ETB) | Total value of reported savings in RuSACCO, VSLA, bank, and home savings, winsorized at the 98th percentile. | ETB |
| Productive hours | Total productive labor hours from work-activity questions, calculated as days worked times typical hours per day for regular wage work, casual wage work, non-farm business, livestock, and crop farming. Wage, business, and livestock components use the past-30-day recall window; crop-farming components use the past Mehr/Meher season and are divided by 3 to express them as month-equivalent hours. | Hours |
| **Childcare outcomes** | | |
| Childcare activities, 0–5 years old | Sum of 7 binary indicators for activities with children aged 0–5: reading, singing, taking outside, playing, drawing or naming/counting, bathing, and feeding. | [0, 7] |
| Childcare activities, 6–15 years old | Sum of 5 binary indicators for activities with children aged 6–15: helping prepare for school, helping with schoolwork, eating together, telling stories or reading, and playing. | [0, 5] |
| Child neglect, 4–15 years old | Sum of 5 binary indicators of neglect toward children aged 4–15. | [0, 5] |
| Any violent child discipline, 4–15 years old | Binary indicator equal to 1 if any psychological aggression or any physical assault toward children aged 4–15 was reported in the past 30 days, and 0 if neither was reported. | [0, 1] |
| **HEW access and services** | | |
| Any core HEW service | Binary indicator equal to 1 if the household reports receiving any core HEW service, including medical exam, vaccination, other medical services, health education, counseling, or hygiene and sanitation. | [0, 1] |
| Accessed care for physical ailment | Binary indicator equal to 1 if the household accessed care for a physical ailment. | [0, 1] |
| Physical ailment: HEW | Binary indicator equal to 1 if care for a physical ailment was accessed from a HEW; coded as 0 when no physical-ailment care was reported or care came from another provider. | [0, 1] |
| Accessed routine care | Binary indicator equal to 1 if the household accessed routine care. | [0, 1] |
| Routine care: HEW | Binary indicator equal to 1 if routine care was accessed from a HEW; coded as 0 when no routine care was reported or care came from another provider. | [0, 1] |

Table A3: Household survey baseline balance

| | HEW+M | HEW+F | LFF+M | LFF+F | LFM+M | Control mean | p-value joint hyp on all T groups | N |
|---|---|---|---|---|---|---|---|---|
| Respondent's age (years) | -0.572 | -0.866 | 0.121 | 0.896 | 0.799 | 45.28 | 0.109 | 3,477 |
| | (0.656) | (0.706) | (0.790) | (0.790) | (0.623) | | | |
| Respondent married | -0.019 | -0.001 | 0.007 | -0.007 | 0.006 | 0.64 | 0.894 | 3,477 |
| | (0.022) | (0.033) | (0.024) | (0.036) | (0.020) | | | |
| Years of education (respondent) | 0.074 | 0.342* | -0.067 | -0.399* | -0.197 | 2.12 | 0.045 | 3,477 |
| | (0.231) | (0.194) | (0.236) | (0.228) | (0.198) | | | |
| Household size | -0.324** | -0.251* | -0.015 | 0.004 | -0.143 | 5.60 | 0.058 | 3,477 |
| | (0.148) | (0.128) | (0.171) | (0.161) | (0.147) | | | |
| HH has an improved roof | 0.016 | 0.010 | -0.002 | -0.044 | -0.023 | 0.64 | 0.772 | 3,477 |
| | (0.034) | (0.032) | (0.038) | (0.044) | (0.035) | | | |
| HH has improved flooring | 0.007 | -0.013 | 0.030 | -0.024 | 0.026 | 0.15 | 0.826 | 3,477 |
| | (0.024) | (0.029) | (0.028) | (0.042) | (0.026) | | | |
| Number of times met a HEW last year | 0.013 | -0.024 | 0.231 | -0.381* | 0.014 | 2.03 | 0.319 | 3,476 |
| | (0.190) | (0.187) | (0.232) | (0.197) | (0.179) | | | |
| PHQ-9 total score | -0.058 | -0.208 | 0.062 | 0.130 | -0.078 | 12.95 | 0.423 | 3,477 |
| | (0.136) | (0.146) | (0.156) | (0.157) | (0.136) | | | |
| Anxiety (GAD-7) score | -0.166 | -0.416 | -0.065 | 0.016 | -0.255 | 9.07 | 0.586 | 3,476 |
| | (0.260) | (0.265) | (0.309) | (0.289) | (0.269) | | | |
| Stress (PSS) score | 0.726** | -0.266 | 0.616* | -0.009 | 0.529* | 21.05 | 0.241 | 3,477 |
| | (0.319) | (0.376) | (0.327) | (0.424) | (0.313) | | | |
| PTSD (PCL-5) score | 0.697 | -0.224 | -0.217 | 0.254 | 0.365 | 12.77 | 0.538 | 3,477 |
| | (0.437) | (0.494) | (0.578) | (0.521) | (0.434) | | | |
| Child's SDQ score | 0.103 | -0.457 | 0.117 | -0.881** | 0.287 | 9.73 | 0.445 | 2,636 |
| | (0.361) | (0.401) | (0.372) | (0.445) | (0.350) | | | |
| Displaced | 0.063** | 0.015 | 0.006 | -0.024 | 0.040 | 0.21 | 0.349 | 3,477 |
| | (0.031) | (0.031) | (0.034) | (0.040) | (0.029) | | | |
| Armed group present during the conflict | -0.011 | 0.023 | -0.082* | -0.085** | -0.046 | 0.80 | 0.035 | 3,471 |
| | (0.035) | (0.036) | (0.044) | (0.037) | (0.039) | | | |
| Injury/death due to the conflict | 0.036 | -0.021 | -0.007 | -0.075** | 0.015 | 0.23 | 0.148 | 3,476 |
| | (0.026) | (0.030) | (0.032) | (0.037) | (0.027) | | | |
| p-val for joint hyp on $\beta$ | .038 | .252 | .519 | .12 | .171 | | | |

