Sharing is caring: Redistributing unpaid domestic and care work improves mental well-being for fathers and mothers*

Shivani J. Gupta,[1a] Ellen B. McCullough,[1a] Tamara McGavock,[1b]

Thomas Woldu Assefa[c] and Tigabu Degu Getahun[d]

January 2026

Click here for latest posted version.

Across cultures, women still do most unpaid domestic and care work and also experience depression about twice as often as men, but evidence linking the two is typically correlational. In ultra-poor rural Ethiopia, we provide causal evidence: a light-touch phone-based behavior change intervention that engaged men in seven biweekly, personalized discussions reduced depression scores by 20% for both mothers and fathers. Treated men increased domestic chores by 25 minutes per day—a 50% rise from baseline but only 5% of the total household load. Women's absolute domestic workload did not fall, but their share declined by 10 percentage points as others, especially men, took on more of the load. Still, given their 70% baseline share, substantial disparities persist. Our results suggest that improving women's mental health need not require reducing their total work hours; engaging men around unpaid work can improve both partners' psychological well-being.

JEL codes: D13, I15, J16, O12

Keywords: mental well-being, unpaid domestic and care work, gender, behavior change communication

---

*Corresponding Author: Tamara McGavock (mcgavock@grinnell.edu). [1]First authorship is shared among Gupta, McCullough, & McGavock, who are listed alphabetically. [a]University of Georgia Agricultural and Applied Economics, [b]Grinnell College, [c]International Food Policy Research Institute (IFPRI), [d]University of Bonn and Policy Research Institute Ethiopia.

We thank our data collection team at EconInsight Center for Development Research, especially Abiy Samuel, Yordanos Tenaw, Senay Asres, and Filsen Yadete. We are grateful for funding from Strengthen PSNP4 Institutions and Resilience Development Food Security Activity II (SPIR II, an IFPRI program funded by USAID), Grinnell College, the University of Georgia, and the Center for the Economics of Human Development (CEHD) at the University of Chicago. We thank World Vision and CARE Ethiopia's gender staff for consultations and materials that informed the intervention. For helpful comments, we thank numerous seminar and conference participants. We thank Joyce Ceballos, Alair Ferguson-Hautzinger, Jungbin Im, Utsah Kalra, and Amedeus Dsouza for excellent research assistance. Pre-registered hypotheses are at the AEA RCT Registry (doi.org/10.1257/rct.11907-1.0) and a de-identified dataset will be on the Harvard Dataverse.

Highlights

- Engaging men on unpaid domestic and care work (UDCW) lowers depression scores.
- Seven personalized discussions w/ men on family and UDCW changed their behavior.
- Reducing women's UDCW load: not a necessary condition for depression prevention.
- Men's contributions increased by 25m per day (5% of household total).
- Sharing is caring: relative work shares matter for women's well-being.

1 Introduction

Economists increasingly recognize that mental health and poverty can reinforce one another. In their framework for "depression for economists," de Quidt & Haushofer (2016) model depression as a consequence of adverse socioeconomic shocks that reduce individuals' optimism about the returns to effort, leading to lower labor supply, reduced investment, and the potential for poverty traps. Yet the causal relationship may also run in the opposite direction, with individuals' mental health, and therefore economic outcomes, being shaped by the nature and burden of work itself—particularly when it is excessive, undervalued, or inequitably shared (Baranov et al., 2020). In this article, we ask two related questions. First, can work arrangements, including the allocation of unpaid labor within households, influence mental health? Second, can interventions that affect work arrangements improve psychological well-being? We provide new experimental evidence on the mental health impacts of unpaid domestic and care work (UDCW), a form of labor that accounts for an estimated 16.4 billion hours of work each day worldwide (Addati et al., 2018) but remains largely under-measured in economic analysis and is strongly shaped by social and gender norms across cultures (Dinkelman & Ngai, 2022; Browning & Gørtz, 2012; Borra et al., 2020; Bordalo et al., 2019). Our research takes place in rural Ethiopia, where extreme poverty and depression rates are high, gender norms are especially rigid, and access to mental health services is severely lacking (Fan et al., 2025). Our results demonstrate the preventative mental health benefits of a novel, light-touch behavioral intervention that uses low-cost technology to engage men in self-reflection about their family roles and to encourage small actions that deepen their support.

While causal evidence remains limited, systematic reviews in public health and psychology show strong correlations between UDCW burden and poor mental health outcomes (Ervin et al., 2025, 2022; Magaña et al., 2020). The negative link is especially strong for women, who perform three times as much UDCW as men on average globally (Addati et al., 2018; Boye, 2009) and face twice the lifetime risk of clinical depression and substantially higher rates of other mental health challenges like anxiety (Kuehner, 2017; Riecher-Rössler, 2017; Fisher et al., 2013). The majority of existing statistical evidence shows only cross-

sectional associations between time contributed to UDCW and mental health outcomes (Ervin et al., 2022). Though longitudinal evidence can improve this by controlling for individual-level characteristics (Ervin et al., 2023), the causal mechanism is still unclear because depressive symptoms can also affect cognition, executive functioning, efficiency, motivation, and even time perception (Andersen et al., 2021). Even quasi-experimental settings, such as COVID-induced school closures (Hammarberg et al., 2020; Xue & McMunn, 2021), cannot isolate changes in unpaid work from concurrent shocks to income, employment, or social connectedness. Insights from across disciplines underscore the intricate interplay between interpersonal relationships, psychosocial processes, subjective well-being, and socioeconomic conditions in shaping both the allocation of work within households and mental health outcomes. Yet, due to substantial challenges in measurement and policy design, relatively little is known about how to address these interconnected dynamics at scale to improve well-being.

In this article, we report on an experiment that encouraged fathers of young children (aged 2–5) in rural Ethiopia to share the load of their households' UDCW. Our study population is ultra-poor households who, lacking electricity or piped water, regularly perform arduous and time-consuming physical labor to meet basic household needs. At baseline, women in our sample carried 80% of their households' chore burden (share of minutes), and four of every ten fathers performed no task at all—not even a few moments of care for any child. Only 20% of male and female respondents experienced no symptoms of depression at baseline, and 40% experienced sufficiently numerous or severe symptoms to be flagged for clinical therapy by a standard screening tool (the PHQ-9). Our study population mirrors millions of households globally who are targeted by government and NGO anti-poverty programs. Multifaceted interventions frequently target women's outcomes in particular, seeking especially to empower them directly with small business loans or jobs, sometimes by discussing gender norms and UDCW burdens with both men and women (Ngai & Petrongolo, 2017; Rendall, 2018; Alderman et al., 2025). While on average men tend to increase engagement in UDCW when women engage more in economic activities (Hook, 2006), which is correlated with improved mental health for female spouses (Baek et al., 2024; Harryson et al., 2012; Glass & Fujimoto, 1994), programs promoting equitable labor divisions seldom measure mental health and they rarely intervene directly on men's behavior (Dinkelman & Ngai, 2022; Goldin,

2014; Blau, 2025).[1] Thus, we know little from causal evidence about how redistribution interventions affect time use, relationships, or mental health.

Our behavior change communication (BCC) intervention departs from previous approaches by engaging men directly in personalized conversations that prompted routine reflection on the status quo division of labor in their own households, their goals and influence of their own fathers on their identity as husbands and fathers, and critical consideration of whether norms about gendered tasks reflect innate differences in ability. By grounding these discussions in men's own experiences and aspirations over a thirteen-week timeline, we operationalized key elements of successful behavior change interventions—including the repeated setting of specific, achievable goals linked to new behaviors, which can build capabilities and update beliefs about the returns to effort (Petit, 2019; Sirois & Pychyl, 2013).

Our findings suggest substantially improved depression outcomes among treatment households relative to the control—the standard depression scores for clinical screening and monitoring (PHQ-9) are lower by more than 20% for both fathers and mothers. The effects are *preventative* rather than curative, since they appear for individuals with minimal symptoms at baseline, though we note that we lack power to detect effects for the most severely depressed. The mental health gains are also *complementary* for spousal pairs — the treatment effect for couples is exactly double the treatment effect for each spouse separately.

The intervention also induced men to participate more in unpaid domestic work, but we find no evidence of an absolute decline in women's total time spent on UDCW on average or an increase in men's contributions to childcare. Specifically, treated men are more likely to participate in domestic work by 16 percentage points, which is nearly 50% of the relatively low baseline participation rate. We also find that men increase their contributions on the intensive margin by a similar percentage gain (number of chores or minutes) relative to the control group, an effect that equates to about 25 minutes per day. These effects are

[1] One notable exception is Alderman et al. (2025), who conducted cooking demonstrations and other activities intended to increase men's engagement, but they did not track mental health or well-being outcomes as a result of the intervention. They find that men's engagement in domestic work increased in the treatment group one year later, but required additional reinforcement to sustain beyond that. They found no effects on men's childcare. They also cannot rule out that the men's engagement effects might instead have been the result of improvements in women's mental health as a result of interventions with women who had mild-to-severe depression at baseline.

concentrated among households with high chore loads at baseline, where on average all men increased their UDCW contributions, whether or not they participated at baseline. In addition, sons in treated households increase time spent on domestic chores while daughters' overall unpaid work remains largely unchanged.

Taken together, the results offer a striking picture of the connection between intra-household allocation of UDCW and mental well-being, and the mental health benefits of an intervention aiming to redistribute the burden. By modestly increasing men's (and to some extent sons') contributions in duration rather than by reducing women's, the intervention generates joint gains in mental well-being within households. That is, the mental health improvements for one partner are not offset by losses for the other even though one partner's UDCW increased and the other's was not reduced. The households that benefit from mental health improvements are not always the same as those where husbands are induced to perform more chores. This brings up important and fundamental questions about the relationship between mental health and workload.

To understand how the intervention improved mental health, we examine six potential mechanisms related to household production, time allocation, effort, relative burden, heterogeneity by baseline loads, and perceptions. For this analysis, we leverage a rich set of data from both spouses, including novel measures of linked time allocation and subjective well-being data from more than 20,000 randomly-timed, high-frequency surveys about experiences in the moment. We find no evidence that the intervention increased total household production of UDCW, altered time spent on income-generating activities by either spouse, or changed time use patterns across other categories. Instead, the intervention reshaped the relative distribution of UDCW within the household and also how the work is experienced by those doing it. Differences in subjective well-being across UDCW tasks indicate that some chores are preferred to others (see Figure 1), consistent with work from across disciplines that the composition of unpaid work matters, not just absolute quantity. While men took on more domestic work, women's *relative* burden declined, the gender balance of UDCW work improved, effort shifted away from women through greater load-sharing, and women became more accurate in perceiving men's contributions. These changes operate differently

depending on baseline household chores performance patterns but consistently point to improved intangible mechanisms like support, balance, and recognition within households as potential drivers of the observed mental health gains.

We contribute to the literature on the socioeconomic determinants of mental health and well-being in several ways. First, we provide novel causal evidence that mental health is driven by UDCW, and that depression prevention does not necessarily require a lightening of UDCW loads. To the contrary, we find that increasing men's UDCW by small amounts of time can actually improve their own mental health. This result points to important intra-household dynamics that are often overlooked and not well understood by economists (Buzard et al., 2025). While total hours performing UDCW are associated with worse mental health outcomes in many studies,[2] our evidence emphasizes the importance of *relative* differences in chores performance between a woman and others around her (Harryson et al., 2012; Glass & Fujimoto, 1994; Boye, 2009). The type of UDCW performed can also matter,[3] since some improvements for women are associated with re-allocating women's least preferred chores to their husbands.

Second, we provide compelling evidence of efficacy for a scalable policy tool for improving well-being by preventing depression symptoms from occurring or worsening. Our intervention leverages simple technology, recruiting men to be proactive instead of labeling them as part of the problem, embracing them as part of a means to improve not only their families' well-being but their own as well. In doing so, we contribute to the literature on behavior change interventions targeting outcomes shaped by strong gender norms.[4] We note especially that the relatively small additional contributions of time by men highlight the importance

[2]See Piovani & Aydiner-Avsar (2021); Boye (2009); Glass & Fujimoto (1994); Xue & McMunn (2021); Väänänen et al. (2004); Hammarberg et al. (2020); Prins et al. (2019); Owoo & Lambon-Quayefio (2021); Molarius & Metsini (2023).

[3]Studies find that different types of unpaid labor have different associations with mental health — performing more household chores is negatively associated with health outcomes for women and men, while performing more childcare is positively associated with mental health outcomes for women and performing more outdoor chores is positively associated with mental health outcomes for men (Ervin et al., 2023; Daminger, 2019; Musick et al., 2016; Sayer, 2005).

[4]A large behavioral literature shows that psychologically informed interventions—such as message framing (Asensio & Delmas, 2016; Madajewicz et al., 2007), commitment devices (Avdeenko et al., 2019; Dupas, 2011; Luoto et al., 2014), recognizing role models (Shah & Ludwig, 2016), and correcting misperceptions (Bursztyn et al., 2020; Dupas, 2011; Fitzsimons et al., 2016; Madajewicz et al., 2007)—can shift behavior in domains such as health, personal finance, and labor supply.

of relative burdens but also reinforce a conceptualization of the intervention as a form of empathy training and relational attunement building, through which men benefit from enhanced empathy, while their wives on the receiving end experience corresponding gains as well (Wiseman, 1996; Banmen & Maki-Banmen, 2014; Bastos et al., 2025). We intentionally designed this phone intervention as a series of one-on-one, personalized conversations led by a well-trained facilitator, which can be a promising mode of delivery for interventions that aim to influence relatively complex behavioral responses. However, we caution that these benefits may not carry over to text-based delivery, a more common BCC delivery mode,[5] since it limits the scope for tailoring content, addressing misperception in real time, or engaging respondents in reflection, and it requires basic literacy and sustained attention.[6]

From a policy perspective, our findings illustrate how norm-sensitive, personalized behavior change communication can generate improvements in mental health while also promoting more equitable divisions of unpaid work, suggesting potential welfare gains for both women and men rather than trade-offs between them. To guide our empirical analysis, we integrate insights from economics, psychology, and sociology in a theory of change explaining how unpaid domestic and care work, gender norms, and household relationships can jointly shape mental well-being (see Section 2.3). Moving beyond standard economic models that focus exclusively on concrete outcomes (e.g., hours worked), this synthesized framework outlines how interventions that change the nature of work can alter not only time allocation but also relational dynamics, perceptions of fairness, and shared responsibility within households. These insights align closely with shared policy and survey methodology priorities by institutions like the World Bank and the International Labour Organization.[7] Our work points

[5]For example, text-based or highly standardized communications can assist in reinforcing health workers' adherence to clinical guidelines (Zurovac et al., 2011), building savings (Rodríguez & Saavedra, 2019; Kerbaj et al., 2017), adopting hygiene behaviors (Kerbaj et al., 2017), or completing prescribed antibiotics courses (Ostervall, 2017). However, these are relatively more simple behaviors, where simple reminders and/or information can enhance adoption of episodic or rule-based behaviors.

[6]Improved delivery of AI-based BCC interventions could achieve similar effects to our phone-based two-way conversations with lower costs if they allow for personalization, feedback, and adaptation to the respondent's circumstances, an area for experimentation we leave to future work.

[7]The World Bank's 2025–2030 Gender Strategy emphasizes shifting social norms, engaging men and boys, and leveraging digital technologies to address persistent gender inequalities (Narayan, 2023). The Living Standards and Measurement Surveys team convened a group of experts in 2025 to discuss the importance and complexity of measuring UDCW and socioeconomic skills like empathy and self-regulation, including a new framework from a World Bank team on the latter (ESTEEM) (Delavallade & Das, 2024). The

to future research opportunities that account for intangible dimensions of work—such as cognitive, relational, and emotional labor (Buzard et al., 2025; Baranov et al., 2020)—in both research and policy design. Emerging digital tools can enable scalable measurement and intervention in these domains, especially in low-resource settings.

We explain our intervention and present a theory of change in Section 2; detail our study setting, experimental design, attrition, and outcome variables in Section 3; describe our empirical strategy in Section 4; report effects on core mental well-being and chores participation outcomes in Section 5; explore pathways of impact drawing on additional data, including unique high-frequency data on time use and subjective well-being, in Section 6; and conclude by pointing to new opportunities for research in Section 7.

2 Context, Intervention Details, and Theory of Change

Our intervention took place in rural Amhara, a region characterized by deeply entrenched gender norms that not only assign almost all UDCW to women but also outline rigid guidelines on spousal relationships (e.g., both men and women commonly hold beliefs that wives and husbands should not call each other by name, greet each other warmly, or walk side-by-side).[8] The region thus regularly features gender-sensitive programs and policies to reduce early marriage, increase educational attainment, and improve women's bargaining power (EDHS, 2016; Asfaw et al., 2025). This setting was therefore well suited for piloting an intervention targeting men to engage more in UDCW and for testing the hypothesis that high UDCW loads cause mental health issues for women in particular.

To provide direct descriptive motivation for our focus on UDCW as a plausible upstream determinant of mental well-being for both spouses in this context, Figure 1 Panel (b) shows baseline distributions from our sample for a standard measure of depressive symptoms (PHQ-9 scores),[9] for primary males (PMs) and primary females (PFs), stratified by their respective

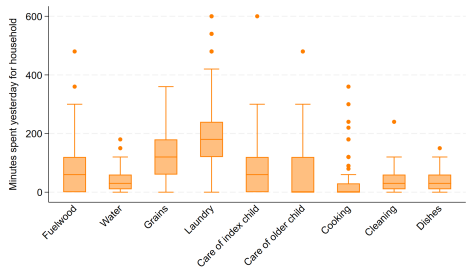
International Labour Organization launched a light time-use module to guide National Statistics Offices in measuring own-household production of goods and services (including care work) that are difficult to observe accurately (Heggeness et al., 2025).

[8]Source: focus group discussions during scoping activities and consultations with NGO staff.

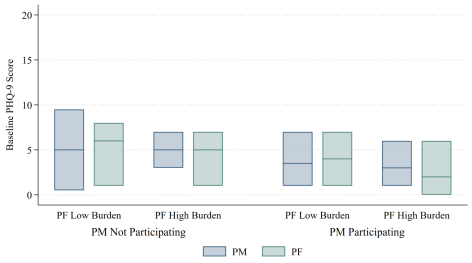
[9]We report more on this and other outcomes later in this article and explain the data completely, but for

Figure 1: Baseline UDCW loads and well-being of PFs and PMs

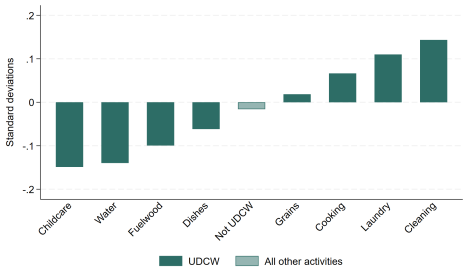
(a) Household UDCW time burden by chore



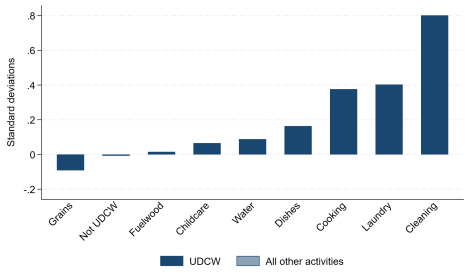
(b) PHQ-9 scores by baseline chores load



(c) PFs' subjective well-being by chore



(d) PMs' subjective well-being by chore



Note: The figures present baseline summaries of household UDCW loads in minutes by chore for the nine we tracked specifically (panel A); the depression scores for PM and PF by their baseline chore loads (panel B), and normalized subjective well-being (SWB) ratings for the PF (panel C) and PM (panel D). In panel A, chores performance minutes and number of chores for the top two figures are from the PF's reports during the baseline phone interview on chores completion, when her PHQ-9 was also completed. Panel B reports inter-quartile ranges of the PMs' and PFs' self-reported PHQ-9 scores at baseline, omitting outside values for simplicity. For more details on these interviews and measures, see Section 3. In Panels C and D, SWB data are from the high-frequency time use surveys collected at baseline and normalized separately by gender. All data are introduced in Section 3.5. The tasks are ordered based on average SWB (from least to most happy) for each gender.

baseline chores loads. Notably, depressive symptoms are better for PFs in households where PMs participate (right half) compared to those where PMs do not (left half). In fact, in households where the PM does not participate, both the median PM and PF could be classified as having mild depression (PHQ-9>4). Men's participation in UDCW seems to be especially associated with better mental health in households where PFs also have an above-median total chores burden (far right).

2.1 The Intervention

Our intervention engaged each father in the treatment group through personalized conversations encouraging him to reflect on the gendered division of unpaid labor within his household and family of origin and to participate (more) in household chores and childcare. During each intervention call with PMs assigned to the treatment group, a well-trained female enumerator followed a carefully designed outline of talking points. The focus, content, and style of the calls are summarized in Table B1, and the full text of enumerators' on-screen prompts appear in Appendix B. We prepared these outlines in consultation with gender experts from CARE Ethiopia and World Vision Ethiopia, with most of the content inspired by a World Vision intervention in which men engaged in group discussions to become allies in empowering women (Alderman et al., 2025). Each call was customized to the PM and his family (e.g., using family members' actual names) to ensure a natural conversation about the respondent's own personal experiences while centering on a clear objective.

We cycled through three script outlines twice. The first asked the PM to consider the burden of time that his wife spends on UDCW compared to him; the second encouraged his involvement by explicitly asking him to reflect on his own upbringing and household of origin, his identity as a spouse and father, and the importance of fathers in children's lives; and the third asked him to reflect on social norms, including his assumptions about the capabilities of men versus women, and any attitudes or misconceptions that might hinder his participation. After repeating these again, the seventh call reviewed all prior discussions.

the purposes of this illustration it is enough to know that lower scores correspond to better mental health (fewer and less severe depressive symptoms).

We emphasize that the intervention was designed to engage men in the behavior change process through self-reflection in a safe and non-judgmental social setting rather than to manipulate men's beliefs about gender norms or insist that they reject these norms (Petit, 2019).[10] Accordingly, our primary hypothesis concerns behavior rather than informedness or beliefs *per se*. To this end, the scripts utilize open-ended questions, short vignettes, and guided discussions to elicit personal reflection on family life, fatherhood, and household work (see Table B1). To preserve conversational flow and confidentiality, enumerators did not record men's responses to questions during these reflective conversations.

To reinforce behavior change and provide light-touch accountability, each call concluded by asking the PM to commit to performing a small household chore or childcare task during the following two weeks, with follow-up in the subsequent call. These commitments were the only data we recorded from the men during intervention calls. Because these data were elicited in a script-specific manner and were not collected from men in the placebo group, they do not permit systematic tracking of the dynamics of men's participation and are therefore not reported in this article.[11]

2.2 The Placebo

Each PM in the control group also received calls on the same schedule but discussed unrelated placebo content—a review of the PM's own food consumption in the previous 24 hours. We selected this content as it would be unlikely to cause the PM to take action to change the allocation of chores across household members or affect his mental well-being. This placebo calls protocol served two important purposes. First, by placing a call (lasting a similar duration of time) to all men regardless of treatment group, we maintained the same level of regular contact with all households, thus preventing differences in attrition,

[10] Members of our research team have a companion article from a similar sample in Amhara dealing with issues of pluralistic ignorance (the idea that men might misperceive the strength of their peers' beliefs about gender norms), so we did not address that issue in this study (McGavock et al., 2020).

[11] We do have some information from these intervention calls that indicates men's willingness to try new things and confirms that many of them were engaged enough to remember the task they committed to from one week to the next. For more details, see Appendix B.

familiarity with our enumeration staff, or building of rapport.[12] Second, our protocol ensured we gave instructions for the same respondent (the PM) to carry the phone during the intervention period for both treatment and control households, thus preventing differences in within-household phone access that could have differentially affected outcomes.