*Notes:* Each cell reports the coefficient from an OLS regression of the baseline characteristic (row) on indicators for the five treatment-by-respondent-gender groups (HEW, female LF, and male LF, each interacted with respondent gender), a respondent-gender control, and baseline-sub-district fixed effects; standard errors clustered at the baseline village are in parentheses ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$). The Control-mean column reports the mean of the characteristic among control-arm respondents. The penultimate column reports the p-value for a joint test that all five treatment-group coefficients are zero for that characteristic; the final row reports, for each treatment group, the p-value for a joint test that its coefficient is zero across all characteristics (estimated jointly by seemingly unrelated regression). Following Kerwin et al. (2024), we also report an omnibus randomization-inference test of the joint orthogonality of treatment assignment and all baseline characteristics, which yields p = 0.108 (any treatment), p = 0.304 (HEW vs. control), and p = 0.008 (LF vs. control), based on 500 permutations of the village-level assignment within sub-district strata.

Table A4: Facilitator characteristics

| | N | HEWs | Female LFs | Male LFs | HEWs vs Female LFs | HEWs vs Male LFs |
|---|---|---|---|---|---|---|
| **Panel A: Comparing HEWs and LFs** | | | | | | |
| Age of respondent | 150 | 34.89 | 25.07 | 30.41 | 0.00 | 0.00 |
| | | (7.60) | (4.13) | (5.95) | | |
| Years educ. | 150 | 12.73 | 14.27 | 13.67 | 0.01 | 0.10 |
| | | (2.43) | (3.18) | (3.52) | | |
| Any prior work experience | 150 | 0.06 | 0.07 | 0.20 | 0.85 | 0.04 |
| | | (0.25) | (0.26) | (0.40) | | |
| Currently married | 150 | 0.60 | 0.24 | 0.50 | 0.00 | 0.29 |
| | | (0.49) | (0.43) | (0.51) | | |
| Count index: assets | 150 | 9.84 | 5.80 | 5.96 | 0.00 | 0.00 |
| | | (6.80) | (3.48) | (4.86) | | |
| Self-efficacy | 150 | 52.25 | 51.32 | 53.15 | 0.40 | 0.47 |
| | | (5.08) | (6.10) | (7.83) | | |
| Index: Knowledge mental illness | 150 | 11.40 | 10.80 | 11.22 | 0.11 | 0.58 |
| | | (1.84) | (1.81) | (1.56) | | |
| Index: Pro-social motivation | 150 | 51.65 | 52.73 | 53.26 | 0.37 | 0.17 |
| | | (6.57) | (4.74) | (5.04) | | |
| Inclusion of others in self (IOS) scale | 150 | 6.24 | 5.27 | 5.35 | 0.00 | 0.00 |
| | | (1.19) | (1.45) | (1.39) | | |
| Dictator game: Amount retained | 150 | 21.03 | 23.90 | 21.20 | 0.37 | 0.96 |
| | | (16.07) | (14.30) | (15.75) | | |
| PHQ-9 total score | 150 | 4.89 | 4.68 | 5.00 | 0.79 | 0.89 |
| | | (4.33) | (4.57) | (4.08) | | |
| **Panel B: LF Characteristics** | | | | | | |
| Years 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. HEW = Health Extension Worker, LFF = female local facilitator, and LFM = male local facilitator. P-values in Panel A come from tests of mean differences between HEWs and local facilitators, controlling for stratification level.

Table A5: Primary Indices

| | (1) | (2) | (3) |
|---|---|---|---|
| | Mental health index | | Child development score |
| | 1-month follow-up | 12-month follow-up | 12-month follow-up |
| HEW x Female | -0.166** | 0.092 | 0.025 |
| | (0.065) | (0.071) | (0.092) |
| HEW x Male | -0.049 | -0.025 | 0.056 |
| | (0.056) | (0.071) | (0.072) |
| LF x Female | -0.348*** | -0.092 | 0.156 |
| | (0.080) | (0.096) | (0.095) |
| LF x Male | -0.088* | -0.029 | 0.057 |
| | (0.052) | (0.062) | (0.062) |
| Average effect on M | -0.076 | -0.028 | 0.057 |
| | (0.048) | (0.059) | (0.059) |
| Average effect on F | -0.237*** | 0.020 | 0.077 |
| | (0.061) | (0.069) | (0.080) |
| p-value (M = F) | 0.046 | 0.636 | 0.852 |
| Control mean | -0.57 | -0.90 | 0.10 |
| $N$ | 3,429 | 3,391 | 2,702 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the follow-up outcome on HEW and LF treatment indicators interacted with respondent gender, respondent gender, the baseline value of the outcome, and baseline-sub-district fixed effects. Missing baseline values are imputed to zero with a missingness indicator. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Average effects on M and F refer to male and female respondents, respectively, and weight the HEW and LF effects by treatment shares within each respondent gender. Outcome definitions are provided in Table A1 and Table A2.

Table A6: SDQ Raw Scores

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Emotional problems (out of 10) | Conduct problems (out of 10) | Hyperactivity problems (out of 10) | Peer problems (out of 10) | Total SDQ score (out of 40) |
| HEW x Female | 0.184 | -0.002 | 0.061 | -0.183 | 0.112 |
| | (0.162) | (0.118) | (0.133) | (0.118) | (0.410) |
| | [1.000] | [1.000] | [1.000] | [1.000] | [1.000] |
| HEW x Male | 0.100 | 0.071 | 0.129 | -0.035 | 0.247 |
| | (0.123) | (0.096) | (0.113) | (0.108) | (0.322) |
| | [1.000] | [1.000] | [1.000] | [1.000] | [1.000] |
| Female LF x Female | 0.199 | 0.166 | 0.139 | 0.104 | 0.689 |
| | (0.170) | (0.124) | (0.143) | (0.129) | (0.425) |
| | [1.000] | [1.000] | [1.000] | [1.000] | [1.000] |
| Female LF x Male | 0.119 | -0.027 | 0.087 | 0.020 | 0.179 |
| | (0.131) | (0.101) | (0.129) | (0.133) | (0.360) |
| | [1.000] | [1.000] | [1.000] | [1.000] | [1.000] |
| Male LF x Male | 0.124 | 0.030 | 0.275*** | -0.107 | 0.288 |
| | (0.107) | (0.091) | (0.103) | (0.098) | (0.297) |
| | [1.000] | [1.000] | [0.329] | [1.000] | [1.000] |
| Control mean | 1.67 | 2.62 | 4.02 | 1.86 | 10.17 |
| *N* | 2,702 | 2,702 | 2,702 | 2,702 | 2,702 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the follow-up outcome on HEW, female LF, and male LF treatment indicators interacted with respondent gender where applicable, respondent gender, the baseline value of the outcome, and baseline-sub-district fixed effects. Missing baseline values are imputed to zero with a missingness indicator. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Sharpened two-stage q-values as described in (Anderson, 2008) are reported in square brackets. Outcome definitions are provided in Table A1 and Table A2.