2.3 Theory of Change

Figure 2 describes the complex web of processes within families that could be influenced by our intervention (bottom) when shaping broader outcomes (top) like time use agency, material standards of living, relational attunement, mental health, and child development. We highlight a few key intermediate outcomes linked to these, which we selected to align with principles on measurement described below. To conceptualize the intervention itself, we draw on the COM-B framework—capabilities-opportunities-motivation & behavior—and the "behavior change wheel" (Michie et al., 2011). These frameworks support our emphasis that our intervention is not merely about information—it is a bundled experience that can simultaneously build capability (know-how and confidence to help), reshape motivation (identity, purpose, perceived returns to effort), and create opportunity (promoting empathetic observation, conversation, and coordination within the household).

[12] We explicitly test for and can reject concerns about enumerator effects for our measurement of outcomes and estimated treatment effects. See Appendix D.4.

Figure 2: Unpaid Domestic and Care Work and Mental Health: A Theory of Change



To connect the intervention and the impacts, we primarily draw on two dynamic frameworks from psychology: the transtheoretical model of behavior change progression (especially pre-contemplation and contemplation stages) (Prochaska, 1984) and Satir's emphasis on disruption and integration when moving from an existing status quo to a new one to fulfill deep human needs ("yearnings") for relational change and acceptance (Banmen & Maki-Banmen, 2014). The diagram highlights these dynamics from intervention to impacts: reflection (understanding the status quo) which could happen between calls not just during them; learning about self and others, their preferences, constraints, and beliefs (the economic modeling fundamentals involved in intra-household production and relational attunement)[13]; opportunities for experimentation (early-responding behaviors) with changing time and/or effort involved in UDCW, livelihood activities, and/or interpersonal interactions (actions that convey empathy, listening, responding to bids for connection)[14]; and finally the experience of feedback and reinforcement that can contribute to new opportunities, motivation, and capabilities for behaviors that turn into routines (Petit, 2019; Sirois & Pychyl, 2013).

We emphasize three important aspects of the theory of change. First, it is dynamic. Short-run changes in time and effort (e.g., men trying chores) can generate feedback and reinforcement through lived experience—connection and understanding with others, confidence with tasks, and updated beliefs about the social costs of violating norms (Narayan, 2023), so that early experimentation can evolve into new routines or identity shifts (Bénabou & Henkel, 2025; Swerbenski et al., 2023). Intermediate outcomes may then include shifts in the balance of tasks, changes in depression symptoms, and downstream impacts. Second, heterogeneous impacts are likely based on the status quo at baseline (such as the division of labor and mental health) and the specific sequences of actions and feedback. These can impact not only the ultimate outcomes but also the nature of the process. Finally, diverse disciplines suggest "circular causality" (Banmen & Maki-Banmen, 2014) since mechanisms are "entangled" (Petit, 2019), so we do not treat pathways as mutually exclusive; rather, we take an "abductive" approach to considering the available evidence (Heckman & Singer, 2017).

[13]See for example Bastos et al. (2025), Browning et al. (2020), and Haghighi et al. (2024).

[14]See for example Sinharoy et al. (2023), The Gottman Institute (2018), Vieten et al. (2024), Verhofstadt et al. (2016), and Martingano et al. (2025).

To make progress empirically, we follow several core principles. First, for estimating treatment effects, we prioritized measuring core outcomes —the absolute and relative performance of UDCW tasks and depression scores (see Section 5). Then, in Section 6 we leverage our rich data to test different pathways that are both policy-relevant and measurable in this low-cost, phone-based design. These are: total consumption of household-produced goods and services, time allocation across activities, effort expended in performing UDCW and load-sharing, relative burden within households, heterogeneity by baseline UDCW, and understanding of others' UDCW within the household. Across these analyses, we rule out changes in total household production, income, or aggregate time use as primary mechanisms. Instead, the evidence supports pathways centered on redistribution and recognition—men's increased participation, shifts in relative UDCW burdens away from women, modest reductions in women's effort through load-sharing, and improved understanding of men's contributions. Together, these findings align with a theory of change in which mental health improvements arise from enhanced perceived support, fairness, and shared responsibility rather than from mechanical reductions in the number of tasks they performed or their minutes performing chores (Buzard et al., 2025; Baranov et al., 2020).

Rather than disentangling all of these mechanisms, our explorations provide strong evidence that can serve as a rallying point for future research to explore. We also emphasize limitations in our ability to illuminate the many social constructs that comprise the intervening steps connecting the intervention to core outcomes. While validated instruments exist for many constructs like agency, socio-emotional skills, empathy, marital satisfaction, and norms, they would be time-consuming and possibly error-prone to collect over the phone. Recognizing the tradeoffs between survey length and data quality, for this early work we chose to prioritize outcomes that are relevant, sensitive to change, and feasible to collect at scale by phone to test the core hypothesis about causality in general. These measures are introduced and explained in Section 3.5, and we discuss limitations and promising areas for future work in Section 7.

3 Experimental Design and Sample

Our sample was drawn from ultra-poor households participating in the Ethiopian government's Productive Safety Net Program (PSNP), which aims to permanently "graduate" households from poverty by collaborating with local and international NGOs to provide multiple channels of support.[15] Conditioning on this ensured that we have a relatively homogeneous sample in terms of living standards. Overall, we sampled 192 households across 48 *gotts* (villages) within 16 *kebeles*, and assigned treatment at the household level by stratifying on *gott*, which we refer to as either "village" or "cluster" hereafter. Additional details on the sampling frame of villages and logistics appear in Appendix C. With attrition, which is balanced on treatment (see Section 3.3), our final analysis sample consists of 163 households across 48 clusters.

3.1 Eligibility Criteria

Within selected villages, we randomly sampled households that met the following eligibility criteria—the household had at least one child between 24 and 59 months (ages 2 to 5 years); there was at least one adult male *and* at least one adult female household member identified as being responsible for their care (usually the married parents of the child); the household had *no* children under 24 months, nor any pregnant women; and the household was participating in the PSNP. We refer to the youngest member aged 24 to 59 months as the "index child" and their caretakers as the primary male (PM) respondent and the primary female (PF) respondent. These sampling criteria were developed in order to ensure that sampled households faced a childcare burden (i.e., young children were present) while eliminating overlap between our sample and that of a parallel program evaluation by IFPRI, which sampled PSNP households with pregnant women or children under 24 months. That program (SPIR II implemented by World Vision) rolled out between the in-person baseline and phone-based baseline phases of our study, recruited all PSNP households to join Village

[15] PSNP eligibility is based on a combination of chronic food insecurity and income and asset poverty. Targeting is based on geography, administrative data, community identification, and other criteria (Sabates–Wheeler et al., 2022).

Economic and Savings Association (VESA) groups, and assigned additional treatments at the sub-county (*kebele*) level related to nutrition and targeting outcomes for infants under 24 months or pregnant women. Because treatment status for our study was assigned at the household level and balanced within villages, the existence of this broader intervention poses no threat to our identification strategy.[16]

3.2 Data Collection and Study Timeline

For each participating household, data collection and intervention activities for our study rolled out in four distinct phases, as summarized in Figure A1. First, during the *recruitment phase* (October 2022), survey staff in the field identified eligible households, established their consent to participate in the study, and conducted separate in-person baseline interviews with both the PM and PF covering basic demographic and household information. We provided all study households with a new, basic mobile phone to ensure that prior phone ownership was not a condition for study participation. All subsequent contact with households occurred by calling these phones from our call center in Addis Ababa, which was staffed by a small group of highly trained female enumerators. Additional details on call center staff, training, and logistics such as maintenance calls, reminder protocols, and randomized assignment of households to survey waves and call times appear in Appendix C.

Over a continuous twenty-week period, each household progressed through the next three phases of the study: *baseline*, *intervention*, and *endline*, as described in the following paragraphs. The start of each household's study period was randomly assigned at the *gott* level in order to meet capacity constraints in the call center.

The *baseline phase* of phone data collection lasted four weeks for each household. First, each PM and PF separately completed one phone baseline survey covering mental health outcomes and completion of chores, lasting about 20 minutes. These occurred a few days apart, and the order was randomly assigned at the household level. During each respondent's baseline phone interview, the respondent first reported on completion of chores for the household. For each of nine specific chores, the respondent was asked whether the chore

[16] See Appendix C for more details.

was completed during the previous 24 hours in service of the household. The nine chores we tracked are: collecting fuelwood, fetching water, processing grains, washing clothes, cleaning the house, cooking, washing dishes, providing care to the index child (between 24 and 60 months, see Section 3.1), and providing care to any other child.[17] These chores were selected based on qualitative exploration conducted during the scoping phase, including informal focus group discussions (FGDs), consultations with gender specialists from CARE and World Vision, and discussions with local NGO partners, and to maintain consistency with the chore definitions used in a prior study (McGavock et al., 2020). If a chore was completed, we asked by whom and recorded how much time was spent by each contributor according to the respondent assigned for that call. We asked about chores performance in this manner rather than by going through the household roster person-by-person in order to limit any social desirability or nudging bias. While we collected chores performance data in this manner from both the PF and the PM, our core chores performance results in this article focus on outcomes collected from the PF, as we pre-specified, to limit any bias in response by the PM that might have been induced by his treatment status. We also present results based on PM reports in Section 5.2.

During the baseline phase, we also collected data on the number of depression symptoms the respondent experienced and their frequency during the two weeks prior to the phone interview using the Patient Health Questionnaire-9 (PHQ-9), a valid and reliable diagnostic instrument widely used in clinical settings for screening and long-term tracking of outcomes (Kroenke et al., 2001; Fried et al., 2022; Glischinski et al., 2021). More details on the questions and construction of outcomes appear in Section 3.5.

Separately, in the weeks following the baseline phone interviews, we also conducted two weeks of high-frequency time use (HFTU) interviews by phone, one week with each primary respondent, with order randomized as with the baseline interviews. Each day for seven days, five calls were placed at randomly assigned times within strata covering daylight hours.

[17] Other children are present in 77% of households and are, on average, 9 years old. In 18% of households, there is at least one non-index child under the age of five. By construction, the index child is the youngest child in the household at baseline, with rare exception of any children born after the in-person baseline survey, who are classified as non-index children. This is rare because our eligibility criteria excluded households with pregnant PFs.

During each call, the respondent reported on their own activity at the time, the other primary respondent's activity at the time of the call (to the best of their knowledge), and a few other simple questions. For more details on the data collected for this article, which we use to explore mechanisms in Section 6, see Appendix E and McGavock et al. (2025) (who provide more details on this survey methodology for a similar but separate sample).

Third, the *intervention phase* lasted thirteen weeks, during which the PM was instructed to carry the phone with him. Every two weeks, an enumerator called him for a brief conversation using either the intervention script (see Section 2.1) or questions covering placebo content (see Section 2.2). As mentioned previously, treatment status was assigned at the household level, stratified on cluster and therefore balanced across waves.

The final four weeks of each household's participation was the *endline phase*, during which we repeated the mobile phone data collection from the baseline phase. In endline phone interviews (one each with the PF and the PM), we again collected information about chores performance and depressive symptoms. As during the baseline phase, these occurred on separate occasions a few days apart with an order that was cross-randomized with the order during the baseline phase.[18] We also repeated a week of HFTU surveys for each primary respondent during the endline phase. All study activities were completed by October 2023.

3.3 Attrition

Of the 192 households recruited to join our study, we collected complete baseline and endline data for chores completion and PHQ-9 from the PF respondent in 163 households.[19] Of these households with complete PF responses, we also collected complete baseline and endline data from PMs in 147 households.[20] Most attrition occurred between the initial

[18]That is, there were four possible sequences for the baseline-endline chores interviews: MF-MF, MF-FM, FM-MF, and FM-FM, and these assignments were cross-randomized with the treatment status for the intervention.

[19]For the PF-collected data, 5 households were never reached after in-person recruitment, 7 could not be reached at baseline only, and 15 could not be reached at endline only. In two cases, the incorrect respondent was interviewed (i.e., PM instead of PF).

[20]Of the attrited PMs, the PM could not be reached at all for 3 households, was only interviewed at baseline in 8 households, and was only interviewed at endline in 1 household. In four cases, the incorrect respondent was interviewed.

recruitment phase (in-person surveys during which the phone was provided) and the baseline phone interview, which was completed just prior to the household's intervention phase.

Table A2 reports attrition rates between the in-person recruitment stage and the completion of phone surveys for PFs and also for PMs, with p-values for tests of balanced attrition. While households in the control arm were slightly more likely to miss either the baseline or endline interview than households in the treatment group (17% versus 14% for the PF sample and 26% versus 21% for the PM sub-sample), attrition rates did not differ significantly between treatment groups. That is, after regressing survey retention on assigned treatment status, we fail to reject the null hypothesis that the coefficient equals zero.

For PF-reported outcomes (chores performance), we use observations from the 163 households with complete PF responses to form a unbalanced panel as described in Section 4. For the self-reported mental health outcomes, we form an unbalanced panel using the 163 completed PF observations and 147 completed PM observations. We address additional attrition for the HFTU dataset in Section 6.2 and Appendix E. Additional robustness checks around sample selection are discussed in Appendix D.1.

3.4 Sample Descriptive Statistics and Balance

Table A3 presents the characteristics of the analysis sample of PFs, PMs, and their households. In the next section, we discuss the construction of outcome variables and present results on balance of pre-treatment outcomes. As shown in Table A3, the treatment and control groups are balanced on baseline characteristics, as indicated by the p-values for t-tests of the difference in means between the treatment and control groups.

Our primary male respondents are 34 years old, on average, with limited educational attainment. Less than half completed any formal education, and less than 15% completed secondary school. About half of them were involved in agricultural production as their main activity during the seven days prior to the phone-based baseline survey. The primary female respondents are slightly younger, with an average age of 28 years. The PFs have even lower educational attainment than the PMs, with about one-third of the sample having attended

any formal school and less than 5% completing secondary school. The vast majority of PFs (∼90%) report unpaid household work as their main activity at baseline.

About half of the households have a young female child under the age of five, and around 60% of them have a young male child. Households are also likely to have older children between the ages of 5 and 18, with an average of 0.7 older male children per household and 0.6 older female children per household. There are few adult males or adult females (apart from the PM and PF) present in the households.[21] For over 95% of the households, the PM and PF are spouses, and for 99% of households, they are the parents of the index child. There are few female-headed households in the sample (less than 7%), as eligibility criteria required presence of at least one adult male caretaker. Households possess limited asset stocks, according to approximate values given by the person most knowledgeable during the in-person recruitment phase interview. That said, over 70% of the households already owned a mobile phone at the time of recruitment.

3.5 Outcomes Variables and Pre-Treatment Balance

Our first outcome category relates to mental health, as self-reported at baseline and endline by the PM and the PF during phone interviews. As pre-specified, our core results use the PHQ-9 score, the standard metric for screening and monitoring for depressive symptoms using the clinical diagnostic instrument we implemented.[22] Table 1 reports average PHQ-9 scores for PMs and PFs at baseline by treatment group, and Figure A2 depicts their cumulative densities. The average PHQ-9 score is around 4, corresponding with occasional presence of four symptoms in the past two weeks or more frequent presence of fewer symptoms. Though clinicians sometimes classify a patient's depression severity based on PHQ-9 score, using such categories presents clinical and empirical challenges, and use of the PHQ-9

[21] We note that it is not common for a spousal pair to live with the husband's mother in this setting—only five households (2 treatment and 3 control) had a mother-in-law as a household member—thus heterogeneity analysis on this dimension is not warranted in this setting but would be a relevant and interesting dimension for other regions.

[22] The composite score variable (PHQ-9) ranges from 0 to 27 and is the sum of scores for the nine individual symptoms (e.g., "little interest or pleasure in doing things"). Each symptom's score is based on the respondent's self-reported perception of its frequency during the two week reference period, using a scale of 0 (not at all), 1 (several days), 2 (more than half the days), and 3 (nearly every day).

score as a continuous measure is preferred (Glischinski et al., 2021; Fried et al., 2022).[23]

Table 1: Baseline depression outcomes by person type and treatment arm

	Primary Male	Primary Female
PHQ-9 Score		
Treatment	4.70	4.35
	(4.16)	(3.92)
Control	4.20	3.93
	(3.62)	(3.62)
P-value	0.44	0.48
Number of Symptoms Reported		
Treatment	3.96	3.76
	(2.97)	(3.14)
Control	3.77	3.48
	(2.89)	(2.86)
P-value	0.69	0.53
PHQ-9>4 (Mild+ Severity)		
Treatment	0.45	0.49
Control	0.41	0.41
P-value	0.63	0.24
PHQ-9>6 (Borderline Moderate+ Severity)		
Treatment	0.32	0.34
Control	0.27	0.18
P-value	0.49	0.02**
PHQ-9>9 (Moderate+ Severity)		
Treatment	0.05	0.08
Control	0.07	0.10
P-value	0.66	0.73
Observations	147	163

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The table summarizes PHQ-9 score outcomes for the PM and PF (i.e., the person-type level) by treatment arm. Mean values are reported. The table also reports the p-value from a t-test of the null hypothesis that each pre-treatment outcome does not differ by treatment status. Pre-treatment balance tests for specific symptoms appear in Table A4, with an alternative measure of severity in Table A5. See Footnote 24 regarding the term "borderline moderate."

Acknowledging these caveats, we proceed to categorize the respondents' baseline PHQ-9 scores, symptom counts, and symptom severity for descriptive purposes. As shown in Table 1, half of PMs and PFs have a PHQ-9 score corresponding with mild or worse depression categories, and these are balanced between treatment and control groups. About one third of

[23]The cutoffs are: 0–4 (minimal), 5–9 (mild), 10–14 (moderate), 15–19 (moderately severe), and 20-27 (severe).

the sample has symptoms corresponding with borderline moderate[24] or worse depression, and there is a higher frequency of borderline moderate or worse depression among PFs assigned to the treatment group compared to those assigned to the control group. Less than 10% of the sample reports symptoms corresponding with moderate or worse depression (a PHQ-9 score of 10 or more). Table A4 reports pre-treatment balance on the presence of each of the nine possible symptoms, and Table A5 reports balance on the cumulative count of symptom frequency.[25]

We have several outcomes related to performance of UDCW, for which we use reports from the PF. The first outcome in this category is a binary measure of whether a person of type j performed any of the seven domestic chores or two childcare chores on behalf of the household on the previous day, where the person types are: PM (the primary male), PF (the primary female), sons (any male child aged 5-17, inclusive), daughters (any female child aged 5-17, inclusive), other adult males (any other adult males aged 18 or older), and other adult females (any other adult females aged 18 or older).[26] As reported in Table 2, nearly all PFs perform domestic chores and childcare at baseline (99% and 95%, respectively). By contrast, only about one-third of men participate in domestic work and just over half participate in childcare. There are few adults apart from the PM and the PF in study households (10 have an adult male, 12 have an adult female).

The second and third chores outcomes are also summarized in Table 2 and measure participation on the intensive margin. Using the count of the number of chores covered by a given person type (from zero to nine), PFs perform eight times as many domestic chores as PMs on average. The total minutes spent by a person type shows that PFs spend more than five hours per day (about 330 minutes) performing domestic work and more than three hours performing childcare per day (about 180 minutes) compared to less than two hours of combined UDCW for men (39 minutes for domestic work and about an hour of childcare).

Our fourth and fifth chores outcomes, reported in Table A6, measure *relative* contribu-

[24] We include the categorization "borderline moderate+" (PHQ-9> 6) since the "moderate+" group is small in our sample, but we note this is not a standard threshold.

[25] There is one small exception—62% of PMs in the treatment group report feeling tired at least some of the time, while only 48% of PMs in the control group do.

[26] We exclude performance of chores by any children under the age of 5, which happens extremely rarely.

tions rather than *absolute* contributions. We define, for each person type, a share of total household person-chores[27] and each person-type's share of total household chore minutes,[28] thus easing concerns about recall of duration by ensuring that any mis-estimates of time that are common across all duration reports by a respondent do not affect the outcome. In practice, these yield similar results. All chore-related outcomes are also reported at the level of the specific chore (e.g., fuelwood collection) in Tables A7, A8, and A9.

In general, all outcomes are balanced between treatment and control groups for each person type, with just three exceptions: PMs' minutes for domestic work and childcare as well as PFs' share of minutes on childcare. Given the large number of t-tests this represents, some imbalance would occur by chance even in the absence of imbalance suggesting non-random assignment. We explain our ANCOVA strategy which addresses these issues by controlling for baseline values and selecting covariates in Section 4.

[27] That is, we count the household total of person-chores that were performed yesterday, which sums the number of contributors for each chore over all chores in the category. If, for example, the household's chores performed on the previous day included only fuelwood collection (with two contributors), cooking (one contributor), and grains processing (two contributors), then the household's person-task total for domestic work would equal five. If the PF contributed to all three of these chores, then her share of the person-task total would be 0.6.

[28] Some of these tasks may be completed together and/or in parallel, i.e., simultaneously, which we explore in Section 6.3 using the HFTU data but cannot address with the chores dataset.

Table 2: Baseline chores performance outcomes by person type and treatment arm

	Primary Male	Primary Female	Sons	Daugh -ters	Other Adult Males	Other Adult Females
Domestic Work						
Any Chore						
Treatment	0.29	1.00	0.14	0.42	0.17	0.57
Control	0.40	0.98	0.17	0.30	0.25	0.20
P-value	0.17	0.16	0.76	0.35	0.79	0.23
Number of Chores						
Treatment	0.52	4.30	0.26	0.76	0.17	1.14
	(0.99)	(1.51)	(0.78)	(1.08)	(0.41)	(1.35)
Control	0.71	4.33	0.22	0.73	0.25	0.20
	(1.17)	(1.53)	(0.57)	(1.28)	(0.5)	(0.45)
P-value	0.28	0.92	0.82	0.92	0.79	0.15
Person-chore Minutes						
Treatment	38.83	317.90	15.57	38.08	20.00	53.57
	(81.79)	(178.60)	(57.66)	(81.46)	(48.99)	(54.67)
Control	38.10	338.69	24.15	48.11	15.00	12.00
	(72.74)	(182.76)	(65.38)	(95.56)	(30.00)	(26.83)
P-value	0.95	0.44	0.56	0.68	0.85	0.15
Childcare						
Any chore						
Treatment	0.49	0.96	0.11	0.24	0.00	0.29
Control	0.59	0.94	0.02	0.22	0.00	0.20
P-value	0.25	0.46	0.16	0.83		0.73
Number of Chores						
Treatment	0.60	1.65	0.11	0.26	0.00	0.43
	(0.68)	(0.55)	(0.32)	(0.50)		(0.79)
Control	0.78	1.54	0.02	0.27	0.00	0.20
	(0.75)	(0.62)	(0.16)	(0.56)		(0.45)
P-value	0.17	0.24	0.16	0.96		0.51
Person-chore Minutes						
Treatment	46.54	194.43	6.86	19.47	0.00	34.29
	(73.45)	(130.63)	(19.37)	(37.92)		(58.55)
Control	77.35	177.13	7.32	23.78	0.00	24.00
	(98.51)	(112.18)	(46.85)	(56.98)		(53.67)
P-value	0.02**	0.38	0.96	0.70		0.73
Observations	163	163	76	75	10	12

Note: * $p < 0.10$, ** $p < 0.05$,*** $p < 0.01$. The table summarizes outcomes described in Section 3.5 at the person-type level by treatment arm. Mean values are reported, and, for all continuous variables, standard deviations are reported in parentheses below the mean. The table also reports a t-test of the null hypothesis that the pre-treatment outcome does not differ by treatment status. For each person type, we observe chores outcomes for all 163 households for whom we have complete interviews with PFs.