Table A7: Psychosocial Outcomes

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| **Panel A: 1-month follow-up** | | | | | |
| | Self-efficacy | WHODAS score (%) | Tension reduction | Brief COPE score | MSPSS |
| HEW x Female | 0.147*** | -2.828*** | 1.279** | 0.213 | 1.417*** |
| | (0.057) | (0.916) | (0.611) | (0.289) | (0.434) |
| | [0.020] | [0.006] | [0.047] | [0.213] | [0.005] |
| HEW x Male | 0.108** | 1.273 | 1.606*** | 0.432 | 0.562 |
| | (0.049) | (0.836) | (0.422) | (0.297) | (0.479) |
| | [0.040] | [0.088] | [0.002] | [0.093] | [0.138] |
| Female LF x Female | 0.088 | -3.780*** | 1.525** | 0.462 | 0.995* |
| | (0.083) | (1.073) | (0.675) | (0.330) | (0.575) |
| | [0.158] | [0.003] | [0.035] | [0.097] | [0.076] |
| Female LF x Male | 0.090 | 1.395 | 1.928*** | 0.746** | 0.900 |
| | (0.062) | (0.867) | (0.452) | (0.364) | (0.587) |
| | [0.093] | [0.081] | [0.001] | [0.048] | [0.088] |
| Male LF x Male | 0.152*** | -0.487 | 1.369*** | 0.604** | 0.540 |
| | (0.047) | (0.844) | (0.406) | (0.254) | (0.408) |
| | [0.005] | [0.227] | [0.004] | [0.030] | [0.108] |
| Control mean | 3.35 | 40.22 | 11.97 | 28.27 | 37.99 |
| *N* | 3,429 | 3,429 | 3,429 | 3,429 | 3,429 |
| **Panel B: 12-month follow-up** | | | | | |
| HEW x Female | 0.038 | 0.590 | 0.421 | 0.155 | 0.394 |
| | (0.072) | (1.011) | (0.533) | (0.257) | (0.524) |
| | [1.000] | [1.000] | [1.000] | [1.000] | [1.000] |
| HEW x Male | 0.058 | 0.685 | 0.450 | -0.049 | 0.419 |
| | (0.062) | (0.810) | (0.489) | (0.263) | (0.519) |
| | [1.000] | [1.000] | [1.000] | [1.000] | [1.000] |
| Female LF x Female | 0.041 | -0.562 | 0.503 | 0.278 | 0.456 |
| | (0.084) | (1.039) | (0.621) | (0.323) | (0.647) |
| | [1.000] | [1.000] | [1.000] | [1.000] | [1.000] |
| Female LF x Male | 0.114* | -0.060 | 0.091 | 0.111 | 0.722 |
| | (0.068) | (0.900) | (0.550) | (0.343) | (0.582) |
| | [0.679] | [1.000] | [1.000] | [1.000] | [1.000] |
| Male LF x Male | 0.049 | 0.696 | 0.017 | -0.174 | 0.048 |
| | (0.058) | (0.709) | (0.454) | (0.259) | (0.524) |
| | [1.000] | [1.000] | [1.000] | [1.000] | [1.000] |
| Control mean | 3.33 | 37.39 | 12.19 | 28.62 | 39.12 |
| *N* | 3,391 | 3,391 | 3,391 | 3,391 | 3,391 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the follow-up outcome on HEW, female LF, and male LF treatment indicators interacted with respondent gender where applicable, respondent gender, the baseline value of the outcome (except for column 3), and baseline-sub-district fixed effects. Missing baseline values are imputed to zero with a missingness indicator. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Sharpened two-stage q-values as described in (Anderson, 2008) are reported in square brackets. Outcome definitions are provided in Table A1 and Table A2.

Table A8: Cognitive and Behavioral Outcomes

| | (1) | (2) | (3) |
|---|---|---|---|
| **Panel A: 1-month follow-up** | | | |
| | At least 2 drinks/week | Any IPV | Forward digit span |
| HEW x Female | -0.015 | -0.002 | 0.056 |
| | (0.014) | (0.032) | (0.038) |
| | [1.000] | [1.000] | [1.000] |
| HEW x Male | 0.017 | 0.009 | -0.002 |
| | (0.025) | (0.027) | (0.045) |
| | [1.000] | [1.000] | [1.000] |
| Female LF x Female | -0.013 | -0.015 | -0.005 |
| | (0.015) | (0.035) | (0.047) |
| | [1.000] | [1.000] | [1.000] |
| Female LF x Male | 0.021 | 0.034 | -0.028 |
| | (0.025) | (0.030) | (0.051) |
| | [1.000] | [1.000] | [1.000] |
| Male LF x Male | 0.002 | -0.003 | -0.007 |
| | (0.023) | (0.025) | (0.041) |
| | [1.000] | [1.000] | [1.000] |
| Control mean | 0.04 | 0.17 | 2.47 |
| $N$ | 3,429 | 1,803 | 3,429 |
| **Panel B: 12-month follow-up** | | | |
| HEW x Female | -0.016 | 0.032 | -0.003 |
| | (0.014) | (0.030) | (0.038) |
| | [0.880] | [0.880] | [1.000] |
| HEW x Male | -0.012 | 0.008 | -0.058* |
| | (0.027) | (0.020) | (0.035) |
| | [1.000] | [1.000] | [0.713] |
| Female LF x Female | -0.015 | -0.009 | -0.017 |
| | (0.019) | (0.039) | (0.048) |
| | [1.000] | [1.000] | [1.000] |
| Female LF x Male | -0.028 | 0.034 | -0.027 |
| | (0.027) | (0.025) | (0.046) |
| | [0.880] | [0.880] | [1.000] |
| Male LF x Male | -0.039 | -0.006 | -0.010 |
| | (0.024) | (0.018) | (0.037) |
| | [0.713] | [1.000] | [1.000] |
| Control mean | 0.05 | 0.13 | 2.54 |
| $N$ | 3,391 | 2,322 | 3,391 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the follow-up outcome on HEW, female LF, and male LF treatment indicators interacted with respondent gender where applicable, respondent gender, the baseline value of the outcome (except for column 3), and baseline-sub-district fixed effects. Missing baseline values are imputed to zero with a missingness indicator. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Sharpened two-stage q-values as described in (Anderson, 2008) are reported in square brackets. Outcome definitions are provided in Table A1 and Table A2.