4 Empirical Estimation

We estimate treatment effects using an Analysis of Covariance (ANCOVA) model, estimating differences in each outcome variable at endline between the treatment and control groups while including the baseline (pre-treatment) value of the outcome variable as a control. Most of our main outcomes, such as depressive symptom score and chores performance, are observed at the person-type level. For these outcomes, because not all person types are present in all households, we structure our dataset as an unbalanced panel of person-type observations.[29]

Thus, for each person type j, we estimate a treatment effect (β_j) as follows:

$$Y_{ij,t=1} = \alpha + \sum_{j=1}^{6} \beta_j (T_i * I_j) + \gamma Y_{ij,t=0} + \mathbf{X}'_{\mathbf{i,t=0}} \lambda + \rho_j + \delta_v + \tau_d + \epsilon_{ij}, \quad (1)$$

where $Y_{ij,t=1}$ is an outcome variable measured at endline ($t = 1$) for person type j in household i. The six person types are defined in Section 3.5 (PM, PF, son, daughter, other adult male, and other adult female). The variable $Y_{ij,t=0}$ is the baseline outcome and T_i is an indicator that equals one if household i is assigned to the treatment group, meaning that the PM received the intervention calls, and zero if the PM received the control calls with placebo content. I_j is an indicator that equals one if the person type is j. The treatment variable is interacted with an indicator for each person type j, and so the coefficient of interest β_j shows the intent-to-treat effect on the outcome for person type j. The vector of household level covariates ($\mathbf{X_{i,t=0}}$) contains additional pre-treatment covariates selected using post double selection (PDS) lasso (Belloni et al., 2014), as further discussed below. We include fixed effects at the village level (δ_v), which is the level at which we stratified treatment assignment, and person type fixed effects (ρ_j). We also control for time of day fixed effects based on the part of the day when the endline interview was completed (τ_d). The error term for person

[29] By using an unbalanced panel, we avoid introducing zeros that indicate non-performance for households without a given person type. For person-type level chore outcomes, the estimation sample includes 499 PF-reported observations (163 PMs, 163 PFs, 76 sons, 75 daughters, 10 adult males, and 12 adult females). For chore-level outcomes, the unit of observation is the person-type–by–chore level (9 chores), equivalent to 4,491 observations. For mental health outcomes, which are self-reported, the sample includes 310 observations (163 PFs and the 147 PMs who also completed the surveys).

type j in household i is denoted by ϵ_{ij}. We compute robust standard errors clustered at the village level to correct for within-village correlations, since that is the level at which the treatment assignments were stratified. We also report randomization-inference (RI) p-values, which exploit the experimental assignment mechanism and offer inference that is valid under the original stratified randomization, independent of asymptotic assumptions.

We adapt this model for chore-specific outcomes Y_{ijk}, such as whether any person type j performed a specific chore k and the share of total household chore minutes contributed by anyone of person type j for chore k. In these cases, we replace the person-type fixed effects with a full set of person-by-chore-type fixed effects.[30]

Given the random assignment of the treatment, any time-invariant and/or baseline characteristics that would predict outcomes should not be correlated with treatment assignment. Table A3 confirms balance of baseline characteristics, while Tables 1, 2, and A4 –A9 report balance for pre-treatment outcomes. We note that, in a few cases, pre-treatment outcomes at the person-type level are correlated with the household-level treatment assignment.

We select the pre-treatment covariates to include in $\mathbf{X_{i,t=0}}$ for each regression using the PDS lasso procedure. This procedure uses a least absolute shrinkage and selection operator (lasso) twice—once to select covariates that best predict the outcome and again to select covariates that best predict the treatment status. Covariates that are selected by both procedures are included in the ANCOVA regression (Belloni et al., 2014). PDS lasso can improve precision of treatment effect estimates when baseline covariates help to predict the outcome variables, and it can reduce bias if the covariates are not balanced between treatment groups. In practice, this procedure selects few covariates for inclusion, and our estimates are quite similar whether or not we include a covariate vector ($\mathbf{X_{i,t=0}}$). We report estimates from regressions both with and without covariates in all regression results tables

[30] Specifically, we estimate:

$$Y_{ijk,t=1} = \alpha + \sum_{j=1}^{J}\sum_{k=1}^{K} \beta_{jk} \left(T_i * I_j * I_k\right) + \gamma Y_{ijk,t=0} + \mathbf{X}'_{i,t=0}\lambda + \rho_{jk} + \delta_v + \tau_d + \epsilon_{ijk},$$

where I_j and I_k are indicators for person type j and chore type k, respectively. The coefficients of interest β_{jk} capture the intent-to-treat effects of the intervention for chore (k) outcomes and member type j. In addition to the other terms defined in equation (1), we replace person-type fixed effects with person-by–chore-type fixed effects (ρ_{jk}).

(see Appendix A).

5 Core Results: Depressive Symptoms and Chores Performance

5.1 Treatment Effects: Depressive Symptoms

Figure 3 Panel (a) reports the effect of the treatment on depressive symptoms for both PMs and PFs based on the specification described in Equation 1, with coefficients and control means in Table A10. We find that the intervention substantially improved (decreased) both PMs' and PFs' total PHQ-9 scores relative to the control group, with a one point improvement for PMs and an improvement for PFs of nearly the same magnitude. This magnitude change in PHQ-9 score corresponds with either one less symptom of a mild nature or a decrease in the severity level (frequency) of one symptom (see Footnote 23). The treatment effect on the PMs' PHQ-9 scores is meaningfully large (22% of the control group mean, with a p-value=0.06). The direction and magnitude of the estimates are consistent with a clinically meaningful improvement, and we reject the null hypothesis in a one-sided test that the intervention had no impact or worsened mental health (p-value=0.03).[31] Similarly, for PFs the PHQ-9 score improves by a meaningful magnitude (21% of the control group mean), and is significant at the 5% level (one-sided test p-value=0.04) using our pre-specified hypothesis test[32] as discussed in Appendix D.5.

Recognizing that it is possible for one spouse's mental health to improve at the expense of the other's, or that one spouse's improved mental health could enhance the other's, we also examine the effect of the intervention on the *sum* of both the PF's and PM's PHQ-9 scores within each household.[33] As depicted in Figure 3 Panel (a), the coefficient on the

[31] For PMs, we pre-specified a two-sided hypothesis since we did not have a strong prior on how the intervention would affect their mental health. The results are statistically significant with a two-sided hypothesis test (rejecting the null hypothesis of no change in PHQ-9 scores) though under the more conservative specification that includes covariates it is noisier and significant at the 10% level.

[32] The two-sided hypothesis test is significant at the 10% level with $p = 0.085$. Details on our pre-registered analysis plan appear in Appendix F.

[33] We note that this is a non-traditional use of these scores intended to assess whether the scores of spouses tend to move jointly or in opposite directions and does not hold clinical significance.

household's sum PHQ-9 score is slightly larger than the sum of the effects for the PF and PM. The household sum PHQ-9 score improves by 2, suggesting a reduction in presence of one mild symptom for two household members or a reduction in presence of two mild symptoms for one household member (p-value<0.01). This result is consistent with additive improvements in mental health across both spouses within the same households rather than mental health improving for one spouse in a household at the other spouse's expense.

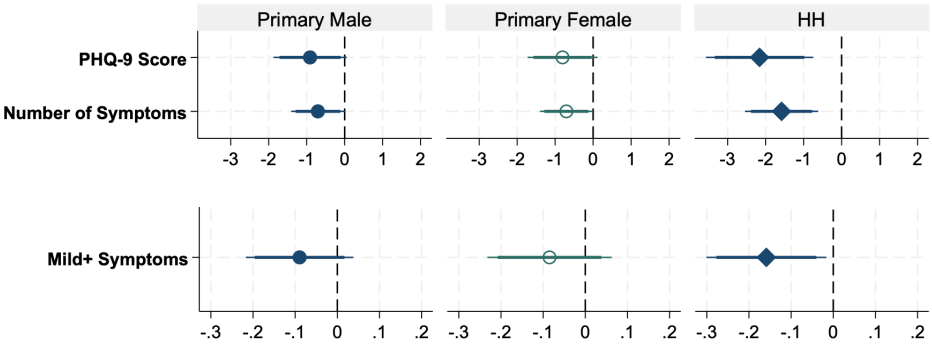
Improvements are also observable using other outcomes constructed from the PHQ-9 questions, such as reducing the number of symptoms present by 0.7 for both PMs and PFs (19% and 20% of the PM and PF control group means, respectively, one-sided p-values<0.05). Because the magnitude of the effect on symptom count is quite similar to the magnitude of the effect on PHQ-9 score for both PMs and PFs, the mental health impacts are consistent with a lower likelihood of a symptom being present in a mild form rather than a reduction in the severity of present symptoms. The intervention also lowered the share of respondents whose PHQ-9 scores correspond with mild or worse depression (PHQ-9> 4) by 9 percentage points for PMs and by 8.5 percentage points for PFs, though only the PM effect is significant (p-value<0.1 with a one-sided test). However, at the household level (far right), the treatment results in a 15 percentage point decline in the likelihood of at least one primary respondent having a score in that range and the effect is statistically significant at the 5% level, pointing to the important role of intra-household dynamics for mental health.[34] Appendix D.5 confirms statistically robust improvements in both PMs' and PFs' mental health, relative to the control group, after implementing adjustments for multiple hypothesis testing.

Figure A3 Panel (a) reports treatment effects on the presence and severity of each of the 9 depressive symptoms that comprise the PHQ-9 score, grouped into cognitive, motivational, emotional, and somatic[35] (see Table A4 for symptom wording). The intervention reduced the probability that both PMs and PFs reported low interest in activities (motivational) and

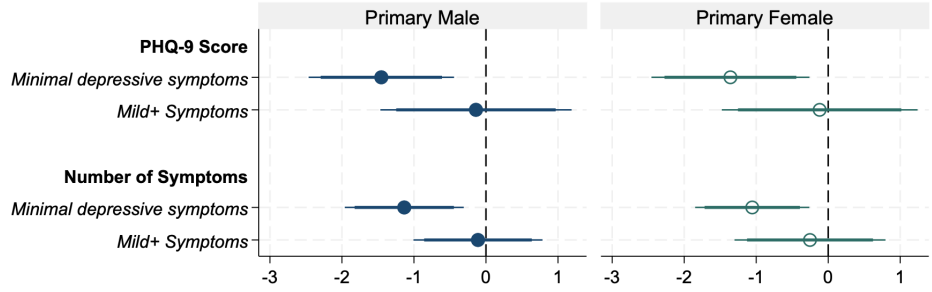
[34]Indeed, if the same estimated treatment effect magnitudes, distribution of the data, and balanced attrition occurred in a quadrupled sample with the same stratified random assignment within villages, the cluster-robust standard errors would reduce by half and both would be statistically significant at the 5% level.

[35]We make these groupings based on the framework described in de Quidt & Haushofer (2016).

Figure 3: Treatment effects on presence of depressive symptoms



(a) Treatment effects on core mental health outcomes



(b) Heterogeneous treatment effects by baseline mental health status

Note: Panel (a) reports treatment effects for three mental health outcomes—PHQ-9 score, number of symptoms present, and a binary indicator of whether the PHQ-9 score corresponds with a mild or worse categorization of depressive symptoms (details in Footnote 23). We also report effects on the sum of the PM and PF outcomes within the same household (for scores or symptom counts) or an indicator for at least one respondent having the mild or worse categorization. The figure depicts the coefficient on treatment interacted with person type for the specification that includes covariates selected (as described in Eq. 1). Panel (b) reports treatment effects for two sub-groups of respondents —those whose PHQ-9 score at baseline corresponds with minimal depressive symptoms (a score 0-4) and those with higher PHQ-9 scores, i.e., the mild or worse group (score $>$ 4). All outcomes are self-reported for the 2-week period preceding the interview. The bars around each coefficient depict its 90% and 95% confidence intervals.

feeling down or depressed (emotional). Both were also less likely to report appetite problems, a somatic symptom, with PMs also less likely to report sleep difficulties, though they were more likely to report psychomotor problems (moving or speaking either too slow or too fast). While PMs do not appear to have experienced improvements in cognitive symptoms as a result of the intervention, one treatment effect for PFs (feeling bad about self) is large and negative, but estimated imprecisely. Panel (b) of Figure A3 shows effects on symptom frequency, which are qualitatively similar and where improvements in PFs' feeling bad about self effects are also statistically significant at the 10% level.

To assess heterogeneity by baseline mental health, we interact the assigned treatment with a binary indicator of whether the respondent had scores indicating mild or worse depression at baseline (PHQ-9>4), the clinical threshold between minimal symptoms and "watchful waiting." The results are reported in Figure 3 Panel (b) and Table A11. For all main mental health outcomes, improvements for both PMs and PFs are concentrated among respondents with minimal depressive symptoms at baseline, with no meaningful or statistically significant effects for those already exhibiting mild or worse depression. Among PMs with minimal baseline symptoms, PHQ-9 scores improve by 1.5 points (p-value<0.01), while they improve by only 0.13 points for PMs with more (or worse) symptoms at baseline. Among PFs with minimal symptoms, scores improve by 1.4 points (p-value<0.05) versus 0.12 points for those with higher PHQ-9 scores. This pattern suggests that the intervention operates primarily through a preventative channel, reducing the onset or worsening of depressive symptoms rather than alleviating existing depression.

5.2 Treatment Effects: Performance of Chores

Figure 4 reports treatment effects on UDCW, distinguishing between domestic work (DW) and childcare (see Table A1 for the full list of chores). For visual display of the results, we focus on the PM, PF, sons, and daughters since less than 10 percent of households have other adult members. Tables A12–A13 report all coefficients as well as control group means.

The results show increased participation by PMs in UDCW. Panel (a) of Figure 4 shows that treated PMs are 16 percentage points more likely to participate in DW (at least one

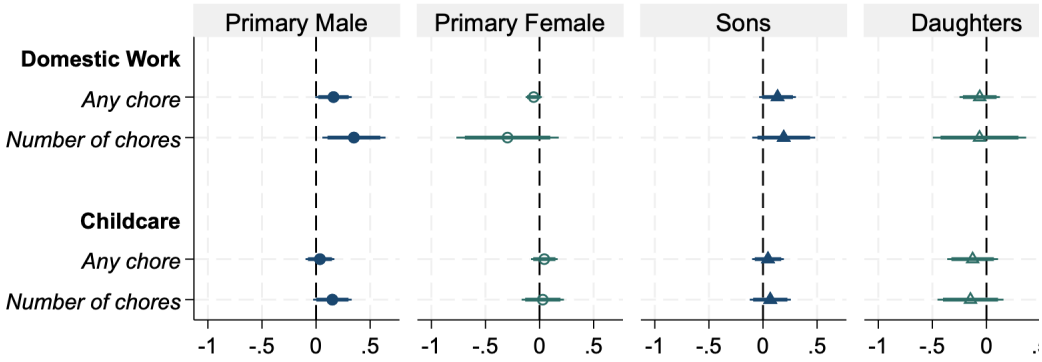
task), a 40% increase relative to the control mean (p-value<0.1). Correspondingly, the average number of DW tasks performed increases, by 0.35 (49% of the control mean, p-value<0.05), and average time spent on DW increases, by 20 minutes (52% of the control group mean, p-value=0.06). However, PMs' contributions do not statistically significantly increase for specific DW tasks except for laundry (p-value<0.1), though all point estimates are positive for 6 of the 7 DW tasks (see Figure A5 and Table A14).[36] This suggests that increases in PMs' contributions are distributed across different tasks rather than concentrated on a specific subset of tasks. Given the highly personalized nature of our intervention, this is consistent and suggests perhaps that the changes will stick as they have become patterns integrated into the household's routine. Treated PMs are no more likely to participate in any childcare (recall the relatively high baseline participation of 60%), but PMs do engage in more childcare tasks (higher number of tasks, by 20% relative to the control mean, p-value<0.1). Total childcare time contributed by PMs does not change on average, however.

As pre-specified, the chores performance results for PMs and all other household member types reported in Figure 4 rely on the PF's reporting of chores performance in the 24 hours prior to the endline interview. Analogous results using PM-reported performance of chores are shown in Figure A4. These PM-reported results suggest different reallocation patterns, so we explore gaps between PMs' and PFs' perceptions of each others' contributions to UDCW in Section 6.6.

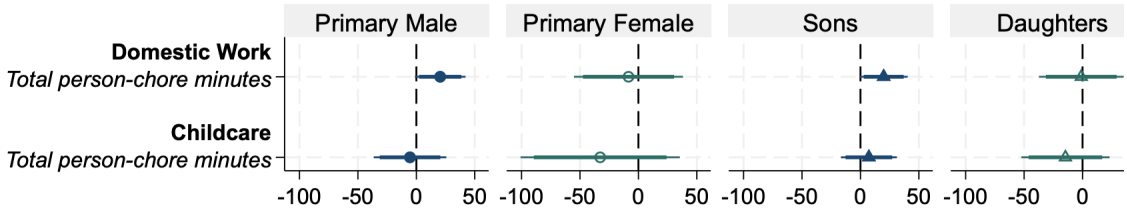
Although the intervention substantially increased PMs' participation in chores, PFs' overall participation declines only modestly (Figure 4 and Tables A12–A13). The probability that PFs perform any DW chore decreases by 5 percentage points (6% of the control mean), and number of DW chores performed by treated PFs declines by 0.3 (7% of the control mean), though neither effect is statistically significant. There is no change in total DW time for treated PFs. PFs do show some changes in probability of performing specific chores, though these also are not statistically significant (see Figure A5 and Table A15). We observe no change in PFs' participation in childcare, number of childcare tasks, or time spent doing childcare as well.

[36]Treated PMs increase their probability of doing laundry by 9 percentage points and spend 11 more minutes on laundry, with both effects exceeding the control-group mean.

Figure 4: Treatment effects on chores performance by person type



(a) Performance of any chore and number of chores performed



(b) Minutes performing chores

Note: The figure reports the impact of the intervention on chores performance outcomes at the person type level. Panel (a) reports the effect on a binary measure of participation in any domestic work chore, a binary measure of participation in any childcare task, the number of domestic work chores, and the number of childcare tasks. Panel (b) reports the effect on total minutes performing domestic work and childcare, respectively. The figure depicts the coefficient on treatment interacted with person-type for the specification that includes covariates selected (as described in Eq. 1). The bars around each coefficient depict its 90% and 95% confidence intervals. These results are also reported in Table A12. This figure relies on chores performance as reported by the PF. PM-reported results are reported in Figure A4.

The intervention also increased sons' contributions to DW, but some of these effects are measured imprecisely perhaps due to limited power since they represent only half of households. Treated sons appear to be more likely to participate in DW by 14 percentage points (82% of the control mean) and perform more DW chores (by 0.19 chores, 86% of the control mean), though these effects are not statistically significant. We estimate a statistically significant increase in time spent by male children on chores on average (20 more minutes, 83% of the baseline control group mean, p-value=0.057). Like PMs, sons' increased DW contributions are spread across many chores (Table A16), with positive effects for 6 of the 7 DW chores, though only fetching water is significant (7 percentage point higher probability where the control mean is 9%, p-value<0.1). The intervention does not change daughters' overall participation in DW or childcare or performance of specific chores (Table A17).

We report on inference for multiple-hypothesis adjustments for a family of related chore outcomes in Appendix D.5, which confirms statistically robust results with increases in PMs' DW contributions and no change in PFs' overall chore participation.

Results for additional adult household members should be interpreted with caution given so few of them are present in the sample. That said, the point estimates suggest that adult males decrease their participation in domestic work, while adult females increase their participation in both domestic work and childcare. Adult females' increases are spread across all DW and childcare tasks, and are statistically significant (p-value<0.05) for 5 of the 7 DW tasks and for care of the index child. These effects are reported in Tables A12, A13, A18 and A19.

6 Engaging Men to Improve Mental Health: Evidence on Causal Pathways

Our results presented thus far show that the BCC intervention caused (1) significant and meaningful improvements in PHQ-9 scores for both spouses; (2) an increase in participation in UDCW tasks by PMs on average; and (3) no significant change in PFs' chores participation, chore counts, or time spent on chores. Taken together, these results raise important

questions about the relationship between well-being and intra-household allocations of time. If women's burdens did not decrease but their mental health improved, this suggests that lightening women's UDCW in itself is not a necessary condition for their own mental health improvement. Indeed, our result for PMs suggests that increasing their own UDCW burden may even be a sufficient condition for improving not only their spouses' mental health but also their own, under certain conditions. These facts invite consideration of how mental health maps to socioeconomic welfare, and provide direction for intra-household models that are equipped to predict the welfare costs of gendered norms on UDCW.

In this section, we evaluate the evidence for possible mechanisms explaining how increasing men's UDCW generated the joint mental health improvements we observed. We take an abductive approach (Heckman & Singer, 2017), leveraging economic intuition and our rich data. The evidence strongly suggests the importance of relatively intangible factors, rather than increased consumption of goods and services.

Classical models would explain the welfare improvements with increased consumption of goods and services and/or leisure (Becker, 1965, 1981). However, we show that the intervention did not increase total household production of unpaid work or change time allocated to economic activities, making this channel unlikely in this setting. Beyond modest shifts in men's time towards domestic work, we otherwise rule out major re-allocation of time across broad activity categories by the PMs and PFs, suggesting changes in income (expenditures) or leisure cannot explain our results.

We then turn to less tangible channels, exploring how sharing the load of specific chores with other family members, having a lower *relative* load if not a lower absolute one, and/or receiving help for especially burdensome loads (given baseline levels) contribute to the mental health improvements. We find evidence suggesting all of these. Finally, we consider the role of perceptions about time allocation, documenting that in some cases primary respondents' ability to report on their spouses' time improved in a manner consistent with increased appreciation for shared burdens. Altogether the results point to multiple, complementary pathways of the intervention's impact on mental health that inspire a closer look at common assumptions in models of intra-household allocation.

6.1 Overall Household Performance of UDCW Tasks

If men's increased participation in household chores is not offset by decreases in others' participation, the household might be able to consume more household-produced goods and services. In a classical framework, this would increase household members' utility, all else equal. To address whether the intervention increased households' overall performance of UDCW, we estimate treatment effects on the number of DW and childcare tasks performed and the household's total time spent on them at the UDCW domain and task levels, using a specification similar to that of Equation 1 estimated at the household rather than person type level.

Overall, we find no evidence that households performed more UDCW tasks or spent more time on them. Figure A7 Panel (a) reports the impact of the treatment on the household level number of DW and childcare tasks performed, and Panel (b) shows the effect on total minutes. All results are null. Although these household-level results may appear to conflict with the person-type results (where PMs and sons increase participation while PFs and daughters do not), we verify that, once effects are appropriately reweighted by each member type's presence, the findings are fully consistent: the intervention redistributes chores within households without increasing total household production.[37]

6.2 Time Allocation Effects

Even without changes in the total performance of UDCW tasks at the household level, the intervention could have affected household members' allocation of time between UDCW and other activities, such as economic activities or leisure, in ways that improve their mental well-being directly (if some ways of spending time promote mental well-being while others degrade it) or indirectly (by altering the household's total income or expenditures). Using the high-frequency time use data (HFTU), further described in Appendix E, we estimate the causal effect of the intervention on PFs' and PMs' time allocation on average. We find that the intervention increased PMs' time spent on DW without inducing large re-allocations

[37]Specifically, the calculation shows that the average time spent on DW increases by 8.5 minutes and is offset by a 9 minute decrease in time on childcare. For details, see Appendix D.6.

across other broad activity categories, and without changing PFs' time allocation.