Table A9: Economic Outcomes

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| **Panel A: 1-month follow-up** | | | | |
| | Any productive activity | Moderate / severe food insecurity | Have any savings | Savings (ETB) |
| HEW x Female | 0.012 | 0.007 | -0.013 | -98.636 |
| | (0.022) | (0.016) | (0.031) | (80.009) |
| | [1.000] | [1.000] | [1.000] | [1.000] |
| HEW x Male | 0.014 | 0.007 | 0.009 | -50.081 |
| | (0.017) | (0.018) | (0.029) | (72.736) |
| | [1.000] | [1.000] | [1.000] | [1.000] |
| Female LF x Female | 0.005 | 0.033* | -0.022 | -96.082 |
| | (0.029) | (0.019) | (0.035) | (77.569) |
| | [1.000] | [0.851] | [1.000] | [1.000] |
| Female LF x Male | -0.014 | 0.000 | 0.043 | 82.804 |
| | (0.019) | (0.026) | (0.033) | (87.868) |
| | [1.000] | [1.000] | [1.000] | [1.000] |
| Male LF x Male | -0.014 | 0.006 | -0.003 | -74.146 |
| | (0.017) | (0.017) | (0.025) | (68.909) |
| | [1.000] | [1.000] | [1.000] | [1.000] |
| Control mean | 0.86 | 0.90 | 0.40 | 402.86 |
| *N* | 3,429 | 3,429 | 3,429 | 3,429 |
| **Panel B: 12-month follow-up** | | | | |
| HEW x Female | 0.002 | -0.005 | -0.030 | -148.409 |
| | (0.012) | (0.028) | (0.029) | (114.617) |
| | [1.000] | [1.000] | [1.000] | [1.000] |
| HEW x Male | -0.005 | -0.031 | -0.016 | -39.359 |
| | (0.005) | (0.030) | (0.028) | (163.298) |
| | [1.000] | [1.000] | [1.000] | [1.000] |
| Female LF x Female | 0.006 | 0.024 | -0.050 | 87.817 |
| | (0.012) | (0.032) | (0.035) | (160.331) |
| | [1.000] | [1.000] | [0.919] | [1.000] |
| Female LF x Male | -0.006 | -0.078** | -0.001 | -15.509 |
| | (0.009) | (0.039) | (0.031) | (203.897) |
| | [1.000] | [0.263] | [1.000] | [1.000] |
| Male LF x Male | -0.013** | -0.009 | 0.023 | -116.847 |
| | (0.006) | (0.028) | (0.026) | (158.240) |
| | [0.263] | [1.000] | [1.000] | [1.000] |
| Control mean | 0.96 | 0.79 | 0.63 | 901.46 |
| *N* | 3,391 | 3,391 | 3,391 | 3,391 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the follow-up outcome on HEW, female LF, and male LF treatment indicators interacted with respondent gender where applicable, respondent gender, the baseline value of the outcome, and baseline-sub-district fixed effects. Missing baseline values are imputed to zero with a missingness indicator. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Sharpened two-stage q-values as described in (Anderson, 2008) are reported in square brackets. Outcome definitions are provided in Table A1 and Table A2.

Table A10: Childcare Outcomes

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| **Panel A: 1-month follow-up** | | | | |
| | Childcare activities | | Child neglect and discipline | |
| | 0-5 years old | 6-15 years old | Neglect | Discipline |
| HEW x Female | -0.101 | 0.166* | -0.169 | -0.004 |
| | (0.186) | (0.097) | (0.103) | (0.030) |
| | [1.000] | [0.613] | [0.613] | [1.000] |
| HEW x Male | 0.113 | 0.054 | -0.047 | -0.025 |
| | (0.171) | (0.072) | (0.097) | (0.031) |
| | [1.000] | [1.000] | [1.000] | [1.000] |
| Female LF x Female | 0.007 | 0.047 | 0.117 | -0.018 |
| | (0.208) | (0.079) | (0.123) | (0.031) |
| | [1.000] | [1.000] | [0.948] | [1.000] |
| Female LF x Male | -0.199 | 0.088 | 0.132 | -0.047 |
| | (0.189) | (0.080) | (0.105) | (0.034) |
| | [0.835] | [0.835] | [0.720] | [0.708] |
| Male LF x Male | -0.245* | 0.016 | 0.048 | -0.043 |
| | (0.147) | (0.060) | (0.084) | (0.028) |
| | [0.613] | [1.000] | [1.000] | [0.671] |
| Control mean | 3.76 | 1.39 | 1.54 | 0.78 |
| $N$ | 1,919 | 2,651 | 2,684 | 2,682 |
| **Panel B: 12-month follow-up** | | | | |
| HEW x Female | -0.301* | 0.002 | -0.158 | 0.015 |
| | (0.162) | (0.098) | (0.106) | (0.027) |
| | [0.314] | [1.000] | [0.453] | [1.000] |
| HEW x Male | -0.210 | -0.159 | -0.092 | 0.024 |
| | (0.182) | (0.103) | (0.095) | (0.030) |
| | [0.710] | [0.453] | [0.859] | [0.897] |
| Female LF x Female | -0.190 | -0.103 | 0.148 | 0.024 |
| | (0.202) | (0.126) | (0.114) | (0.031) |
| | [0.859] | [0.897] | [0.659] | [0.897] |
| Female LF x Male | 0.025 | 0.003 | 0.020 | 0.019 |
| | (0.201) | (0.121) | (0.111) | (0.032) |
| | [1.000] | [1.000] | [1.000] | [1.000] |
| Male LF x Male | -0.073 | -0.166* | 0.046 | 0.050* |
| | (0.165) | (0.091) | (0.082) | (0.027) |
| | [1.000] | [0.314] | [1.000] | [0.314] |
| Control mean | 4.37 | 1.95 | 1.62 | 0.81 |
| $N$ | 1,815 | 2,618 | 2,702 | 2,701 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the follow-up outcome on HEW, female LF, and male LF treatment indicators interacted with respondent gender where applicable, respondent gender, the baseline value of the outcome, and baseline-sub-district fixed effects. Missing baseline values are imputed to zero with a missingness indicator. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Sharpened two-stage q-values as described in (Anderson, 2008) are reported in square brackets. Outcome definitions are provided in Table A1 and Table A2.

Table A11: HEW services

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| **Panel A: 1-month follow-up** | | | | | |
| | Any core HEW service | Accessed care for physical ailment | Physical ailment: HEW | Accessed routine care | Routine care: HEW |
| HEW x Female | -0.015 | 0.028 | 0.016 | 0.018 | 0.002 |
| | (0.032) | (0.029) | (0.016) | (0.020) | (0.006) |
| | [1.000] | [0.598] | [0.598] | [0.623] | [1.000] |
| HEW x Male | -0.001 | -0.057* | 0.041* | -0.010 | 0.029*** |
| | (0.030) | (0.031) | (0.022) | (0.022) | (0.010) |
| | [1.000] | [0.296] | [0.296] | [1.000] | [0.047] |
| Female LF x Female | -0.058 | -0.003 | -0.005 | 0.020 | -0.002 |
| | (0.038) | (0.033) | (0.014) | (0.024) | (0.006) |
| | [0.358] | [1.000] | [1.000] | [0.634] | [1.000] |
| Female LF x Male | -0.027 | 0.073** | 0.000 | 0.007 | 0.014 |
| | (0.035) | (0.035) | (0.017) | (0.029) | (0.008) |
| | [0.673] | [0.277] | [1.000] | [1.000] | [0.317] |
| Male LF x Male | -0.052* | -0.002 | -0.012 | -0.021 | 0.016*** |
| | (0.028) | (0.030) | (0.013) | (0.020) | (0.006) |
| | [0.296] | [1.000] | [0.623] | [0.598] | [0.047] |
| Control mean | 0.56 | 0.46 | 0.02 | 0.10 | 0.01 |
| *N* | 3,429 | 3,429 | 3,429 | 3,429 | 3,429 |
| **Panel B: 12-month follow-up** | | | | | |
| HEW x Female | 0.010 | 0.037 | -0.007 | 0.039 | -0.002 |
| | (0.033) | (0.024) | (0.012) | (0.027) | (0.005) |
| | [0.788] | [0.310] | [0.606] | [0.310] | [0.606] |
| HEW x Male | 0.016 | -0.037 | -0.029 | -0.059* | -0.002 |
| | (0.034) | (0.026) | (0.021) | (0.031) | (0.010) |
| | [0.619] | [0.310] | [0.332] | [0.160] | [0.808] |
| Female LF x Female | -0.033 | -0.024 | -0.012 | 0.032 | -0.006 |
| | (0.034) | (0.028) | (0.014) | (0.033) | (0.005) |
| | [0.550] | [0.550] | [0.550] | [0.550] | [0.425] |
| Female LF x Male | 0.026 | -0.004 | 0.014 | -0.072* | -0.019** |
| | (0.036) | (0.031) | (0.028) | (0.037) | (0.009) |
| | [0.550] | [0.834] | [0.606] | [0.160] | [0.135] |
| Male LF x Male | -0.029 | -0.025 | -0.037* | -0.072** | -0.007 |
| | (0.030) | (0.025) | (0.019) | (0.029) | (0.009) |
| | [0.550] | [0.550] | [0.160] | [0.094] | [0.550] |
| Control mean | 0.51 | 0.70 | 0.04 | 0.15 | 0.01 |
| *N* | 3,391 | 3,391 | 3,391 | 3,391 | 3,391 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the follow-up outcome on HEW, female LF, and male LF treatment indicators interacted with respondent gender where applicable, respondent gender, the baseline value of the outcome, and baseline-sub-district fixed effects. Missing baseline values are imputed to zero with a missingness indicator. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Sharpened two-stage q-values as described in (Anderson, 2008) are reported in square brackets. Outcome definitions are provided in Table A1 and Table A2.