Figure 5 reports treatment effects on PMs' and PFs' time use shares across eight categories. These results closely mirror those from the chores module. PMs increase their share of time spent on DW by 4 percentage points (about 26 minutes per day, p-value<0.05).[38] We note that this effect size is very similar to the 20-minute effect reported in Section 5.2. PFs show a small, noisy decline in DW time use (also consistent with Section 5.2 findings). We find no meaningful changes in childcare time for either spouse from these reports on primary activities, though we note that the vast majority of childcare is performed as a secondary activity in all time use data, including for this sample. Consistent with household-level analyses reported in Section 6.1, combined PM–PF time shares show no significant change in total domestic work or childcare time.[39]

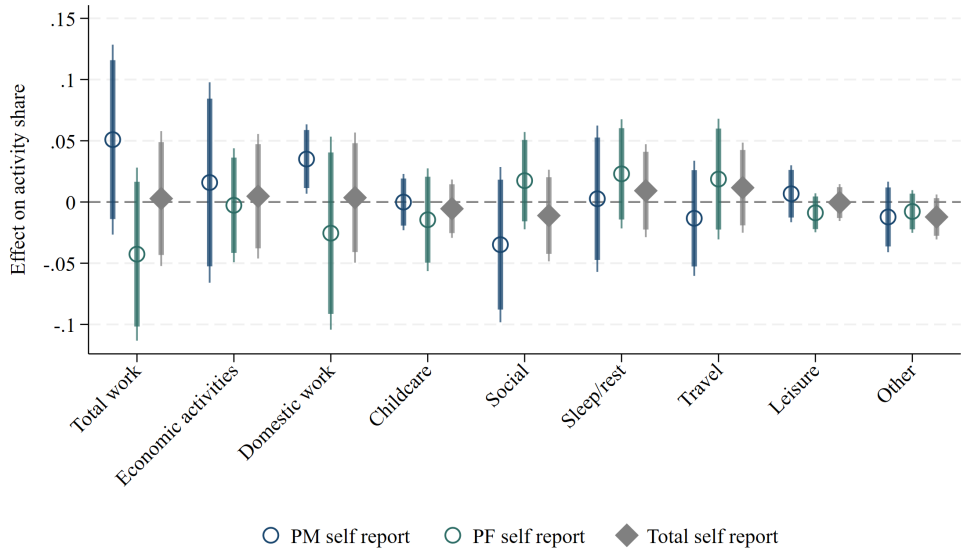
Figure 5 also shows no evidence of substantial changes in the HFTU data for time devoted to economic activities, including self-employment (farm or non-farm enterprise) and wage employment. This rules out income-related substitution as a channel for PM's increased DW. Rather, PMs' additional time spent on DW appears to come from small reductions in other activities—most notably social time—though these effects are not statistically significant. Thus, men's additional DW time is drawn from household-specific changes to other activities that are diluted on average. We also find no significant changes in total work time (which combines all economic activities with DW tasks and childcare) for PMs, PFs, or households overall. PM's total work activity share increases by about 5 percentage points, or 33 minutes per day, though this effect is not statistically significant. Aside from the increase in PMs' domestic work time, there is no evidence of a systemic reallocation across broad activity categories.

Leveraging the HFTU data linking reports of subjective well-being (SWB) to specific activities, we show that the nature of UDCW tasks matters to PFs in particular, providing suggestive evidence that patterns of chores reallocation following treatment may contribute

[38]Given that the HFTU data were collected over an 11-hour period, we convert an activity share into minutes by multiplying it by 660 minutes (the total period of time subject to HFTU sampling).

[39]Though we omit a full discussion of secondary activities here to focus on broad allocation patterns, we estimate treatment effects on the likelihood that respondents report performing DW while in the presence of others and find no impact (see Figure A9).

Figure 5: Treatment effects on activity shares using high frequency time use data



Note: The figure reports the effect of the intervention on the HFTU outcomes: PM's activity shares, PF's activity shares, and "total" activity shares, which combine the PM's and PF's self-reported activity shares. The "total work" category combines economic activities with domestic work and childcare. Otherwise, the remaining 7 categories (economic activities, domestic work, childcare, social, sleep/rest, travel, leisure, and other are mutually exclusive and sum to 1 for each respondent). Treatment effects are estimated using data at the call level so that time of day and household fixed effects are controlled for. Error bars around each coefficient depict their 90% and 95% confidence intervals. The HFTU data and estimation specification are further described in Section 6.2 and Appendix E.

to mental health improvements. While Figure A5 shows no statistically significant impacts on the reallocation of specific chores, the directions of these changes align with baseline rankings of UDCW tasks shown in Figure 1 Panels (c) and (d).[40] At baseline, PFs reported relatively higher SWB while performing cleaning, laundry, cooking, and grain processing compared to non-UDCW tasks, while their SWB was substantially lower when performing childcare, water fetching, fuelwood collection, and dishwashing. Notably, point estimates of treatment effects are negative and sizable for several of these, especially fuelwood collection and dishwashing. On the other hand, PMs reported relatively high SWB for all chores except grain processing, and the point estimates in Figure A5 for PMs' participation in chores are positive and largest for tasks that PFs disliked the most (fuelwood collection, water fetching, and dishwashing). These shifts suggest that comparative differences in experienced well-being may drive the mental health effects in response to the intervention, especially for PFs who do not, on average, change their time spent doing DW.

6.3 Effort effects

Results in Section 6.1 show no change in total household time spent on UDCW production or in overall performance of chores. Mental health improvements could nevertheless arise if chores are performed more efficiently or with different effort levels, or if household members specialize in tasks where they are more productive.[41] Because we observe chore participation and time but not output quality or effort, we cannot directly test these mechanisms. We therefore proceed in two steps. First, using the chores performance data at baseline and endline and some basic assumptions, we perform a back-of-the-envelope (BOE) calculation to

[40] Detailed descriptions of the subjective well-being (SWB) measures collected during the high-frequency time use (HFTU) surveys, including question wording, normalization, and baseline SWB rankings across activities, are provided in Appendix E.

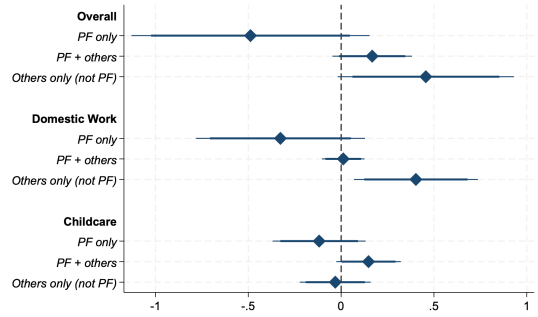
[41] Effort is notoriously difficult to observe in real outcomes outside of lab settings, and one's perceptions of effort (and even time spent) may be measured with substantial error (Charness et al., 2018; Agness et al., 2025). We were also logistically constrained on time in our phone surveys and on resources for field visits, so we did not collect full information on production quantities/quality, nor did we collect simultaneous activity tracker data which might tell us about effort for physically demanding tasks. Moreover, the nature of unpaid domestic and care work is that a lot of labor is performed that cannot be measured based on discrete outcomes from self reports—even self-reported data on the nature of activities with children, for example, does not tell us about the quality of engagement between parents and children. Thus we are limited in our ability to speak to quality and effort and we note this as an area for important future innovations in research and survey methods.

bound the possible effort reduction for PFs afforded by men's increased chores participation. Details on this calculation appear in Appendix G. For this, we ask what is the maximum level of women's effort reduction $e_r \in [0, 1)$ allowed by the increase in men's contribution to household chores. For this calculation, we assume that women exert maximum effort at baseline ($e = 1$). Under different scenarios related to men's effort, we find that the reduction in PF's effort can be no more than about 12.5%.

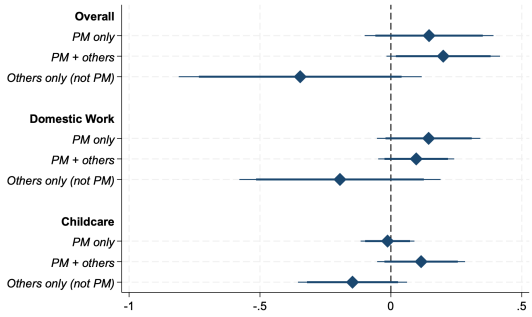
Second, we use the chores data to examine two indirect indicators of effort reduction—greater load-sharing across household members and changes in the simultaneous co-performance of chores. Defining load-sharing as multiple household members contributing to the same chore, we then classify chores performance from each primary respondent's perspective as sole load-bearing, shared, or fully relieved (others perform it). The results, shown in Figure 6, indicate decreased effort for PFs and increased effort for PMs. PFs in treated households are less likely to solely bear domestic work (DW) chores and benefit from greater load-sharing in childcare. In particular, PFs are significantly less likely to be the sole performer of DW chores (by more than 0.4 chores per day, p-value$<$0.05), consistent with a reduction in effort (whether cognitive or physical) even in the absence of large changes in total chores performed. Correspondingly, PMs increase effort, with higher sole load-bearing of domestic work and greater shared participation in both domestic work and childcare, and they are less likely to be fully relieved from chores. Turning from respondent-specific measures to household-level patterns, we next examine whether the intervention increased load-sharing more broadly within households. Figure A8 shows that treated households perform about 0.25 more chores per day with contributions from multiple household members (p-value$<$ 0.05), with the increase driven primarily by childcare rather than domestic work.

While the chores data indicate greater load-sharing, they cannot distinguish between simultaneous co-performance and contributions at separate times, which can affect the mental (or emotional) load of performing duties (Buzard et al., 2025). Using the HFTU data to examine companionship while performing chores, we find no evidence of increased co-performance, more time spent with others, or more spousal companionship during UDCW

Figure 6: Treatment effects on load-sharing of chores within households



(a) Load-sharing from the PF's perspective



(b) Load-sharing from the PM's perspective

Notes: The figures report the impact of the intervention on chore allocation patterns. Panel (a) and Panel (b) shows allocation categories defined from the PF and PM perspective, respectively. The figure depicts the coefficient on treatment at the household level for the specification that includes covariates selected (as described in Eq. 1). The bars around each coefficient depict its 90% and 95% confidence intervals. This figure relies on chores performance as reported by the PF.

activities (Figure A9).[42] These results suggest that mental health improvements are not driven by increased co-presence during chores, motivating our exploration of other dimensions of relational attunement in Section 6.6.

6.4 Relative UDCW burden effects

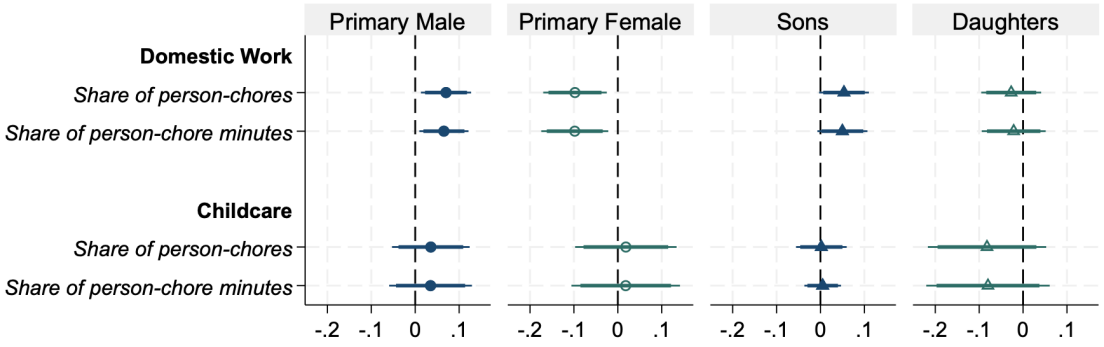
We now examine how the intervention affected *relative* contributions to UDCW, complementing our earlier focus on *absolute* measures. Prior work shows that mental health is often linked to relative intra-spousal UDCW burdens even when absolute burdens are unchanged (Harryson et al., 2012; Glass & Fujimoto, 1994), consistent with broader evidence on the importance of relative comparisons for perceived well-being (de Quidt & Haushofer, 2016). We find that the intervention redistributed relative UDCW burdens away from PFs and toward PMs (and sons), with no change in daughters' shares. Combined with the absence of changes in total household chores, these results are consistent with PFs perceiving increased support with the household's UDCW load.

We first examine treatment effects on each person type's *share* of the household's total chore burden, using the same specification as in Section 5.2. As shown in Figure 7 and

[42]Note that the PM's results are substantially more precise due to a low overall share in UDCW activities.

Table A20, the intervention reduces PFs' share of domestic work (DW) chores and minutes by 10 percentage points (12% of the control mean, p-values<0.01) and increases PMs' share of chores and share of chore minutes by 7 and 6.5 percentage points, respectively (over 60% of each control mean, p-values<0.05). Sons' shares also rise by 5 percentage points (more than the control mean, p-values<0.1), while daughters' shares are unchanged. We find no systematic changes in relative childcare burdens.

Figure 7: Treatment effects on relative chores burden by person type

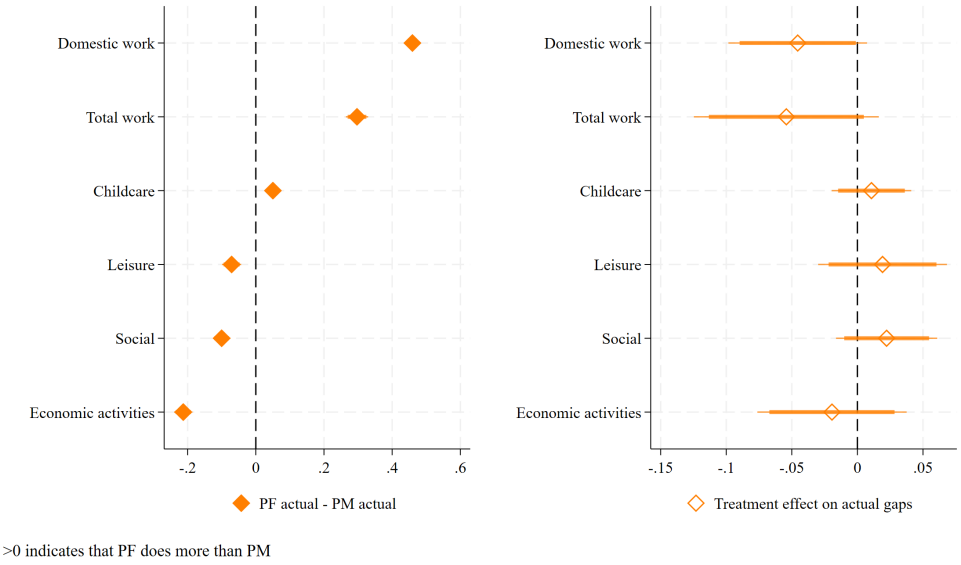


Note: The figure reports the impact of the intervention on relative chores burden outcomes at the person type level. The top portion reports the effect on each person type's share in the household's total number of DW chores performed and the share in the household's total DW chore minutes. The bottom portion reports the effect on on each person type's share in the household's total number of childcare tasks performed and the share in the household's total childcare minutes. The figure depicts the coefficient on treatment interacted with person-type for the specification that includes covariates selected (as described in Eq. 1). The bars around each coefficient depict its 90% and 95% confidence intervals.

Figure 8 presents baseline gaps between PFs' and PMs' activity shares in the HFTU data and the treatment effects on these gaps. At baseline, PFs spent substantially more time on domestic work (DW) than PMs—by about 300 minutes per day—which is only partly offset by men's greater time in economic activities. On balance, PFs' total work burden is 30 percentage points higher than PMs. PMs also devoted more time to leisure and social activities. Treatment effects estimated using an ANCOVA specification indicate modest narrowing of the DW and total work gaps (by about 23 minutes for DW), though these changes are not statistically significant. By contrast, leisure and social gaps remain largely unchanged, suggesting that men continue to retain relatively more time in these activities after the intervention.

Using the PF's reports about chores completion, we next classify households into five

Figure 8: Gender gaps in self-reported activity shares, baseline averages and treatment effects

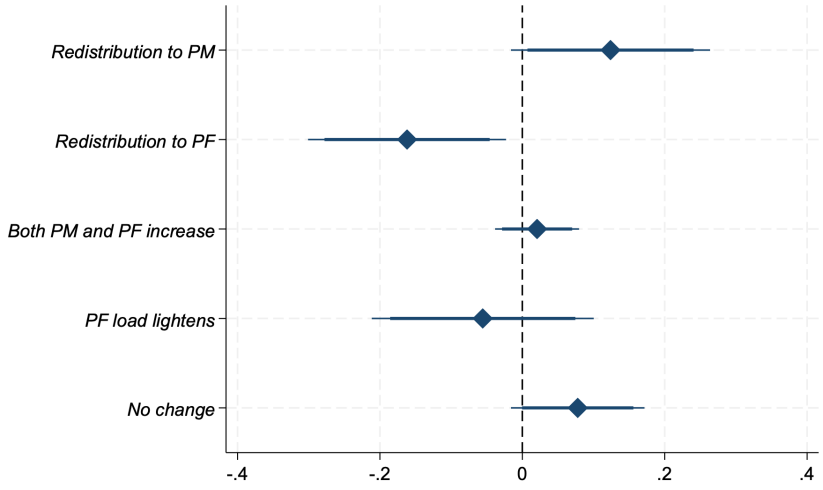


Note: The figure shows baseline gender gaps for HFTU activity shares across categories based on self reports of the PF and the PM (left sub-figure, > 0 if PF>PM) and the treatment effects on these gender gaps (right sub-figure). The bars around each coefficient depict its 90% and 95% confidence intervals. The treatment effects are estimated using a household level version of Eq. 1. Data are from the HFTU data for the sample of households retained using the 40% compliance threshold. For more details, see Appendix E.

mutually exclusive redistribution patterns between baseline and endline and estimate treatment effects on the probability of each pattern (see Appendix D.6 for the definitions of these patterns). As shown in Figure 9, the intervention significantly lowered the probability that chores were redistributed *towards* PFs (by 18 percentage points, p-value<0.05) and increased the probability that chores were redistributed towards the PM (by 12 percentage points, p-value<0.1). Together, this suggests that the intervention effectively reallocated UDCW away from women and towards men in some households.

To see how these changes map to depression outcomes, Figure A10 summarizes the change in each household's total PHQ-9 score from baseline to endline by treatment status, by their redistribution pattern (see Figure 9). There were substantial improvements in the PHQ-9 score between baseline and endline for households in many of these categories, including some whose probability of occurring was not significantly affected by the intervention. In particular, PHQ-9 scores improve substantially in households with mutual increases in chores and in households where PFs' loads lightened, indicating that mental health gains are not

Figure 9: Treatment effects on chores redistribution



Note: The figure reports the impact of the intervention on whether each household's chores redistribution pattern fell into five mutually exclusive categories, with the category definitions described in Section 6.4. The household redistribution outcomes are based on chores completion on the day before the interview at baseline and endline as reported by the PF. The bars around each coefficient depict its 90% and 95% confidence intervals. The treatment effects are estimated at the household level and include baseline outcomes and baseline covariates selected using PDS lasso.

confined to a single redistribution pathway but arise across a range of chore redistribution patterns.

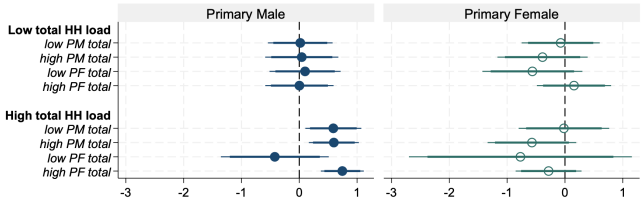
6.5 Heterogeneity by baseline chores participation

We examine heterogeneity in mental health effects by baseline chore burdens at both the household and individual levels to clarify how the intervention operates through UDCW loads. We classify households as having a high or low burden of chores at baseline using the median of seven chores. Further, we stratify by whether the PM performed any chore (baseline participation), yielding four groups that differ in initial chores allocation. We conduct an analogous split for PFs using the PF median baseline burden of six chores. For each subgroup, we estimate treatment effects on chore performance, relative burdens, and mental health.

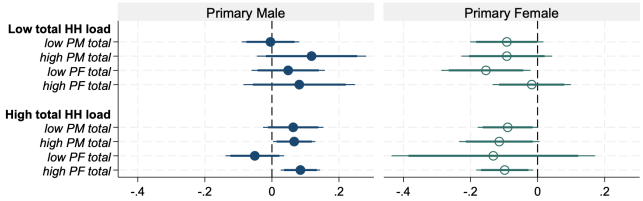
As reported in Figure 10 Panel (a), in high-burden households, PMs respond to the treatment by increasing the total number of chores they perform, and this increase occurs

both for men who did not do chores at baseline and those who did. In high-burden households where PMs performed any chores at baseline, PFs in treatment households perform 0.6 fewer chores, suggesting some displacement of PFs' burden, though this effect is not statistically significant. Turning to heterogeneity by the PF's baseline load, for high-burden households, PMs significantly increase the number of chores performed in households where the PF's baseline burden is also large (by 0.7 chores) but not in households where the PF's burden is lower than the median.

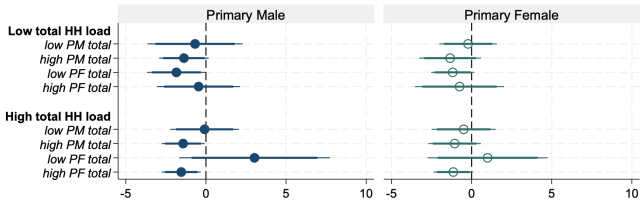
Figure 10: Treatment effects on chores performance, relative burden, and PHQ-9 score based on patterns of chores performance at baseline.



(a) Total number of DW chores performed by respondent



(b) Respondent's share of household total DW chore minutes



(c) Respondent's PHQ-9 score

Notes: The figures report treatment effects on three outcomes for the PM and PF—total number of domestic work chores performed, the respondents' share of total domestic work chore minutes, and the respondent's PHQ-9 score—with each outcome reported for four subgroups of households. The subgroups are defined by the interaction of total household chore burden at baseline (high versus low) and the baseline chore burden of the primary male (PM) or primary female (PF). A high total household chore burden is defined as the household completing seven or more chores on the day prior to the baseline interview. A high individual chore burden at baseline is defined for the PM his completing at least one chore in the previous day and for the PF as her completing six or more chores in the previous day. Bars depict 90 percent and 95 percent confidence intervals.

Figure 10 Panel (b) shows heterogeneous effects on the relative distribution, using each

spouse's share of household DW minutes. In high-burden households, particularly those where PMs already participated or PFs had high baseline loads, the intervention shifts about 10 percentage points of the DW time share from PFs to PMs. In high-burden households where PMs did not participate at baseline, PFs' share also falls (p-value<0.1), though PMs' share does not rise significantly. Effects are noisier in high-burden households with low baseline PF participation. In low-burden households, PMs' shares do not change, but PFs' shares decline in subgroups where PMs did not participate at baseline or PFs had low initial burdens. In these cases, the reallocation appears to reflect increased contributions from other household members rather than greater PM participation.

Figure 10 Panel (c) shows heterogeneous treatment effects on PHQ-9 scores. Across all subgroups, effects for PMs and PFs move in the same direction, reinforcing complementarities in mental health improvements (as discussed in Section 5.1). In high-burden households, PHQ-9 scores improve significantly for PMs—especially when they participated at baseline or when PFs had high initial loads—while PFs' scores also improve, though less precisely. In low-burden households, mental health improves for both spouses when PFs' baseline participation is low (p-value<0.1); for PFs, these gains coincide with significant reductions in their DW time share, consistent with reallocation of burden to other household members.

Taken together, the results suggest that baseline chore burdens shape the channels through which the intervention operates. In highly burdened households, particularly those where PFs are heavily loaded and PMs are already engaged, the intervention shifts the relative work burden toward PMs and away from PFs, even without large reductions in PFs' absolute chore performance. In low-burden households, mental health improvements emerge despite little change in total chore participation, consistent with alternative channels such as increased contributions from other household members, improved organization of tasks, or reductions in PFs' effort and feeling less alone rather than a direct reallocation of chores from PFs to PMs.

6.6 Gaps in perceptions of time allocation

The mental health improvements observed for both PFs and PMs—despite no absolute reduction in PFs' chores and increased chores for PMs—may also reflect improved feelings or perceptions of how others are helping. Using HFTU data that capture both respondents' own current activities and their beliefs about their spouses' current activities, we examine whether the intervention affected within-household perception gaps between respondents and their spouses.