Table A12: Primary Indices: Lasso-Selected Baseline Controls

| | (1) | (2) | (3) |
|---|---|---|---|
| | Mental health index | | Child development score |
| | 1-month follow-up | 12-month follow-up | 12-month follow-up |
| HEW x Female | -0.166** | 0.100 | 0.026 |
| | (0.065) | (0.071) | (0.094) |
| HEW x Male | -0.062 | -0.024 | 0.059 |
| | (0.056) | (0.069) | (0.072) |
| Female LF x Female | -0.359*** | -0.103 | 0.164* |
| | (0.078) | (0.093) | (0.096) |
| Female LF x Male | -0.052 | -0.078 | 0.039 |
| | (0.076) | (0.085) | (0.082) |
| Male LF x Male | -0.092* | -0.003 | 0.069 |
| | (0.053) | (0.064) | (0.066) |
| Average HEW effect | -0.115*** | 0.039 | 0.042 |
| | (0.041) | (0.047) | (0.054) |
| Weighted LF effect | -0.219*** | -0.066 | 0.111** |
| | (0.044) | (0.051) | (0.053) |
| Average Female LF effect | -0.206*** | -0.091 | 0.101* |
| | (0.048) | (0.057) | (0.060) |
| p-value (Avg HEW = Avg. Weighted LF) | 0.011 | 0.048 | 0.222 |
| p-value (Avg HEW = Avg Female LF) | 0.045 | 0.026 | 0.351 |
| Control mean | -0.57 | -0.90 | 0.10 |
| Lasso-selected baseline controls | 4 | 2 | 1 |
| *N* | 3,429 | 3,391 | 2,702 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column re-estimates the corresponding column of Table 1 adding baseline covariates selected by the post-double-selection lasso (Belloni et al., 2014) from a pool of 37 candidate baseline characteristics (51 terms once each is imputed to zero with a missingness indicator) spanning demographics, mental health, psychosocial skills, child development, intimate partner violence, economic activity, food security, savings, alcohol use, and prior contact with HEWs. The HEW, female LF, and male LF treatment indicators interacted with respondent gender, respondent gender, the baseline value of the outcome (imputed to zero with a missingness indicator), and baseline-sub-district fixed effects are always included and not subject to selection; the penultimate row reports the number of additional baseline controls selected. Standard errors clustered at the baseline-village level are in parentheses. The control mean is the mean of the follow-up outcome among control-arm female respondents (the omitted gender category). Reported average effects use realized composition shares. The weighted LF effect weights the female-LF female-respondent, female-LF male-respondent, and male-LF male-respondent coefficients by 0.50, 0.167, and 0.333 respectively, reflecting that female respondents are served only by female LFs and male respondents roughly one-third by female and two-thirds by male LFs. Outcome definitions are provided in Table A1 and Table A2.

Table A13: Definition of facilitator-related outcomes

| Outcome | Definition | Range |
|---|---|---|
| **Facilitator implementation and engagement** | | |
| Number of sessions conducted | Total number of core gPM+ sessions conducted across the two therapy groups assigned to a facilitator. | [0, 10] |
| Mean session attendance (conditional) | Average attendance percentage across conducted sessions and groups, calculated relative to 8 expected participants per session. | [0, 100] |
| Mean session attendance (unconditional) | Average attendance percentage across all planned sessions and groups, calculated relative to 8 expected participants per session and assigning 0 attendance to sessions that were not conducted. | [0, 100] |
| Attended refresher training | Binary indicator equal to 1 if the facilitator attended the post-trial refresher training. | [0, 1] |
| Facilitated new groups | Binary indicator equal to 1 if the facilitator formed or facilitated new gPM+ groups after the trial cycle. | [0, 1] |
| Motivation score | Total score from 12 gPM+-related motivation items, each coded on a 1 to 5 scale. | [12, 60] |
| **Facilitator quality and rapport** | | |
| Prior interaction with facilitator | Binary indicator equal to 1 if the respondent reports having interacted with the facilitator before gPM+. | [0, 1] |
| Highly satisfied with provider | Binary indicator equal to 1 if the respondent reports being highly satisfied with the facilitator. | [0, 1] |
| Comfortable speaking with facilitator | Binary indicator equal to 1 if the respondent reports being comfortable or very comfortable speaking with the facilitator. | [0, 1] |
| Comfortable because they knew facilitator | Binary indicator equal to 1 if the respondent reports being comfortable because they already knew the facilitator. | [0, 1] |