We find large baseline misperceptions about partners' domestic work, and show that the intervention closes these gaps: PFs in treated households more accurately perceive PMs' contributions to domestic work, pointing to improved recognition and perceived support as an additional pathway.

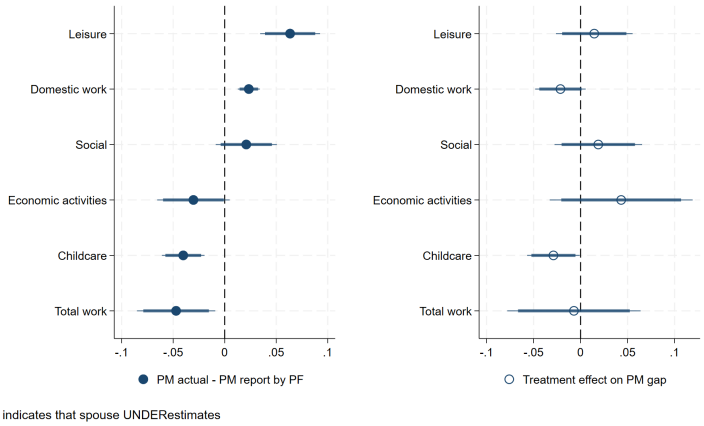
Panel (a) of Figure 11 shows PFs have substantial baseline misperceptions of PMs' time use. PFs underestimate the spouses' leisure, domestic work, and social time, and overestimate their time spent on economic activities, childcare, and total work. The treatment closes the perception gap for domestic work, with the estimated effect fully offsetting the baseline gap, implying that PFs in treated households accurately perceive PMs' DW contributions at endline (p-value<0.1). There is also a noisy indication that treated PFs reduce their overestimation of PMs' economic activity. By contrast, PFs continue to underestimate PMs' leisure and social time. The apparent increase in childcare overestimation likely reflects limits of using only primary activities from the HFTU data.[43]

Panel (b) of Figure 11 shows that PMs also have substantial baseline misperceptions of PFs' time use. PMs also underestimate their wives' time spent on DW, and treated PMs lower this misperception gap (though the effect is very noisy and not statistically significant). PMs also underestimate their wives' time spent on leisure, social, and economic activities and over-estimate their wives' time spent on childcare. Treated PMs also lower their underestimation of their wives' leisure time while further exacerbating their underestimation of their wives' time spent on economic and social activities (though none of these effects is

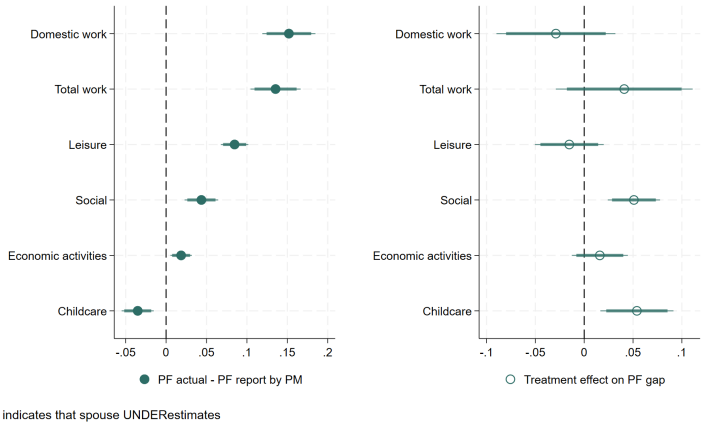
[43]While we did collect secondary activity data, we have focused our analysis on primary activities here for the purpose of exploring mechanisms that could explain the welfare impacts.

statistically significant). Treated PMs do significantly lower their overestimation of the wives' time spent on childcare. Together, these results suggest that the intervention improved PFs' recognition of PMs' DW contributions, as well as the reverse (though noisier), pointing to improved recognition and perceived support as an additional pathway.

Figure 11: Perception gaps between self-and proxy-reported activity shares



(a) Perception gap for PM activities (Self - Proxy report)



(b) Perception gap for PF activities (Self - Proxy report)

Note: Panels (a) and (b) show average perception gaps at baseline (left) and treatment effects on these gaps (right) for the PM and PF, respectively. Perception gaps are the difference activity shares based on self- and a proxy-reports, i.e., a gap>0 if the respondent reports performing an activity more often than their spouse reports for them. The treatment effects are estimated using a household level version of Eq. 1. Data are from the HFTU data for the sample of households retained using the 40% compliance threshold. For more details, see Appendix E.

We note one important limitation of the data from the call center that could affect perceptions as well as estimates of treatment effects on activity categories. Our HFTU method captured time allocation (self- or proxy-reported) only during the call center's operating

hours. That is, some workload changes could occur outside of the call center's operation hours but would not be captured by our surveys. However, we believe this limitation to be minor: the HFTU sampling period covered the entire relevant time period for a majority of households because it corresponds to the year-round daylight hours and most households lack electricity.

7 Discussion

Depression is a major public health challenge in Sub-Saharan Africa, where prevalence is high, access to mental health care remains extremely limited, and large treatment gaps make low-cost, scalable approaches to prevention and symptom reduction especially important (Fan et al., 2025). Interpersonal conflict and social isolation—both closely tied to household dynamics (The Gottman Institute, 2018)—have been highlighted as central drivers of mild depression in low and middle income countries (World Health Organization & Columbia University, 2016). Because depressive symptoms may exacerbate conflict over household labor, while unequal work burdens may themselves contribute to psychological distress, it has been difficult to disentangle the relationship between mental health and unpaid domestic and care work (UDCW). Our study provides new evidence establishing a causal link between UDCW and mental health.

By experimentally varying exposure to a behavior change intervention that directly induced more UDCW performance by men, we provide causal evidence that reallocating unpaid work within households improves mental health outcomes for both spouses. This study adds novel experimental evidence to a literature that is largely correlational or reliant on broad shocks—such as COVID-related school closures—that simultaneously affected employment, income, and social isolation (Piovani & Aydiner-Avsar, 2021; Boye, 2009; Glass & Fujimoto, 1994; Xue & McMunn, 2021; Väänänen et al., 2004; Hammarberg et al., 2020; Prins et al., 2019). We find that our intervention reduces the severity of depressive symptoms for both men and women, with effects primarily operating through prevention of symptom onset or worsening rather than through alleviation of existing symptoms. These improvements are

additive across spouses, suggesting that rebalancing unpaid work can generate shared welfare gains within households instead of redistributing well-being from one partner to another in a zero-sum dynamic.

By demonstrating that a personalized, interactive behavior change communication (BCC) intervention can meaningfully increase men's participation in unpaid work in a setting characterized by especially strong and restrictive gender norms, we also contribute to the literature on gender norms and behavior change (Blackden & Wodon, 2006; Jayachandran, 2015; Dinkelman & Ngai, 2022; Bursztyn et al., 2020; Borrell-Porta et al., 2019; Alderman et al., 2025; Nguyen & Tarp, 2022; Dhar et al., 2022). Delivered through repeated, individualized phone conversations, our intervention encouraged men to reflect on their own household roles and their spouses' contributions. The intervention substantially increased fathers' participation in domestic work, shifted household chore burdens toward greater equity, and also increased sons' time spent on domestic chores, without increasing daughters' UDCW burden and without decreasing women's UDCW burden. The increase in sons' time contributions suggests that such interventions could accelerate the softening of gender norms among younger cohorts, perhaps through a role modeling effect if sons mimic their fathers' increased engagement or through direct communication that shapes beliefs around acceptable behavior and norms (Jayachandran, 2015; Dhar et al., 2019; Bandura & Walters, 1977; Trommsdorff, 2009; Perales et al., 2023). Our findings also show that engaging men directly, rather than focusing exclusively on women's economic participation or on abstract attitudes, offers a promising approach to re-balance entrenched inequities in household labor distributions (Dinkelman & Ngai, 2022; Evans & Jakiela, 2024; Goldin, 2014). More broadly, our results point to important questions about why social norms persist if they lead to sub-optimal allocations, an exciting area of ongoing work in economics (Bicchieri, 2016). That is, if mental health benefits accrue to both spouses in a household apparently without reducing consumption, our results seem to imply Pareto improvements that households are otherwise unable to make given prevailing norms.

The mechanisms underlying the observed mental health effects are likely multifaceted. The evidence shows a strong shift in the relative UDCW burden away from PFs and towards

PMs. We also find suggestive evidence that the intervention improved spouses' perceptions of each other's contributions, which may have reduced misunderstanding and conflict (Bastos et al., 2025). More broadly, increased participation by men may have fostered a sense of shared responsibility and cooperation within households, improving mental well-being for both partners even as men took on more work and perhaps relieving the invisible mental or emotional load of UDCW (Buzard et al., 2025).

While our exploration of the rich data on matched spouses provides fodder for further inquiry, we acknowledge that without longitudinal data on the dynamics of behavior change during the course of the intervention, we are limited in our ability to parse the channels of behavior change. In particular, we cannot quantify the relative importance of—or complementarities in—capabilities, opportunities, and motivation in the COM-B framework for this setting (Michie et al., 2011). Moreover, the mechanisms supported by our results suggest complicated inter-personal processes of social cognition (e.g., communication patterns, updating beliefs about preferences of self and others), which we did not prioritize for measurement in this early causal work. Thus we highlight this as an especially exciting avenue for future work in economic measurement, modeling, and empirical investigation. Indeed, improving our understanding of the intangible aspects of work will be important as artificial intelligence renders some cognitive skills more substitutable (Edin et al., 2022; Drozd & Tavares, 2025) and/or technology aids skill-building in adults' socio-emotional ("non-cognitive") skills like empathy, regulation, and executive functioning (Bloomfield et al., 2025; Cunha et al., 2023). The role of these in caregiving arrangements of all kinds[44] and the capacity for growth in these skills over the life cycle are not yet incorporated well by work in economics but warrant further inquiry.

We note a few additional limitations pointing to directions for future research. We do not directly measure changes in beliefs, attitudes, or self-perceived relationship quality, but these may be important for next steps in unpacking mechanisms for changes in UDCW to have causal impacts on mental health. It would be especially interesting to understand

[44]Such as between romantic partners, between home visitors and new parents, among contributors to child and elder care within and across households, between service professionals and those in their charge, among colleagues.

how these changes might be mediated by or could lead to broader changes in social norms among peer networks. We also do not measure effects on anxiety, which is distinct from, though sometimes correlated with, depression. Mental health measurement in economics remains relatively coarse, and further work establishing causal links to intra-household behavior would be valuable (de Quidt & Haushofer, 2016). Our sample is also specific—poor households in rural Ethiopia eligible for a safety net program—and some estimates are noisy because we lacked the resources for a larger sample. Future research would be helpful for revealing whether sustained increases in men's participation eventually translate into reductions in women's total unpaid workloads and/or transitions to participation in "economic activities" that enhance livelihoods. In contexts where fertility rates are quite low, fertility preferences and outcomes may also be affected by changes in the allocation of UDCW in families (Doepke & Kindermann, 2019). Finally, we did not examine children's outcomes in this study but future work should. While we found no impact of the intervention on the involvement of fathers in childcare time, the mental health impacts of the intervention could affect children's development, which could affect their flourishing over the entire life-cycle (Heckman et al., 2026).

We also note that understanding the UDCW burden's impact on women's economic participation is an ongoing area of work. In particular, given that many women in the developing world are engaged in home-based livelihoods activities that go unobserved by traditional enterprise surveys (Hardy et al., 2025), it is important to understand how the productivity of these activities depends on the UDCW burden and gains from its relief. For example, increasing contributions to childcare provision by other household members or by external programs might substantially enhance productivity. Alternatively, opportunities for women to work part-time or from home can allow them to work in ways that enhance mental well-being as well (Carvalho et al., 2025). Better understanding the dynamics of UDCW and labor supply in such settings depends also on high-quality measurement of multitasking, which our HFTU methodology can contribute (McGavock et al., 2025).

Overall, our findings highlight the complexity of the links between household labor allocation, gender norms, and mental well-being. They also suggest an important insight for

economists and policymakers—re-balancing unpaid domestic and care work need not entail trade-offs between spouses in a zero-sum framework. Instead, norm-sensitive interventions that engage men directly can generate joint gains in well-being within households, challenging classical frameworks that view intra-household allocations as primarily efficiency-driven (Becker, 1965; Folbre, 2006) and opening new possibilities for improving welfare, especially in low-resource settings.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve the clarity of exposition during the editing phase. After using this tool/service, the authors reviewed and edited all content in the manuscript and take full responsibility for everything contained in this publication.

References

Addati, L., Cattaneo, U., Esquivel, V., & Valarino, I. (2018). *Care work and care jobs for the future of decent work*. International Labour Organisation (ILO).

Agness, D., Baseler, T., Chassang, S., Dupas, P., & Snowberg, E. (2025). Valuing the time of the self-employed. *The Review of Economic Studies*, *92*, 3471–3503.

Alderman, H., Gilligan, D. O., Hidrobo, M., Leight, J., Mulford, M., & Tambet, H. (2025). Men can cook: Effectiveness of a men's engagement intervention to change attitudes and behaviors in rural Ethiopia. *World Development*, *185*, 106781.

Andersen, L. M. B., Rasmussen, A. N., Reavley, N. J., Bøggild, H., & Overgaard, C. (2021). The social route to mental health: A systematic review and synthesis of theories linking social relationships to mental health to inform interventions. *SSM - Mental Health*, *1*, 100042.

Anderson, M. L. (2008). 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*, *103*(484), 1481–1495.

Asensio, O. I., & Delmas, M. A. (2016). The dynamics of behavior change: Evidence from energy conservation. *Journal of Economic Behavior & Organization*, *126*, 196–212.

Asfaw, M. N., Mengesha, A. K., Gugerell, K., Mansberger, R., & Damyanovic, D. (2025). Land rights for women in rural Ethiopia: Legal framework and reality. *World Development Perspectives*, *39*.

Assefa, T., Getahun, T., McCullough, E., & McGavock, T. (2023). Enhancing the impacts of gender training: Encouraging men to share the load through repeated remote reinforcement messaging. *AEA RCT Registry. August 10*.

Avdeenko, A., Bohne, A., & Frölich, M. (2019). Linking savings behavior, confidence and individual feedback: A field experiment in Ethiopia. *Journal of Economic Behavior & Organization*, *167*, 122–151.

Baek, S. U., Lee, Y. M., Won, J. U., & Yoon, J. H. (2024). Association between husband's participation in household work and the onset of depressive symptoms in married women: A population-based longitudinal study in South Korea. *Social Science and Medicine*, *362*.

Bandura, A., & Walters, R. H. (1977). *Social Learning Theory*, vol. 1. Prentice-Hall.

Banmen, J., & Maki-Banmen, K. (2014). What has become of Virginia Satir's therapy model since she left us in 1988? *Journal of Family Psychotherapy*, *25*(2), 117–131.

Baranov, V., Bhalotra, S., Biroli, P., & Maselko, J. (2020). Maternal depression, women's empowerment, and parental investment: Evidence from a randomized controlled trial.

American Economic Review, *110*, 824–859.

Bastos, M. A. V. J., Braz, D. F., Porto, A. L. M., da Silva Cordeiro, K. S., Portella, R. B., & Granger, D. A. (2025). Physiological attunement and flourishing: Understanding the influence of relationships on health. *Frontiers in Psychiatry*, *16*, 1614379.

Becker, G. S. (1965). A theory of the allocation of time. *The Economic Journal*, *75*(299), 493–517.

Becker, G. S. (1981). *A Treatise on the Family*. NBER Books. Cambridge, MA: National Bureau of Economic Research.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

Bénabou, R., & Henkel, L. (2025). Identity as self-image. Working Paper 34297, National Bureau of Economic Research.

Bicchieri, C. (2016). *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press.

Blackden, C. M., & Wodon, Q. (2006). *Gender, time use, and poverty in Sub-Saharan Africa*. The World Bank.

Blau, F. (2025). Gender inequality in the labor market: Continuing progress? *ILR Review*.

Bloomfield, J., Balsa, A. I., Cid, A., & Oreopoulos, P. (2025). Calling all parents: Leveraging behavioral insights to boost early childhood outcomes in the developing world. Working Paper 33338, National Bureau of Economic Research.

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2019). Beliefs about gender.

Borra, C., Browning, M., & Sevilla, A. (2020). Marriage and housework. *Oxford Economic Papers*, *73*, 479–508.

Borrell-Porta, M., Costa-Font, J., & Philipp, J. (2019). The 'mighty girl' effect: Does parenting daughters alter attitudes towards gender norms? *Oxford Economic Papers*, *71*(1), 25–46.

Boye, K. (2009). Relatively different? how do gender differences in well-being depend on paid and unpaid work in Europe? *Social Indicators Research*, *93*, 509–525.

Browning, M., Donni, O., & Gørtz, M. (2020). Do you have time to take a walk together? private and joint time within the household. Working Paper 01/20, Center for Economic Behavior and Inequality (CEBI).

Browning, M., & Gørtz, M. (2012). Spending time and money within the household. *The Scandinavian Journal of Economics*, *114*, 681–704.

Bursztyn, L., Egorov, G., & Fiorin, S. (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review*, *110*(11), 3522–3548.

Buzard, K., Gee, L. K., & Stoddard, O. (2025). Who you gonna call? gender inequality in external demands for parental involvement. *The Quarterly Journal of Economics*, *140*(4), 2805–2849.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*, vol. 100. Cambridge University Press.

Carvalho, L., de Walque, D., Lund, C., Schofield, H., Somville, V., & Wei, J. (2025). Psychological barriers to participation in the labor market: Evidence from rural Ghana. Tech. rep., World Bank Policy Research Working Paper 11271.

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort

in economics experiments. *Journal of Economic Behavior & Organization*, *149*, 74–87.
Csikszentmihalyi, M. (2014). *Flow and the Foundations of Positive Psychology*. Springer.
Cunha, F., Hu, Q., Xia, Y., & Zhao, N. (2023). Reducing bullying: Evidence from a parental involvement program on empathy education. Working Paper 30827, National Bureau of Economic Research.
Daminger, A. (2019). The cognitive dimension of household labor. *American Sociological Review*, *84*, 609–633.
de Quidt, J., & Haushofer, J. (2016). Depression for economists. Working Paper 22973, National Bureau of Economic Research.
Delavallade, C., & Das, S. (2024). How socio-emotional skills matter for economic empowerment: Theory of change & research agenda. ESTEEM (Effective Socio-emotional Skills to Gain Economic Empowerment) project page; accessed December 30, 2025. URL https://poverty-action.org/ses-theory-of-change
Dhar, D., Jain, T., & Jayachandran, S. (2019). Intergenerational transmission of gender attitudes: Evidence from India. *The Journal of Development Studies*, *55*(12), 2572–2592.
Dhar, D., Jain, T., & Jayachandran, S. (2022). Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in India. *American Economic Review*, *112*(3), 899–927.
Dinkelman, T., & Ngai, L. R. (2022). Time use and gender in Africa in times of structural transformation. *Journal of Economic Perspectives*, *36*, 57–80.
Doepke, M., & Kindermann, F. (2019). Bargaining over babies: Theory, evidence, and policy implications. *American Economic Review*, (p. 3264–3306).
Drozd, L. A., & Tavares, M. M. (2025). Generative AI: A turning point for labor's share? Working paper, Federal Reserve Bank of Philadelphia, Research Department.
Dupas, P. (2011). Health behavior in developing countries. *Annual Review of Economics*, *3*, 425–449.
EDHS, E. (2016). Demographic and health survey 2016: Key indicators report. *The DHS Program ICF*, *363*, 364.
Edin, P.-A., Fredriksson, P., Nybom, M., & Öckert, B. (2022). The rising return to noncognitive skill. *American Economic Journal: Applied Economics*, *14*(2), 78–100.
Ervin, J., Churchill, B., Ruppanner, L., & King, T. (2025). Unpaid labour and mental health-the role of perceived fairness and satisfaction in division amongst working-age adults; a longitudinal analysis using 18 waves of panel data. *Social Science & Medicine*, *384*, 118559.
Ervin, J., Taouk, Y., Alfonzo, L. F., Hewitt, B., & King, T. (2022). Gender differences in the association between unpaid labour and mental health in employed adults: a systematic review. *The Lancet Public Health*, *7*, e775–e786.
Ervin, J., Taouk, Y., Hewitt, B., & King, T. (2023). The association between unpaid labour and mental health in working-age adults in Australia from 2002 to 2020: a longitudinal population-based cohort study. *Articles Lancet Public Health*, *8*, e276–85.
Evans, D. K., & Jakiela, P. (2024). The role of fathers in promoting early childhood development in low- and middle-income countries: A review of the evidence. *The World Bank Research Observer*, *40*(2), 211–228.
Fan, Y., Fan, A., Yang, Z., & Fan, D. (2025). Global burden of mental disorders in 204 countries and territories, 1990–2021: results from the Global Burden of Disease study 2021. *BMC Psychiatry*, *25*(1), 486.

Fisher, J., Herrman, H., de Mello, M. C., & Chandra, P. S. (2013). Women's mental health. In V. Patel, H. Minas, A. Cohen, & M. J. Prince (Eds.) *Global Mental Health: Principles and Practice*, (p. 354). Oxford: Oxford University Press.

Fitzsimons, E., Malde, B., Mesnard, A., & Vera-Hernández, M. (2016). Nutrition, information and household behavior: Experimental evidence from Malawi. *Journal of Development Economics*, *122*, 113–126.

Folbre, N. (2006). Measuring care: Gender, empowerment, and the care economy. *Journal of Human Development*, *7*(2), 183–199.

Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, *1*, 358–368.

Glass, J., & Fujimoto, T. (1994). Housework, paid work, and depression among husbands and wives. *Journal of Health and Social Behavior*, *35*, 179–191.

Glischinski, M. v., Brachel, R. v., Thiele, C., & Hirschfeld, G. (2021). Not sad enough for a depression trial? A systematic review of depression measures and cut points in clinical trial registrations. *Journal of Affective Disorders*, *292*, 36–44.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, *104*(4), 1091–1119.

Haghighi, P., Littler, E. A. L., Mauer-Vakil, D., Miller, M., & Oremus, M. (2024). Exploring the relationship between marital quality and cognitive function: A systematic review. *Social Science & Medicine*, *355*, 117120.

Hammarberg, K., Tran, T., Kirkman, M., & Fisher, J. (2020). Sex and age differences in clinically significant symptoms of depression and anxiety among people in Australia in the first month of covid-19 restrictions: A national survey. *BMJ Open*, *10*.

Hardy, M., Kagy, G., & Jimi, N. (2025). Mind the data gaps: An examination of women-owned enterprise representation. *Available at SSRN*.

Harryson, L., Novo, M., & Hammarström, A. (2012). Is gender inequality in the domestic sphere associated with psychological distress among women and men? results from the Northern Swedish Cohort. *Journal of Epidemiology & Community Health*, *66*(3), 271–276.

Heckman, J. J., & Singer, B. (2017). Abducting economics. *American Economic Review*, *107*, 298–302.

Heckman, J. J., Tian, H., Zhang, Z., & Zhou, J. (2026). Dynamic complementarity. *American Economic Journal: Applied Economics (accepted)*.

Heggeness, M. L., Bommarito, J., & Prewitt, L. (2025). The care board: Version 1.5 [online dataset and dashboard]. https://thecareboard.org. Interactive dashboard visualizing care work and care economy data.

Hook, J. L. (2006). Care in context: Men's unpaid work in 20 countries, 1965–2003. *American Sociological Review*, *71*, 639–660.

Jayachandran, S. (2015). The roots of gender inequality in developing countries. *Annual Review of Economics*, *7*(1), 63–88.

Kerbaj, J., Toure, Y., Soto Aladro, A., Boudjema, S., Giorgi, R., Dufour, J. C., & Brouqui, P. (2017). Smartphone text message service to foster hand hygiene compliance in health care workers. *American Journal of Infection Control*, *45*, 234–239.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613.

Krueger, A. (2007). Does anybody really know what they were doing yesterday? Research

proposal, Princeton University.

Krueger, A. B., Kahneman, D., Schkade, D., Schwarz, N., & Stone, A. A. (2009). *Measuring the Subjective Well-Being of Nations: National Accounts of Time Use and Well-Being*. October. University of Chicago Press.

Kuehner, C. (2017). Why is depression more common among women than among men? *The Lancet Psychiatry*, *4*, 146–158.

Luoto, J., Levine, D., Albert, J., & Luby, S. (2014). Nudging to use: Achieving safe water behaviors in Kenya and Bangladesh. *Journal of Development Economics*, *110*, 13–21.

Madajewicz, M., Pfaff, A., van Geen, A., Graziano, J., Hussein, I., Momotaj, H., Sylvi, R., & Ahsan, H. (2007). Can information alone change behavior? response to arsenic contamination of groundwater in Bangladesh. *Journal of Development Economics*, *84*, 731–754.

Magaña, I., Martínez, P., & Loyola, M. S. (2020). Health outcomes of unpaid caregivers in low- and middle-income countries: A systematic review and meta-analysis.

Martingano, A. J., Konrath, S., Blanch-Hartigan, D., Davis, M. H., Hall, J. A., Ruben, M. A., Sanders, J. J., & Schwartz, R. (2025). Seven empathy myths: Correcting misconceptions about a complex construct. *Social and Personality Psychology Compass*, *19*, e70093.

McGavock, T., McCullough, E., & Assefa, T. (2025). Things take the time they take: Lessons on measurement error from high-frequency time use surveys. *Working Paper*. URL https://shorturl.at/a6yKn

McGavock, T., McCullough, E., Magnan, N. P., & Assefa, T. (2020). Stifled by stigma? experimental effects of updating husbands' beliefs on participation in women's household work. *Working Paper*.

Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implement. Sci.*, *6*(1), 42.

Molarius, A., & Metsini, A. (2023). The association between time spent in domestic work and mental health among women and men. *International Journal of Environmental Research and Public Health*, *20*.

Musick, K., Meier, A., & Flood, S. (2016). How parents fare: Mothers' and fathers' subjective well-being in time with children. *American Sociological Review*, *81*, 1069–1095.

Narayan, D. (2023). Gender equality and collective wellbeing: The power of changing mindsets. Tech. rep., World Bank.

Ngai, L. R., & Petrongolo, B. (2017). Gender gaps and the rise of the service economy. *American Economic Journal: Macroeconomics*, *9*, 1–44.

Nguyen, C. V., & Tarp, F. (2022). Changing male perceptions of gender equality: Evidence from a randomised controlled trial study. *World Development*, *158*, 106019.

Ostervall, L. W. (2017). Nudging to prudence? the effect of reminders on antibiotics prescriptions. *Journal of Economic Behavior & Organization*, *135*, 39–52.

Owoo, N. S., & Lambon-Quayefio, M. P. (2021). Mixed methods exploration of ghanaian women's domestic work, childcare and effects on their mental health. *PLoS ONE*, *16*.

Perales, F., Kuskoff, E., Flood, M., et al. (2023). Like father, like son: Empirical insights into the intergenerational continuity of masculinity ideology. *Sex Roles*, *88*, 399–412.

Petit, V. (2019). The behavioural drivers model: A conceptual framework for social and behaviour change programming. Report, UNICEF.

Piovani, C., & Aydiner-Avsar, N. (2021). Work time matters for mental health: A gender analysis of paid and unpaid labor in the United States. *Review of Radical Political Economics*, *53*, 579–589.

Prins, S. J., McKetta, S., Platt, J., Muntaner, C., Keyes, K. M., & Bates, L. M. (2019). Mental illness, drinking, and the social division and structure of labor in the United States: 2003-2015. *American Journal of Industrial Medicine*, *62*(2), 131–144.

Prochaska, J. O. (1984). *Systems of Psychotherapy: A Transtheoretical Analysis*. Brooks-Cole, 2nd ed.

Rendall, M. (2018). Female market work, tax regimes, and the rise of the service sector. *Review of Economic Dynamics*, (p. 269–89).

Riecher-Rössler, A. (2017). Sex and gender differences in mental disorders.

Rodríguez, C., & Saavedra, J. E. (2019). The persistent effects of youth savings reminders: Experimental evidence from text-message campaigns in Colombia. *Journal of Development Economics*, *139*, 135–156.

Sabates–Wheeler, R., Hirvonen, K., Lind, J., & Hoddinott, J. (2022). Expanding social protection coverage with humanitarian aid: Lessons on targeting and transfer values from Ethiopia. *The Journal of Development Studies*, *58*(10), 1981–2000.

Sayer, L. C. (2005). Gender, time and inequality: Trends in women's and men's paid work, unpaid work and free time. *Social Forces*, *84*(1), 285–303.

Shah, A. K., & Ludwig, J. (2016). Option awareness: The psychology of what we consider. *American Economic Review*, *106*, 425–429.

Sinharoy, S., Cheong, Y. F., Seymour, G., Heckert, J., Johnson, E. R., & Yount, K. M. (2023). The time-use agency scale: Development and validation of a measure for Ghana and beyond. *Feminist Economics*, *29*, 103–132.

Sirois, F. M., & Pychyl, T. (2013). Procrastination and the priority of short-term mood regulation: Consequences for future self. *Social and Personality Psychology Compass*, *7*, 115–127.

Swerbenski, H. G., Sturge-Apple, M. L., Koven, M., & Davies, P. T. (2023). Strengths-based spillover models: Constructive interparental conflict, parental supportive problem solving, and development of child executive functioning. *Journal of Family Psychology*, *37*, 1060–1071.

The Gottman Institute (2018). Disagreements about housework ranked 3rd leading reason for divorce in study.

Trommsdorff, G. (2009). Intergenerational relations and cultural transmission. In U. Schönpflug (Ed.) *Cultural Transmission: Psychological, Developmental, Social, and Methodological Aspects*, (pp. 126–160). Cambridge: Cambridge University Press.

Verhofstadt, L., Devoldre, I., Buysse, A., Stevens, M., Hinnekens, C., Ickes, W., & Davis, M. (2016). The role of cognitive and affective empathy in spouses' support interactions: An observational study. *PLOS ONE*, *11*, e0149944.

Vieten, C., Rubanovich, C. K., Khatib, L., Sprengel, M., Tanega, C., Polizzi, C., Vahidi, P., Malaktaris, A., Chu, G., Lang, A. J., Tai-Seale, M., Eyler, L., & Bloss, C. (2024). Measures of empathy and compassion: A scoping review. *PLOS ONE*, *19*, e0297099.

Väänänen, A., Kevin, M. V., Ala-Mursula, L., Pentti, J., Kivimäki, M., & Vahtera, J. (2004). The double burden of and negative spillover between paid and domestic work: Associations with health among men and women. *Women and Health*, *40*, 1–18.

Wiseman, T. (1996). A concept analysis of empathy. *Journal of Advanced Nursing*, *23*(6), 1162–1167.

World Health Organization, & Columbia University (2016). Group Interpersonal Therapy (IPT) for depression. Tech. rep., WHO.

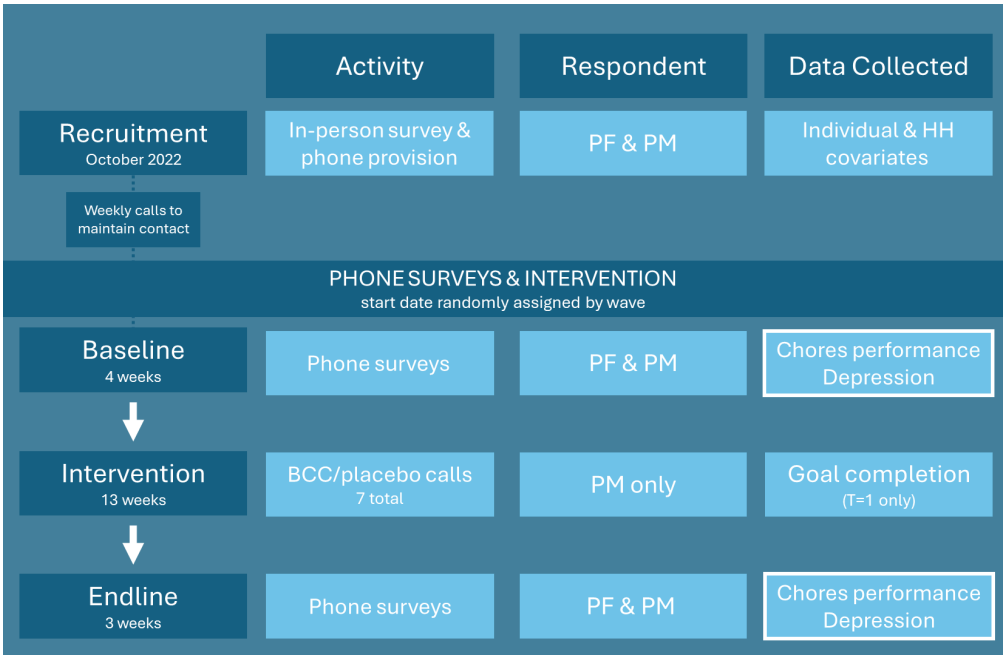
Xue, B., & McMunn, A. (2021). Gender differences in unpaid care work and psychological distress in the UK Covid-19 lockdown. *PLoS ONE*, *16*(3 March).

Zurovac, D., Sudoi, R., Akhwale, W., Ndiritu, M., Hamer, D., Rowe, A., & Snow, R. (2011). The effect of mobile phone text-message reminders on Kenyan health workers' adherence to malaria treatment guidelines: a cluster randomised trial. *The Lancet*, *378*, 795–803.

A Supplementary Figures and Tables

A.1 Supplementary Figures

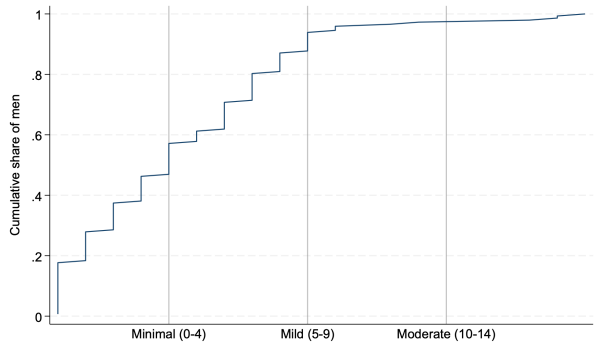
Figure A1: Timeline of survey and intervention activities



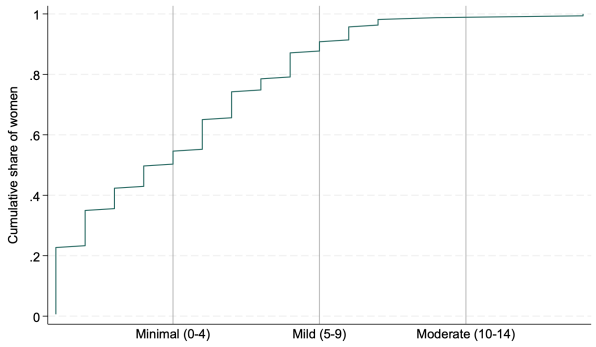
Note: The figure shows activities across the study's four phases indicated in the leftmost column. The recruitment phase consisted of in-person surveys only, during which all households were issued mobile phones by survey staff. All subsequent interactions were via phone. Baseline took four weeks per household, spanning January–June 2023 across all households. The intervention lasted thirteen weeks per household, spanning February–September 2023, and endline took three weeks per household, spanning June–November 2023. PF refers to the primary female respondent, i.e., the primary female caretaker of the index child, the youngest child between ages 24 and 60 months at the time of the recruitment interview that determined household eligibility. The primary male respondent (PM) is that woman's male spouse or, when not available, another male who is primarily responsible for the index child's care. The intervention (a set of seven calls over 13 weeks) was targeted to the PM. We interviewed both the PF and the PM at baseline and endline. We measured depressive symptoms during both baseline and endline phone interviews from both the PF and the PM using the PHQ-9 instrument as described in Section 3.5.

Figure A2: Baseline depression scores (PHQ-9) for the PM and PF

(a) Primary Male

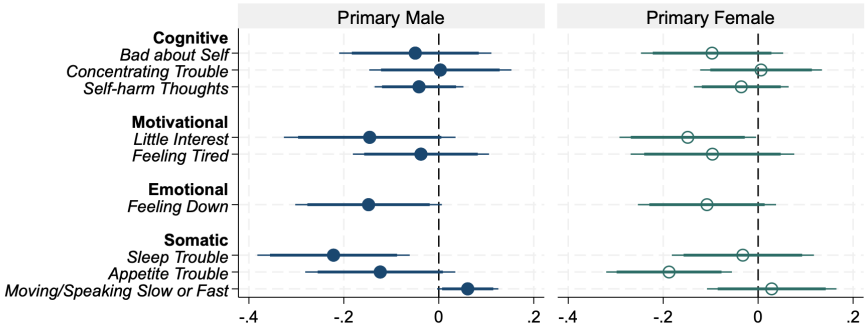


(b) Primary Female

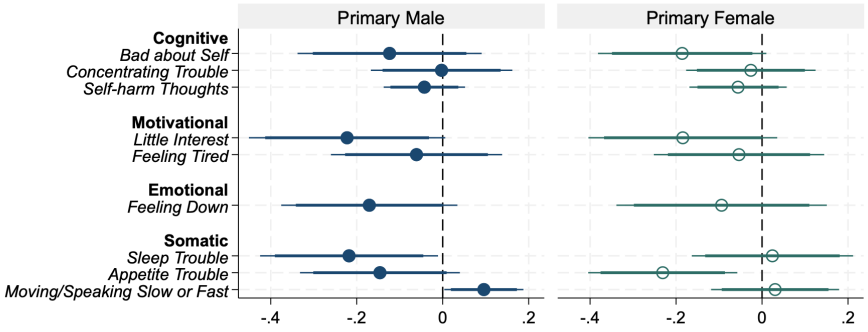


Note: Panels (a) and (b), respectively, show the cumulative density functions of PHQ-9 scores for the PM and PF at baseline. The vertical gray lines show thresholds used as guidelines in clinical settings (Fan et al., 2025). The cutoffs are: 0–4 (minimal), 5–9 (mild), 10–14 (moderate), 15–19 (moderately severe), and 20-27 (severe).

Figure A3: Treatment effects on depression symptom presence and frequency (PHQ-9)



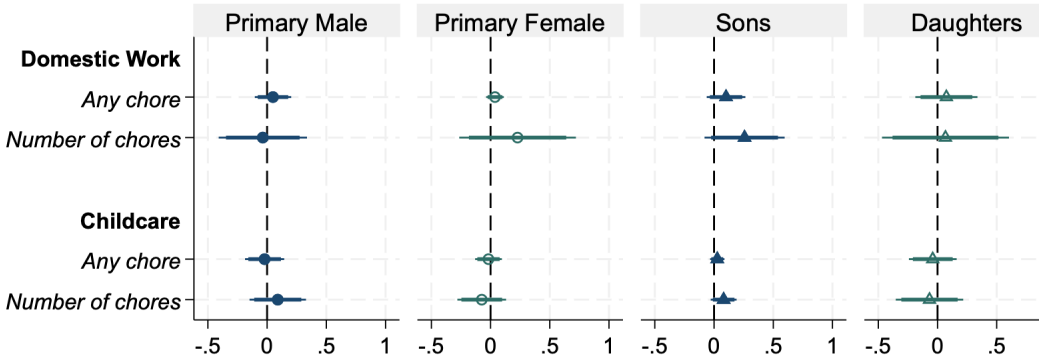
(a) Presence of symptom to any degree



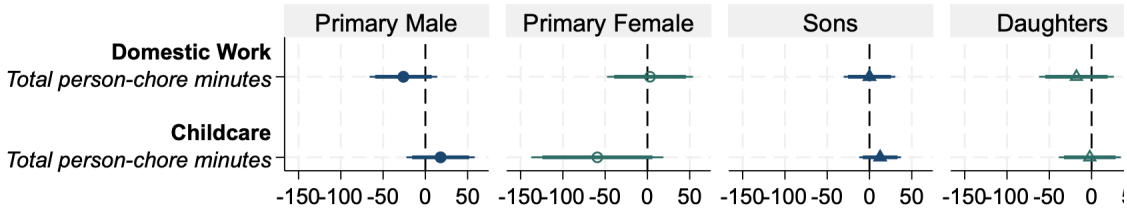
(b) Frequency of symptom experience

Note: Panel (a) reports the impact of the intervention on a binary measure of whether each symptom was observed to any extent by the PM (left column) or PF (right column). Panel (b) reports the impact on the frequency that the symptom is reported by the respondent, between 0 (not at all), 1 (several days), 2 (more than half the days) and 3 (nearly every day). The symptoms have been grouped into categories by type. The bars around each coefficient depict the 90% and 95% confidence intervals. The treatment effects are estimated based on PM and PF self-reporting of their PHQ-9 depression symptoms during the two weeks prior to the interview.

Figure A4: Treatment effects on chores performance (PM reported version of Figure 4)



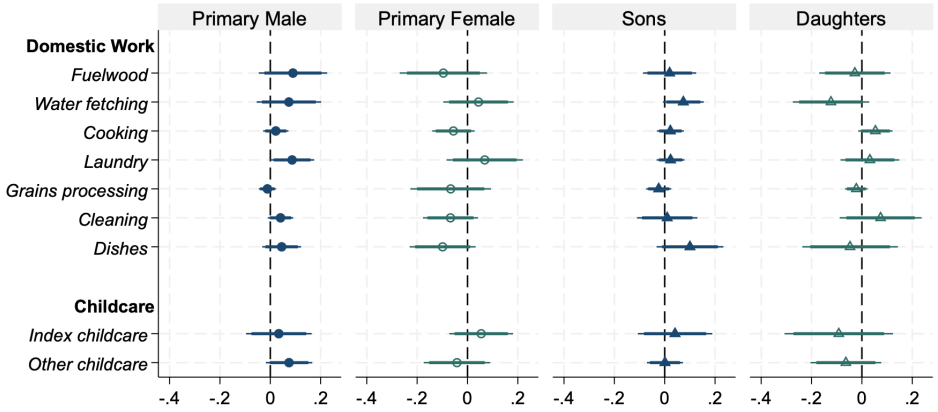
(a) Performance of chores



(b) Minutes performing chores

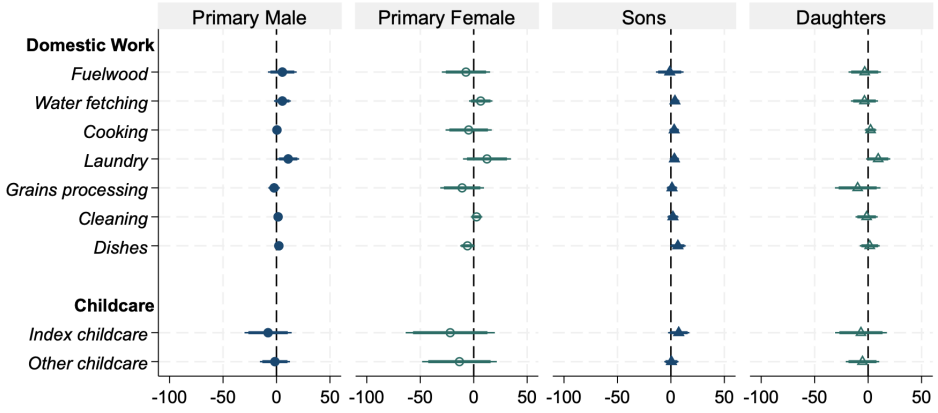
Note: The figure reports the impact of the intervention on chores performance outcomes at the person type level. Panel (a) reports the effect on a binary measure of participation in any domestic work chore, a binary measure of participation in any childcare task, the number of domestic work chores, and the number of childcare tasks. Panel (b) reports the effect on total minutes performing domestic work and childcare, respectively. The figure depicts the coefficient on treatment interacted with person-type for the specification that includes covariates selected (as described in Eq. 1). The bars around each coefficient depict its 90% and 95% confidence intervals. This figure relies on chores performance as reported by the PM. PF-reported results are reported in Figure 4.

Figure A5: Treatment effects on chores performance by person type and chore type (PF reported)



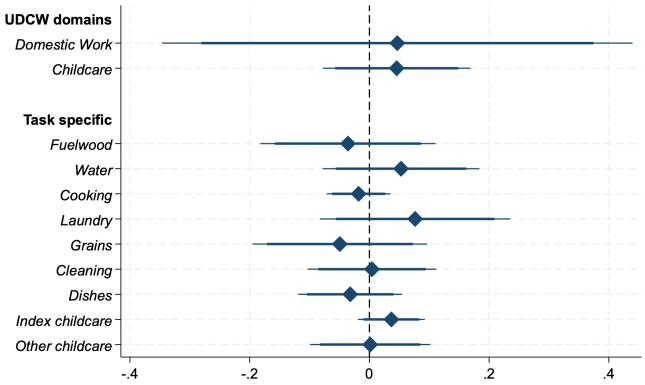
Note: The figure reports the impact of the intervention on a binary measure of participation in chores at the person-type and chore level. The bars around each coefficient depict the 90% and 95% confidence intervals. Each treatment effect reports the coefficient on treatment interacted with the corresponding column's person type and the corresponding row's specific chore for an adapted version of the specification (see, Footnote 30) as described in Eq. 1 that includes covariates selected. The outcome is chore completion on the day before the interview, as reported by the primary female.

Figure A6: Treatment effects on total chore minutes by person type and chore type (PF reported)

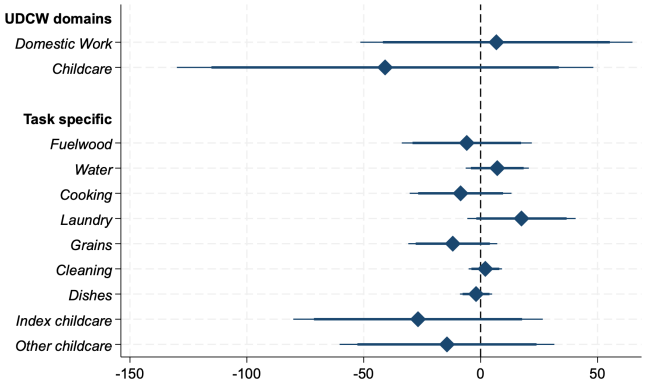


Note: The figure reports the impact of the intervention on total chore minutes at the person-type and chore level. Each treatment effect reports the coefficient on treatment interacted with the corresponding column's person type and the corresponding row's specific chore for an adapted version of the specification (see, Footnote 30) as described in Eq. 1 that includes covariates selected. The bars around each coefficient depict the 90% and 95% confidence intervals. The treatment effects are estimated based on chore completion on the day before the interview, as reported by the primary female.