Table A14: Facilitator engagement post-trial

| | (1) | (2) | (3) |
|---|---|---|---|
| | Attended refresher training | Facilitated new groups | Motivation score |
| Male LF | -0.071 | 0.837*** | 4.573*** |
| | (0.059) | (0.058) | (1.071) |
| | [0.103] | [0.001] | [0.001] |
| Female LF | 0.023 | 0.927*** | 2.444* |
| | (0.040) | (0.042) | (1.275) |
| | [0.233] | [0.001] | [0.045] |
| Average LF effect | -0.027 | 0.879*** | 3.578*** |
| | (0.042) | (0.039) | (1.008) |
| Control mean | 0.95 | 0.02 | 47.69 |
| $N$ | 136 | 136 | 136 |

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HEW denotes health extension worker and LF denotes local facilitator. Each column reports an OLS regression of the facilitator post-endline survey outcome on male and female LF indicators, with HEWs as the reference group. Heteroskedasticity-robust standard errors are in parentheses. The control mean is the mean among HEWs. The average LF effect weights the male- and female-LF coefficients by their shares among LFs in the relevant source dataset. Sharpened two-stage q-values as described in (Anderson, 2008) are reported in square brackets. Outcome definitions are provided in Table A1 and Table A2.

Table A15: Depression-screening inclusion criteria in comparator psychotherapy trials

| Study | Country | Intervention | Enrollment screening criterion |
|---|---|---|---|
| Haushofer, Mudida & Shapiro (2020) | Kenya | PM+ (vs. cash transfer) | No depression screen; poverty-targeted sample, not screened for psychological distress |
| Acarturk et al. (2022) | Turkey | gPM+ (pilot) | Kessler-10 $> 15$ *and* WHO-DAS $> 16$ |
| Greene et al. (2024) | Colombia | gPM+ (pilot) | GHQ-12 $> 2$ *and* WHODAS $> 16$ (women only) |
| Patel et al. (2017); Weobong et al. (2017); Bhat et al. (2022) | India (Goa) | Healthy Activity Program (individual) | PHQ-9 $> 14$ (equivalently $\geq 15$) |
| Vanobberghen et al. (2020); Gibbs et al. (2025) | Pakistan | Thinking Healthy Programme, peer-delivered | PHQ-9 $\geq 10$; pregnant women aged $\geq 18$ |
| Baird et al. (2025) | Uganda | Interpersonal therapy in groups (adolescents) | PHQ-8 $\geq 10$ |
| This study | Ethiopia (Tigray) | gPM+ | PHQ-9 between 10 and 19 (inclusive) |

Notes: The table reports the clinical eligibility threshold used to enroll participants in each comparator trial benchmarked in the cost-effectiveness discussion. Instruments: PHQ-9 / PHQ-8 (Patient Health Questionnaire, 9- and 8-item depression scales); K10 (Kessler-10 psychological distress scale); GHQ-12 (12-item General Health Questionnaire); WHODAS (WHO Disability Assessment Schedule, functional impairment). Two comparators did not screen on a depression-severity threshold: the Kenya PM+ trial enrolled a poverty-targeted sample without any psychological-distress screen, and the Turkey and Colombia pilots screened jointly on psychological distress and functional impairment rather than on a PHQ depression cutoff. For the Pakistan THPP trial, Vanobberghen et al. (2020) report a PHQ-9 $\geq 10$ screen while Gibbs et al. (2025) frame eligibility as a current major depressive episode established by structured clinical interview. The Healthy Activity Program cutoff appears as "PHQ-9 $> 14$" (Patel et al. 2017; Weobong et al. 2017) and equivalently as "PHQ-9 $\geq 15$" (Bhat et al. 2022).

## ALL IFPRI DISCUSSION PAPERS

All discussion papers are available **here**

They can be downloaded free of charge

**INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE**
www.ifpri.org

**IFPRI HEADQUARTERS**
1201 Eye Street, NW
Washington, DC 20005 USA
Tel.: +1-202-862-5600
Fax: +1-202-862-5606
Email: ifpri@cgiar.org