Figure A7: Treatment effects on chores performance at the household level



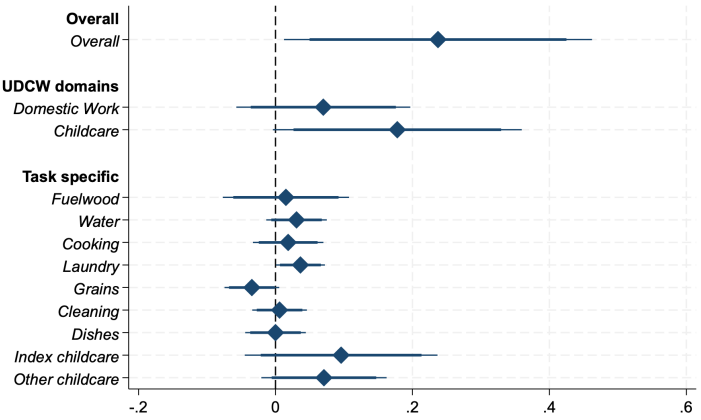
(a) Count of chores performed count by UDCW domain (top) and binary task outcome (bottom)



(b) Minutes by UDCW domain (top) and by task (bottom)

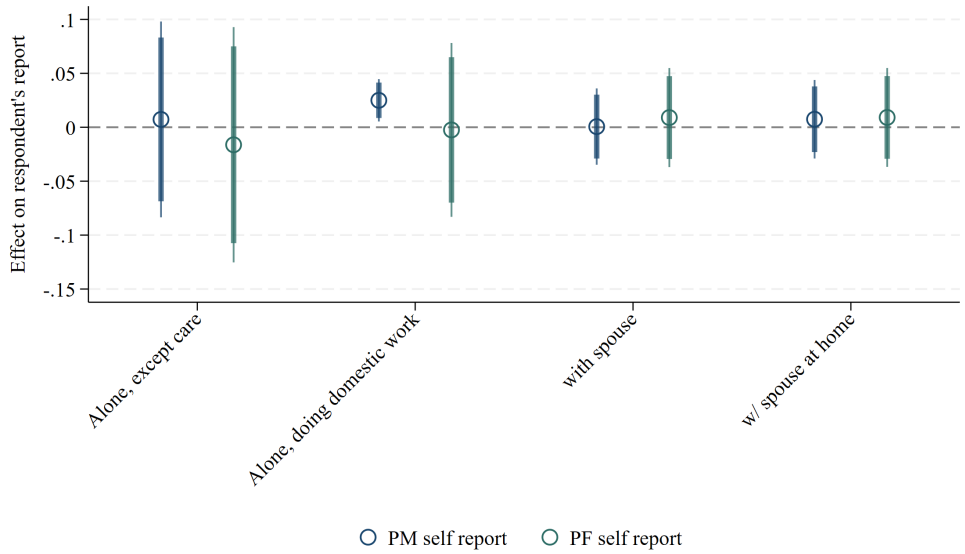
Note: The figure reports the impact of the intervention on performance of chores for the household: (a) reports coefficients on the number of domestic work or childcare tasks (top) and then binary outcomes for each of the specific chores indicating whether they were performed (bottom); (b) reports the total number of minutes spent on domestic work and childcare (top) and then the chores split up (bottom). The chores completion outcomes are based on chores completion on the day before the interview as reported by the PF. The bars around each coefficient depict its 90% and 95% confidence intervals. Estimation is described in Section 6.1 and baseline covariates are selected using PDS lasso.

Figure A8: Treatment effects on chores sharing (performance by > 1 household member)



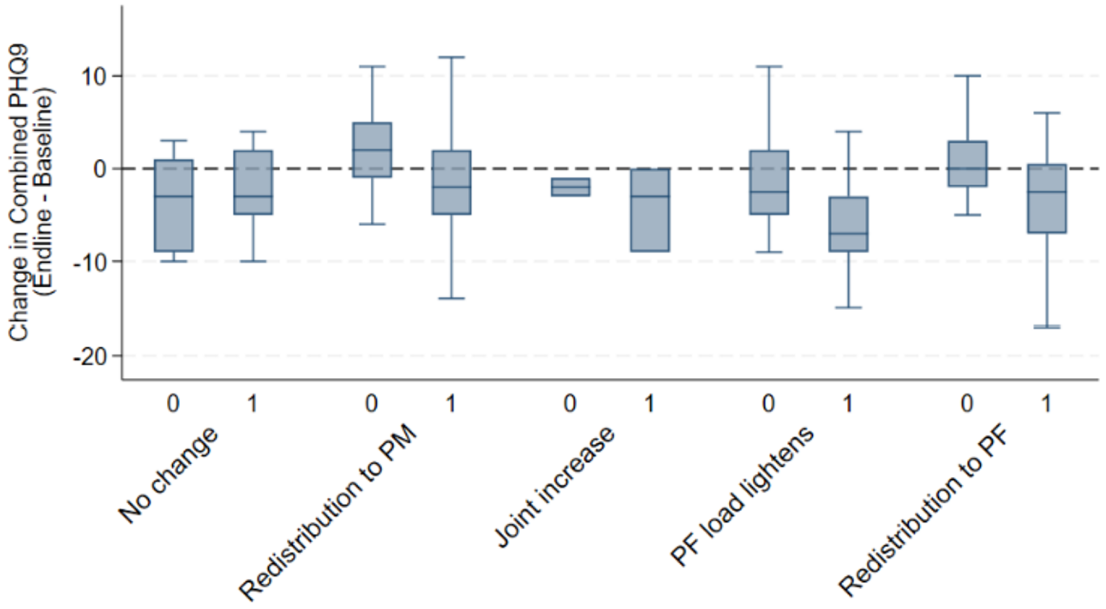
Note: The figure reports the impact of the intervention on whether chores were performed by more than one household member. The figure reports coefficients on the number of all tasks (top section), the number of domestic work or childcare tasks (middle section) and then binary outcomes for each of the specific chores indicating whether they were performed by more than one member (bottom section). The chores completion outcomes are based on chores completion on the day before the interview as reported by the PF. The bars around each coefficient depict its 90% and 95% confidence intervals. Estimation is described in Section 6.3 and baseline covariates are selected using PDS lasso.

Figure A9: Treatment effects on companions by activity type (HFTU data)



Note: The figure reports the effect of the intervention on the number of companions with the PM and the PF generally, while performing DW chores, and while performing childcare chores. The figure also reports effects on the probability that the respondent reports being with their spouse during all activities or while at home. These treatment effects are based on a call level estimation using the HFTU data as further described in Section 6.2 and Appendix E.

Figure A10: Change in depression score (PHQ-9) by chores redistribution category



Note: The figure reports box plots for the change between baseline and endline for the combined household PHQ-9 score (i.e., sum of PM and PF) where the height of bars is the inter-quartile range (bottom of blue box is the 25th percentile and the top is the 75th percentile). Negative values imply either that both spouses experienced a decline (improvement in depression symptoms) or that the decline in one spouse's score exceeds the increase in the other's. Error bars show the adjacent values ($+/-1.5IQR$) and outside values (outliers) are omitted. Each group along the x-axis is defined as in Section 6 based on the change in number of chores for PM and PF between baseline and endline. Within each redistribution category, PHQ-9 scores are reported by treatment (0 corresponds to the control group and 1 to the treatment group). Figure 9 shows treatment effects on the likelihood of each household falling into each redistribution category.

A.2 Supplementary Tables

Design, Attrition, and Baseline Summary Tables

Table A1: Individual chores tracked in the chores module

Domestic Work Chores	**Care Chores**
Fuelwood collection	Childcare (index child)
Water fetching	Childcare (other children)
Cooking	
Laundry	
Grains Processing	
Cleaning	
Dishes	

Note: This table lists the nine chores that were tracked as part of the chores survey module. For the analysis, we have categorized each chore as either "domestic work" in that it is related to unpaid provision of goods and services apart from care and "care" chores that are related to caregiving. Baseline performance of each chore at the household person-type level is summarized in Table A7.

Table A2: Attrition for interviews with the PF and PM, by treatment group

	Treatment	Control	Total	P-value
Sampled Households				
Recruited & received phone	96	96	192	
Baseline survey completed with PF	83	81	164	
Endline survey completed with PF	83	80	163	
Endline survey completed with PF (*and PM*)	76	71	147	
Attrition Rates Relative to Recruited Sample				
% HHs not completing baseline with PF	13.5%	15.6%	14.9%	0.67
% HHs not completing endline with PF	13.5%	16.7%	15.1%	0.52
% HHs not completing endline with PF (*and PM*)	20.8%	26.04%	23.4%	0.303

Note: This table reports the number of households recruited into the study during the recruitment phase (first row), the number of households from which the primary female respondent (PF) completed her baseline chores interview (second row), and the number of households these households from whom the PF *also* completed an endline chores interview (third row). The fourth row reports the number of households in which the primary male respondent (PM) completed the endline interview, conditional on the PF having completed the endline interview, which defines the analysis sample for PM mental health outcomes. Recruitment and phone provision were completed in person in October 2022. Baseline (endline) chores interviews were completed by phone with the PF during the week before (after) all intervention calls took place with the PM. For more details on the timeline of data collection and intervention calls see Section 3.2. The attrition rates report, by treatment group, the percentage of households for whom we do not observe complete data from the PF. Differences in attrition rates across treatment and control groups are not statistically significant at any standard levels as shown in the last column.

Table A3: Baseline characteristics of study sample, by treatment group

	Treatment arm (N=83)	Control arm (N=80)	P-value
Primary male characteristics			
Age in years	33.92	33.54	0.70
	(8.19)	(6.96)	
Completed: any school	0.39	0.30	0.30
primary school	0.18	0.29	0.10
secondary school	0.08	0.15	0.17
Main activity in last 7 days is: ag	0.51	0.45	0.35
non-ag	0.10	0.13	0.51
daily laborer	0.17	0.18	0.91
unpaid housework	0.01		0.32
none/other non-market activity	0.19	0.25	0.31
Primary female characteristics			
Age in years	28.27	28.01	0.71
	(5.98)	(4.86)	
Completed: any school	0.33	0.36	0.66
primary school	0.17	0.13	0.48
secondary school	0.06	0.03	0.34
Main activity in last 7 days is: ag	0.04	0.01	0.34
non-ag	0.06	0.05	0.75
daily laborer		0.01	0.32
unpaid housework	0.90	0.91	0.84
none/other non-market activity		0.01	0.33
Household characteristics			
Age gap between PM and PF	5.65	5.53	0.86
	(6.69)	(5.17)	
No. of young children (under 5): males	0.66	0.59	0.27
	(0.50)	(0.54)	
females	0.48	0.50	0.83
	(0.61)	(0.53)	
No. of older children (5-18): males	0.65	0.74	0.40
	(0.89)	(0.84)	
females	0.61	0.60	0.87
	(0.76)	(0.74)	
No. of adults ($\geq$ 18): males apart from PM	0.10	0.05	0.31
	(0.34)	(0.22)	
females apart from PF	0.13	0.09	0.39
	(0.61)	(0.53)	
Total value (USD, PPP) of: productive assets	57.41	63.16	0.16
	(30.49)	(35.51)	
consumptive assets	175.22	160.71	0.55
	(192.36)	(155.55)	
livestock assets	664.01	584.68	0.28
	(696.97)	(620.84)	
PM/PF are spouses	0.94	0.98	0.20
PM/PF are parents of index child	0.99	0.99	0.84
Primary female is sole hh head	0.07	0.06	0.81
HH owned phone before study	0.71	0.73	0.85

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The table reports the mean of baseline characteristics, by treatment group, for the PM, the PF, and their households. For continuous variables, the standard deviation is reported in parentheses below the mean. All variables with no reported standard deviation are binary. For the asset values reported, we convert 2022 ETB into current international dollars using the PPP conversion factor for private consumption (20.83 ETB per USD). The third column reports the p-value for a t-test of the null hypothesis that the covariate does not differ by treatment status. These variables are all included as possible covariates in the PDS Lasso variable selection procedure discussed in Section 4, with the exception of the index child's relation to the HH head (due to no variation) and the age gap between the PM and the PF (since this is collinear with their respective ages).

Table A4: Baseline presence of specific mental health symptom by person type and treatment arm

	Primary Male	**Primary Female**
Having little interest or pleasure in doing things		
Treatment	0.53	0.47
Control	0.48	0.40
P-value	0.60	0.36
Feeling down, depressed, or hopeless		
Treatment	0.55	0.47
Control	0.56	0.51
P-value	0.89	0.57
Having trouble falling asleep, staying asleep, or sleeping too much		
Treatment	0.59	0.51
Control	0.56	0.46
P-value	0.72	0.58
Feeling tired or having little energy		
Treatment	0.62	0.54
Control	0.48	0.48
P-value	0.07*	0.40
Having a poor appetite or overeating		
Treatment	0.32	0.40
Control	0.41	0.38
P-value	0.26	0.79
Feeling bad about yourself – or that you are a failure or have let yourself or your family down		
Treatment	0.46	0.46
Control	0.48	0.45
P-value	0.82	0.91
Having trouble concentrating on things, such as cooking food, coffee ceremony, or listening to the radio		
Treatment	0.29	0.37
Control	0.32	0.30
P-value	0.58	0.28
Moving or speaking so slowly that other people could have noticed. Or, the opposite – being so fidgety or restless that you have been moving around a lot more than usual		
Treatment	0.32	0.28
Control	0.24	0.24
P-value	0.26	0.56
Having thoughts that you would be better off dead or of hurting yourself in some way		
Treatment	0.29	0.27
Control	0.24	0.26
P-value	0.48	0.97
Observations	147	163

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The table summarizes presence of specific mental health symptoms (to any degree) at the person-type level by treatment arm. Mean values are reported. The table also reports the P-value associated with a t-test of the null hypothesis that each pre-treatment outcome does not differ by treatment status.

Table A5: Baseline measure of whether N or more mental health symptoms are observed by person type and treatment arm

	Primary Male	Primary Female
1+ symptoms		
Treatment	0.86	0.77
Control	0.79	0.78
P-value	0.31	0.94
2+ symptoms		
Treatment	0.68	0.61
Control	0.72	0.68
P-value	0.84	0.90
3+ symptoms		
Treatment	0.61	0.57
Control	0.62	0.57
P-value	0.61	0.41
4+ symptoms		
Treatment	0.53	0.53
Control	0.52	0.46
P-value	0.94	0.31
5+ symptoms		
Treatment	0.43	0.45
Control	0.39	0.40
P-value	0.61	0.48
6+ symptoms		
Treatment	0.38	0.36
Control	0.32	0.26
P-value	0.48	0.14
7+ symptoms		
Treatment	0.28	0.24
Control	0.24	0.14
P-value	0.59	0.12
8+ symptoms		
Treatment	0.13	0.16
Control	0.14	0.13
P-value	0.88	0.61
Observations	147	163

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The table summarizes whether N or more mental health symptoms are observed (to any degree) at the person-type level by treatment arm. Mean values are reported. The table also reports the P-value associated with a t-test of the null hypothesis that each pre-treatment outcome does not differ by treatment status.

Table A6: Baseline chores performance outcomes by person type and treatment arm

	Primary Male	Primary Female	Sons	Daugh -ters	Other Adult Males	Other Adult Females
Domestic Work						
Share of Person-chores						
Treatment	0.09	0.80	0.04	0.16	0.03	0.19
	(0.16)	(0.24)	(0.14)	(0.23)	(0.07)	(0.24)
Control	0.11	0.79	0.05	0.14	0.04	0.05
	(0.18)	(0.25)	(0.12)	(0.26)	(0.08)	(0.11)
P-value	0.32	0.82	0.93	0.78	0.79	0.26
Share of Person-chore Minutes						
Treatment	0.08	0.84	0.03	0.11	0.06	0.12
	(0.16)	(0.22)	(0.10)	(0.19)	(0.14)	(0.15)
Control	0.09	0.83	0.05	0.10	0.04	0.06
	(0.19)	(0.24)	(0.13)	(0.21)	(0.07)	(0.15)
P-value	0.67	0.80	0.55	0.88	0.77	0.56
Childcare						
Share of Person-chores						
Treatment	0.19	0.73	0.05	0.08	0.00	0.10
	(0.20)	(0.26)	(0.18)	(0.16)		(0.16)
Control	0.29	0.64	0.01	0.08	0.00	0.10
	(0.28)	(0.29)	(0.05)	(0.17)		(0.22)
P-value	0.01***	0.10	0.20	0.98		0.97
Share of Person-chore Minutes						
Treatment	0.14	0.78	0.04	0.08	0.00	0.09
	(0.18)	(0.25)	(0.17)	(0.17)		(0.15)
Control	0.27	0.67	0.01	0.08	0.00	0.10
	(0.30)	(0.30)	(0.08)	(0.18)		(0.22)
P-value	0.00***	0.04**	0.32	0.95		0.92
Observations	163	163	76	75	10	12

Note: * $p < 0.10$, ** $p < 0.05$,*** $p < 0.01$. The table summarizes outcomes described in Section 3.5 at the person-type level by treatment arm. Mean values are reported, and, for all continuous variables, standard deviations are reported in parentheses below the mean. The table also reports a t-test of the null hypothesis that the pre-treatment outcome does not differ by treatment status. For each person type, we observe chores outcomes for all 163 households for whom we have complete interviews with PFs.

Table A7: Person–Chore baseline outcome (whether chore was performed) by treatment arm

	Primary Male	Primary Female	Sons	Daughters	Other Adult Males	Other Adult Females
Fuelwood:						
Treatment	0.14	0.52	0.09	0.11	0.17	0.00
Control	0.20	0.60	0.07	0.11	0.00	0.00
P-value	0.32	0.28	0.85	0.97	0.35	
Water:						
Treatment	0.11	0.59	0.09	0.11	0.00	0.29
Control	0.19	0.60	0.10	0.08	0.25	0.00
P-value	0.20	0.89	0.87	0.73	0.33	0.19
Cooking:						
Treatment	0.04	0.99	0.00	0.05	0.00	0.00
Control	0.05	0.94	0.00	0.05	0.00	0.20
P-value	0.67	0.04**		0.97		0.34
Laundry:						
Treatment	0.10	0.23	0.03	0.00	0.00	0.00
Control	0.08	0.35	0.02	0.05	0.00	0.00
P-value	0.62	0.12	0.91	0.17		
Grains:						
Treatment	0.01	0.33	0.00	0.00	0.00	0.00
Control	0.04	0.20	0.00	0.00	0.00	0.00
P-value	0.31	0.08*				
Cleaning:						
Treatment	0.07	0.81	0.03	0.24	0.00	0.43
Control	0.08	0.84	0.00	0.22	0.00	0.00
P-value	0.95	0.63	0.33	0.86		0.04**
Dishes:						
Treatment	0.05	0.84	0.03	0.26	0.00	0.43
Control	0.09	0.80	0.02	0.22	0.00	0.00
P-value	0.29	0.45	0.91	0.70		0.10*
IC Care:						
Treatment	0.46	0.94	0.11	0.24	0.00	0.29
Control	0.51	0.91	0.02	0.22	0.00	0.20
P-value	0.54	0.53	0.16	0.83		0.73
OC Care:						
Treatment	0.14	0.71	0.00	0.03	0.00	0.14
Control	0.26	0.63	0.00	0.05	0.00	0.00
P-value	0.08*	0.25		0.55		0.37
Observations	1467	1467	684	675	90	108
Number of households	163	163	76	75	10	12

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The table summarizes performance of each chore (binary) by each person type, for households assigned to each treatment arm. The table also reports a t-test of the null hypothesis that the pre-treatment outcome does not differ by treatment status. When a chore is not completed by a given person type the participation rates are 0 and no p-value is reported, resulting in blank spaces. The number of households varies across columns when a given person type does not exist in all households.

Table A8: Person–Chore baseline outcome (total minutes) by treatment arm

	Primary Male	Primary Female	Sons	Daughters	Other Adult Males	Other Adult Females
Fuelwood:						
Treatment	16.75	38.92	6.00	5.79	20.00	0.00
	(50.47)	(53.19)	(22.78)	(21.89)	(48.99)	
Control	10.06	48.94	11.71	11.35	0.00	0.00
	(23.74)	(58.33)	(42.89)	(36.98)		
P-value	0.25	0.22	0.48	0.44	0.35	
Water:						
Treatment	5.18	34.30	6.86	5.79	0.00	17.14
	(18.09)	(39.33)	(24.23)	(19.54)		(29.28)
Control	8.13	34.75	8.78	4.32	15.00	0.00
	(20.07)	(40.35)	(30.02)	(20.21)	(30.00)	
P-value	0.37	0.95	0.77	0.75	0.33	0.19
Cooking:						
Treatment	1.63	125.12	0.00	3.16	0.00	0.00
	(9.37)	(69.13)		(13.58)		
Control	2.38	123.38	0.00	6.49	0.00	12.00
	(10.82)	(77.16)		(27.51)		(26.83)
P-value	0.65	0.87		0.39		0.34
Laundry:						
Treatment	13.13	30.48	1.71	0.00	0.00	0.00
	(44.31)	(64.73)	(10.14)			
Control	8.50	53.00	2.93	4.86	0.00	0.00
	(35.40)	(82.13)	(18.74)	(21.81)		
P-value	0.43	0.05*	0.72	0.19		
Grains:						
Treatment	0.36	34.52	0.00	0.00	0.00	0.00
	(3.29)	(60.54)				
Control	4.13	19.56	0.00	0.00	0.00	0.00
	(21.21)	(52.00)				
P-value	0.12	0.12				
Cleaning:						
Treatment	1.18	31.05	0.57	11.89	0.00	14.29
	(6.82)	(37.31)	(3.38)	(30.61)		(22.99)
Control	1.60	29.40	0.00	10.81	0.00	0.00
	(7.57)	(27.08)		(26.39)		
P-value	0.72	0.74	0.33	0.89		0.07*
Dishes:						
Treatment	0.60	23.52	0.43	11.45	0.00	22.14
	(3.53)	(23.32)	(2.54)	(28.85)		(44.52)
Control	3.31	29.66	0.73	10.27	0.00	0.00
	(15.47)	(32.67)	(4.69)	(25.33)		
P-value	0.13	0.19	0.72	0.87		0.25
IC Care:						
Treatment	37.59	133.43	6.86	17.89	0.00	25.71
	(56.26)	(91.01)	(19.37)	(36.55)		(47.21)
Control	61.73	128.63	7.32	20.54	0.00	24.00
	(92.51)	(97.86)	(46.85)	(52.86)		(53.67)
P-value	0.04**	0.74	0.96	0.80		0.95
OC Care:						
Treatment	8.95	61.00	0.00	1.58	0.00	8.57
	(30.79)	(64.73)		(9.73)		(22.68)
Control	15.63	48.50	0.00	3.24	0.00	0.00
	(32.37)	(51.07)		(13.75)		
P-value	0.18	0.19		0.55		0.37
Observations	1467	1467	684	675	90	108
Number of households	163	163	76	75	10	12

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The table summarizes total minutes performing each chore for each person type, by treatment arm. Mean values and standard deviations are reported. The table also reports a t-test of the null hypothesis that the pre-treatment outcome does not differ by treatment status.

Table A9: Person–Chore baseline outcome (share of household chore minutes) by treatment arm

	Primary Male	Primary Female	Sons	Daughters	Other Adult Males	Other Adult Females
Fuelwood:						
Treatment	0.02	0.06	0.01	0.01	0.03	0.00
	(0.07)	(0.08)	(0.03)	(0.03)	(0.08)	
Control	0.01	0.07	0.02	0.01	0.00	0.00
	(0.03)	(0.08)	(0.06)	(0.05)		
P-value	0.20	0.49	0.54	0.45	0.35	
Water:						
Treatment	0.01	0.06	0.01	0.01	0.00	0.03
	(0.04)	(0.07)	(0.04)	(0.03)		(0.05)
Control	0.02	0.05	0.01	0.01	0.02	0.00
	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	
P-value	0.27	0.33	1.00	1.00	0.33	0.19
Cooking:						
Treatment	0.00	0.20	0.00	0.02	0.00	0.00
	(0.03)	(0.11)		(0.08)		
Control	0.00	0.19	0.00	0.01	0.00	0.02
	(0.02)	(0.11)		(0.04)		(0.04)
P-value	0.98	0.42		0.43		0.34
Laundry:						
Treatment	0.02	0.04	0.00	0.00	0.00	0.00
	(0.05)	(0.08)	(0.02)			
Control	0.01	0.07	0.00	0.01	0.00	0.00
	(0.05)	(0.12)	(0.03)	(0.02)		
P-value	0.39	0.04**	0.74	0.18		
Grains:						
Treatment	0.00	0.05	0.00	0.00	0.00	0.00
	(0.00)	(0.09)				
Control	0.00	0.03	0.00	0.00	0.00	0.00
	(0.02)	(0.06)				
P-value	0.13	0.04**				
Cleaning:						
Treatment	0.00	0.05	0.00	0.02	0.00	0.02
	(0.01)	(0.06)	(0.01)	(0.05)		(0.03)
Control	0.00	0.04	0.00	0.01	0.00	0.00
	(0.01)	(0.04)		(0.03)		
P-value	0.89	0.43	0.33	0.61		0.06*
Dishes:						
Treatment	0.00	0.04	0.00	0.02	0.00	0.03
	(0.01)	(0.04)	(0.00)	(0.04)		(0.05)
Control	0.01	0.04	0.00	0.01	0.00	0.00
	(0.04)	(0.05)	(0.01)	(0.03)		
P-value	0.16	0.63	0.86	0.75		0.21
IC Care:						
Treatment	0.06	0.21	0.01	0.03	0.00	0.03
	(0.08)	(0.12)	(0.04)	(0.06)		(0.05)
Control	0.09	0.19	0.01	0.03	0.00	0.05
	(0.13)	(0.14)	(0.04)	(0.07)		(0.11)
P-value	0.05*	0.39	0.42	0.97		0.72
OC Care:						
Treatment	0.01	0.09	0.00	0.00	0.00	0.01
	(0.03)	(0.08)		(0.01)		(0.03)
Control	0.02	0.07	0.00	0.01	0.00	0.00
	(0.05)	(0.08)		(0.02)		
P-value	0.02**	0.29		0.52		0.37
Observations	1467	1467	684	675	90	108
Number of households	163	163	76	75	10	12

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The table summarizes the share of household's person-chores for each chore and person type, for households assigned to each treatment arm. Mean values and standard deviations are reported. The table also reports a t-test of the null hypothesis that the pre-treatment outcome does not differ by treatment status.

Main Results Tables

Table A10: Treatment effects on mental health outcomes by person type

	PHQ-9 Score		**Number of Symptoms**		**Mild+ Symptoms**	
	(1)	(2)	(1)	(2)	(1)	(2)
Treat X PM	-1.008**	-0.910*	-0.769**	-0.703*	-0.086	-0.089
	(0.473)	(0.480)	(0.346)	(0.350)	(0.063)	(0.063)
Treat X PF	-0.977**	-0.805*	-0.774**	-0.705**	-0.088	-0.085
	(0.461)	(0.457)	(0.338)	(0.347)	(0.071)	(0.073)
Covariates	No	Yes	No	Yes	No	Yes
Num. Covariates	–	3	–	3	–	2
Obs.	310	310	310	310	310	310
Control Means						
PM	4.20	4.20	3.77	3.77	0.41	0.41
PF	3.92	3.92	3.48	3.48	0.41	0.41
RI P-value						
PM	0.040	0.060	0.040	0.060	0.25	0.24
PF	0.040	0.090	0.040	0.050	0.25	0.27
One sided P-value						
PM	0.020	0.032	0.016	0.025	0.09	0.08
PF	0.020	0.042	0.013	0.024	0.11	0.13

Note. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered by village, shown in parentheses. All regressions report the treatment effect for both PM and PF using the specification in Equation 1. All columns include village-level fixed effects, endline interview time-of-day fixed effects and the baseline value of the outcome variable. For each outcome, column 1 uses no additional covariates, while column 2 reports results from a regression that also includes a vector of covariates selected using PDS Lasso from the list of baseline characteristics in Table A3. Across all of the results reported in this table, the covariates selected include: PFs main activity is daily laborer in last 7 days; PFs main activity is none/other non-market activity in last 7 days; PMs main activity is unpaid housework in last 7 days; PFs main activity is agriculture in last 7 days. We also report the baseline period control group means for each outcome, one-sided p-value and the randomization inference p-value for a t-test that each treatment effect equals zero, as explained in Section 4.

Table A11: Treatment effects on mental health outcomes by person type

	PHQ-9 Score		**Number of Symptoms**		**Mild+ Symptoms**	
	(1)	(2)	(1)	(2)	(1)	(2)
PM X Mild+ symptoms	-0.303	-0.139	-0.191	-0.111	0.063	0.057
	(0.663)	(0.660)	(0.450)	(0.445)	(0.102)	(0.106)
PF X Mild+ symptoms	-0.299	-0.119	-0.308	-0.253	0.037	0.047
	(0.699)	(0.676)	(0.512)	(0.521)	(0.108)	(0.108)
PM X Minimal depressive symptoms	-1.492***	-1.453***	-1.180***	-1.135***	-0.191**	-0.194**
	(0.497)	(0.502)	(0.408)	(0.411)	(0.076)	(0.076)
PF X Minimal depressive symptoms	-1.524***	-1.358**	-1.130***	-1.055**	-0.187**	-0.188**
	(0.536)	(0.545)	(0.395)	(0.394)	(0.086)	(0.087)
Covariates	No	Yes	No	Yes	No	Yes
Num. Covariates	–	3	–	3	–	2
Obs.	310	310	310	310	310	310

Note. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered by village, shown in parentheses. Each column reports four coefficients from a regression interacting assigned treatment with a binary variable indicating the person type (PM or PF) and a binary variable indicating the respondent's baseline PHQ-9 score corresponded with mild or worse symptoms (PHQ-9 score > 4). All columns include village-level fixed effects, endline interview time-of-day fixed effects and the baseline value of the outcome variable. For each outcome, column 1 uses no additional covariates, while column 2 reports results from a regression that also includes a vector of covariates selected using PDS Lasso from the list of baseline characteristics in Table A3. Across all of the results reported in this table, the covariates selected include: PFs main activity is daily laborer in last 7 days; PFs main activity is none/other non-market activity in last 7 days; PMs main activity is unpaid housework in last 7 days; PFs main activity is agriculture in last 7 days.

Table A12: Treatment effects on the performance of any chore and the number of chores performed by person type

	Domestic Work:				**Childcare:**			
	Any		**Number**		**Any**		**Number**	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Treat X PM	0.168**	0.161*	0.326**	0.350**	0.031	0.036	0.136	0.151*
	(0.082)	(0.084)	(0.136)	(0.145)	(0.064)	(0.066)	(0.086)	(0.089)
Treat X PF	-0.047	-0.054	-0.320	-0.296	0.039	0.044	0.013	0.030
	(0.036)	(0.037)	(0.238)	(0.235)	(0.060)	(0.061)	(0.099)	(0.097)
Treat X CM	0.132	0.135	0.161	0.192	0.037	0.046	0.045	0.067
	(0.087)	(0.085)	(0.145)	(0.145)	(0.076)	(0.073)	(0.095)	(0.094)
Treat X CF	-0.049	-0.062	-0.078	-0.064	-0.135	-0.128	-0.159	-0.148
	(0.095)	(0.093)	(0.214)	(0.215)	(0.115)	(0.117)	(0.149)	(0.152)
Treat X AM	-0.211	-0.195	-0.373*	-0.366*	0.084	0.088	-0.024	-0.017
	(0.199)	(0.195)	(0.205)	(0.202)	(0.126)	(0.126)	(0.135)	(0.134)
Treat X AF	0.452***	0.450***	1.195**	1.190**	0.453**	0.450**	0.530**	0.525**
	(0.140)	(0.146)	(0.582)	(0.579)	(0.184)	(0.189)	(0.231)	(0.235)
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
Num. Covariates	–	4	–	3	–	3	–	3
Obs.	499	499	499	499	499	499	499	499
Control Means								
PM	0.40	0.40	0.71	0.71	0.59	0.59	0.78	0.78
PF	0.97	0.97	4.33	4.33	0.94	0.94	1.54	1.54
CM	0.17	0.17	0.22	0.22	0.024	0.024	0.024	0.024
CF	0.30	0.30	0.73	0.73	0.22	0.22	0.27	0.27
AM	0.25	0.25	0.25	0.25	0	0	0	0
AF	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
RI P-value								
PM	0.030	0.040	0.020	0.020	0.69	0.65	0.19	0.15
PF	0.050	0.040	0.17	0.20	0.50	0.44	0.90	0.78
CM	0.16	0.14	0.25	0.16	0.64	0.54	0.66	0.49
CF	0.63	0.54	0.74	0.78	0.21	0.24	0.27	0.30
AM	0.32	0.37	0.13	0.14	0.63	0.62	0.88	0.90
AF	0.070	0.080	0.16	0.17	0.17	0.18	0.29	0.30

Note. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered by village, shown in parentheses. All regressions report the treatment effect for each person type using the specification in Equation 1. All columns include village-level fixed effects, endline interview time-of-day fixed effects and the baseline value of the outcome variable. For each outcome, column 1 uses no additional covariates, while column 2 reports results from a regression that also includes a vector of covariates selected using PDS Lasso from the list of baseline characteristics in Table A3. Across all of the results reported in this table, the covariates selected include: PMs main activity is unpaid housework in last 7 days; PFs main activity is agriculture in last 7 days; PFs main activity is daily laborer in last 7 days; PFs main activity is none/other non-market activity in last 7 days. We also report the baseline period control group means for each outcome and the randomization inference p-value for a t-test that each treatment effect equals zero, as explained in Section 4.

Table A13: Treatment Effects on Time Performing Chores by person type

	Domestic Work: Person-chore Minutes		**Childcare: Person-chore Minutes**	
	(1)	(2)	(1)	(2)
Treat X PM	13.593	20.430*	-6.316	-5.414
	(10.870)	(10.775)	(15.229)	(15.466)
Treat X PF	-14.865	-8.450	-33.743	-32.671
	(23.616)	(23.211)	(33.392)	(33.877)
Treat X CM	6.952	19.898*	5.866	7.215
	(11.991)	(10.193)	(11.281)	(11.911)
Treat X CF	-10.104	-1.076	-14.865	-14.611
	(21.784)	(18.050)	(18.596)	(18.773)
Treat X AM	-37.585**	-33.882*	13.820	13.994
	(18.543)	(18.864)	(48.555)	(48.739)
Treat X AF	111.599*	107.747*	28.916	28.731
	(56.957)	(54.891)	(44.013)	(44.249)
Covariates	No	Yes	No	Yes
Num. Covariates	–	3	–	3
Obs.	499	499	499	499
Control Means				
PM	38.1	38.1	77.3	77.3
PF	338.7	338.7	177.1	177.1
CM	24.1	24.1	7.32	7.32
CF	48.1	48.1	23.8	23.8
AM	15	15	0	0
AF	12	12	24	24
RI P-value				
PM	0.24	0.060	0.69	0.75
PF	0.56	0.75	0.33	0.35
CM	0.51	0.030	0.64	0.59
CF	0.66	0.96	0.43	0.44
AM	0.12	0.16	0.76	0.76
AF	0.25	0.28	0.65	0.65

Note. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered by village, shown in parentheses. All regressions report the treatment effect for each person type using the specification in Equation 1. All columns include village-level fixed effects, endline interview time-of-day fixed effects and the baseline value of the outcome variable. For each outcome, column 1 uses no additional covariates, while column 2 reports results from a regression that also includes a vector of covariates selected using PDS Lasso from the list of baseline characteristics in Table A3. Across all of the results reported in this table, the covariates selected include: PMs main activity is unpaid housework in last 7 days; PFs main activity is daily laborer in last 7 days; PFs main activity is none/other non-market activity in last 7 days. We also report the baseline period control group means for each outcome and the randomization inference p-value for a t-test that each treatment effect equals zero, as explained in Section 4.

Table A14: Treatment effects on performing chores by PM and chore type

	Any chore		**Total person-chore minutes**		**Share of person-chore minutes**	
	(1)	(2)	(1)	(2)	(1)	(2)
Treat X PM X Fuelwood	0.086	0.090	4.573	5.463	0.017	0.017
	(0.067)	(0.068)	(6.849)	(6.665)	(0.013)	(0.013)
Treat X PM X Water	0.070	0.074	4.612	5.481	0.008	0.008
	(0.063)	(0.064)	(3.819)	(3.922)	(0.012)	(0.012)
Treat X PM X Cooking	0.018	0.022	-0.451	0.423	0.000	0.000
	(0.023)	(0.025)	(1.727)	(1.860)	(0.003)	(0.003)
Treat X PM X Laundry	0.082*	0.087*	10.065*	10.950**	0.012*	0.012*
	(0.044)	(0.044)	(5.098)	(5.110)	(0.007)	(0.007)
Treat X PM X Grains	-0.016	-0.012	-3.092	-2.223	-0.003	-0.003
	(0.017)	(0.017)	(2.625)	(2.694)	(0.003)	(0.003)
Treat X PM X Cleaning	0.037	0.041	0.635	1.510	0.002	0.001
	(0.024)	(0.025)	(0.719)	(0.973)	(0.001)	(0.001)
Treat X PM X Dishes	0.041	0.045	1.325	2.195*	0.002	0.002
	(0.038)	(0.038)	(0.914)	(1.167)	(0.001)	(0.001)
Treat X PM X IC care	0.030	0.034	-8.709	-7.892	-0.000	-0.000
	(0.064)	(0.065)	(10.833)	(10.962)	(0.020)	(0.020)
Treat X PM X OC care	0.070	0.074	-2.342	-1.478	0.001	0.001
	(0.046)	(0.046)	(6.943)	(6.991)	(0.006)	(0.006)
Covariates	No	Yes	No	Yes	No	Yes
Num. Covariates	–	3	–	3	–	4
Obs.	4491	4491	4491	4491	4491	4491

Note. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered by village, shown in parentheses. All regressions report the treatment effect for primary male using the adapted model ofspecification in Equation 1 (see, Footnote 30). All columns include village-level fixed effects, endline interview time-of-day fixed effects and the baseline value of the outcome variable. For each outcome, column 1 uses no additional covariates, while column 2 reports results from a regression that also includes a vector of covariates selected using PDS Lasso from the list of baseline characteristics in Table A3. Across all of the results reported in this table, the covariates selected include: PMs main activity is unpaid housework in last 7 days; PFs main activity is daily laborer in last 7 days; PFs main activity is none/other non-market activity in last 7 days; No. of daughters.

Table A15: Treatment effects on performing chores by PF and chore type

	Any chore		**Total person-chore minutes**		**Share of person-chore minutes**	
	(1)	(2)	(1)	(2)	(1)	(2)
Treat X PF X Fuelwood	-0.100	-0.096	-8.122	-7.264	-0.015	-0.015
	(0.086)	(0.086)	(11.192)	(11.223)	(0.017)	(0.017)
Treat X PF X Water	0.040	0.044	5.673	6.545	0.011	0.011
	(0.070)	(0.070)	(5.464)	(5.451)	(0.009)	(0.009)
Treat X PF X Cooking	-0.060	-0.056	-5.588	-4.726	-0.007	-0.007
	(0.043)	(0.042)	(10.720)	(10.744)	(0.020)	(0.020)
Treat X PF X Laundry	0.065	0.069	11.572	12.404	0.018	0.018
	(0.076)	(0.075)	(11.169)	(11.150)	(0.020)	(0.020)
Treat X PF X Grains	-0.071	-0.067	-11.736	-10.830	-0.009	-0.009
	(0.080)	(0.080)	(10.406)	(10.173)	(0.019)	(0.019)
Treat X PF X Cleaning	-0.072	-0.068	1.814	2.689	0.004	0.004
	(0.055)	(0.054)	(2.673)	(2.661)	(0.005)	(0.005)
Treat X PF X Dishes	-0.103	-0.099	-6.693*	-5.830*	-0.013	-0.013
	(0.066)	(0.065)	(3.429)	(3.412)	(0.008)	(0.008)
Treat X PF X IC care	0.050	0.055	-22.847	-21.969	-0.017	-0.017
	(0.064)	(0.063)	(20.660)	(20.741)	(0.029)	(0.029)
Treat X PF X OC care	-0.046	-0.042	-14.240	-13.341	-0.008	-0.008
	(0.066)	(0.066)	(17.359)	(17.371)	(0.013)	(0.013)
Covariates	No	Yes	No	Yes	No	Yes
Num. Covariates	–	3	–	3	–	4
Obs.	4491	4491	4491	4491	4491	4491

Note. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered by village, shown in parentheses. All regressions report the treatment effect for primary female using the adapted model ofspecification in Equation 1 (see, Footnote 30). All columns include village-level fixed effects, endline interview time-of-day fixed effects and the baseline value of the outcome variable. For each outcome, column 1 uses no additional covariates, while column 2 reports results from a regression that also includes a vector of covariates selected using PDS Lasso from the list of baseline characteristics in Table A3. Across all of the results reported in this table, the covariates selected include: PMs main activity is unpaid housework in last 7 days; PFs main activity is daily laborer in last 7 days; PFs main activity is none/other non-market activity in last 7 days; No. of daughters.

Table A16: Treatment effects on performing chores by CM and chore type

	Any chore		**Total person-chore minutes**		**Share of person-chore minutes**	
	(1)	(2)	(1)	(2)	(1)	(2)
Treat X CM X Fuelwood	0.014	0.020	-2.704	-1.111	0.002	0.002
	(0.054)	(0.053)	(7.528)	(6.443)	(0.007)	(0.007)
Treat X CM X Water	0.069*	0.074*	2.167	3.767*	0.002*	0.002
	(0.040)	(0.040)	(1.780)	(2.086)	(0.001)	(0.001)
Treat X CM X Cooking	0.017	0.023	1.464	3.069	0.000	0.000
	(0.027)	(0.027)	(2.001)	(2.277)	(0.001)	(0.001)
Treat X CM X Laundry	0.018	0.023	1.664	3.265	0.002	0.001
	(0.027)	(0.028)	(1.861)	(2.174)	(0.002)	(0.002)
Treat X CM X Grains	-0.029	-0.024	-0.570	1.035	-0.002	-0.003
	(0.024)	(0.025)	(2.175)	(2.453)	(0.002)	(0.002)
Treat X CM X Cleaning	0.005	0.010	0.311	1.916	0.000	-0.000
	(0.059)	(0.060)	(2.295)	(2.569)	(0.002)	(0.002)
Treat X CM X Dishes	0.095	0.100	4.949	6.553*	0.012*	0.012*
	(0.066)	(0.066)	(3.420)	(3.621)	(0.006)	(0.006)
Treat X CM X IC care	0.036	0.041	5.708	7.313	0.008	0.008
	(0.075)	(0.074)	(5.091)	(4.961)	(0.008)	(0.008)
Treat X CM X OC care	-0.004	0.001	-1.118	0.486	-0.002	-0.003
	(0.035)	(0.036)	(3.097)	(3.318)	(0.004)	(0.004)
Covariates	No	Yes	No	Yes	No	Yes
Num. Covariates	–	3	–	3	–	4
Obs.	4491	4491	4491	4491	4491	4491

Note. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered by village, shown in parentheses. All regressions report the treatment effect for male children using the adapted model ofspecification in Equation 1 (see, Footnote 30). All columns include village-level fixed effects, endline interview time-of-day fixed effects and the baseline value of the outcome variable. For each outcome, column 1 uses no additional covariates, while column 2 reports results from a regression that also includes a vector of covariates selected using PDS Lasso from the list of baseline characteristics in Table A3. Across all of the results reported in this table, the covariates selected include: PMs main activity is unpaid housework in last 7 days; PFs main activity is daily laborer in last 7 days; PFs main activity is none/other non-market activity in last 7 days; No. of daughters.

Table A17: Treatment effects on performing chores by CF and chore type

	Any chore		**Total person-chore minutes**		**Share of person-chore minutes**	
	(1)	(2)	(1)	(2)	(1)	(2)
Treat X CF X Fuelwood	-0.031	-0.028	-4.234	-3.182	-0.008	-0.008
	(0.070)	(0.071)	(7.340)	(7.522)	(0.010)	(0.010)
Treat X CF X Water	-0.126*	-0.123	-4.467	-3.404	-0.008	-0.008
	(0.075)	(0.075)	(6.105)	(6.296)	(0.008)	(0.008)
Treat X CF X Cooking	0.051	0.054	1.256	2.311	0.000	0.001
	(0.034)	(0.034)	(3.354)	(2.707)	(0.005)	(0.005)
Treat X CF X Laundry	0.029	0.032	8.392	9.444	0.014	0.014
	(0.058)	(0.058)	(5.562)	(5.631)	(0.010)	(0.010)
Treat X CF X Grains	-0.025	-0.022	-10.832	-9.771	-0.006	-0.006
	(0.025)	(0.022)	(11.604)	(10.604)	(0.006)	(0.005)
Treat X CF X Cleaning	0.071	0.074	-2.446	-1.381	-0.004	-0.004
	(0.081)	(0.081)	(5.504)	(5.163)	(0.009)	(0.009)
Treat X CF X Dishes	-0.050	-0.047	0.458	1.521	0.001	0.001
	(0.095)	(0.094)	(4.453)	(4.645)	(0.007)	(0.007)
Treat X CF X IC care	-0.095	-0.092	-7.741	-6.682	-0.016	-0.016
	(0.107)	(0.107)	(11.828)	(12.037)	(0.021)	(0.022)
Treat X CF X OC care	-0.067	-0.064	-6.312	-5.255	-0.016	-0.016
	(0.070)	(0.070)	(7.657)	(7.796)	(0.015)	(0.015)
Covariates	No	Yes	No	Yes	No	Yes
Num. Covariates	–	3	–	3	–	4
Obs.	4491	4491	4491	4491	4491	4491

Note. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered by village, shown in parentheses. All regressions report the treatment effect for female children using the adapted model ofspecification in Equation 1 (see, Footnote 30). All columns include village-level fixed effects, endline interview time-of-day fixed effects and the baseline value of the outcome variable. For each outcome, column 1 uses no additional covariates, while column 2 reports results from a regression that also includes a vector of covariates selected using PDS Lasso from the list of baseline characteristics in Table A3. Across all of the results reported in this table, the covariates selected include: PMs main activity is unpaid housework in last 7 days; PFs main activity is daily laborer in last 7 days; PFs main activity is none/other non-market activity in last 7 days; No. of daughters.

Table A18: Treatment effects on performing chores by AM and chore type

	Any chore		**Total person-chore minutes**		**Share of person-chore minutes**	
	(1)	(2)	(1)	(2)	(1)	(2)
Treat X AM X Fuelwood	-0.053	-0.052	-4.251	-3.740	-0.003	-0.002
	(0.036)	(0.036)	(5.447)	(5.478)	(0.004)	(0.004)
Treat X AM X Water	0.021	0.022	-0.994	-0.543	0.004	0.005
	(0.043)	(0.042)	(6.742)	(6.714)	(0.003)	(0.003)
Treat X AM X Cooking	-0.025	-0.024	-2.478	-1.999	0.001	0.002
	(0.021)	(0.021)	(5.571)	(5.592)	(0.001)	(0.002)
Treat X AM X Laundry	-0.025	-0.024	-2.478	-1.999	0.001	0.002
	(0.021)	(0.021)	(5.571)	(5.592)	(0.001)	(0.002)
Treat X AM X Grains	-0.025	-0.024	-2.478	-1.999	0.001	0.002
	(0.021)	(0.021)	(5.571)	(5.592)	(0.001)	(0.002)
Treat X AM X Cleaning	-0.219	-0.217	-6.349	-5.870	-0.005	-0.003
	(0.185)	(0.184)	(6.828)	(6.721)	(0.005)	(0.006)
Treat X AM X Dishes	-0.025	-0.024	-2.478	-1.999	0.001	0.002
	(0.021)	(0.021)	(5.571)	(5.592)	(0.001)	(0.002)
Treat X AM X IC care	-0.025	-0.024	-2.478	-1.999	0.001	0.002
	(0.021)	(0.021)	(5.571)	(5.592)	(0.001)	(0.002)
Treat X AM X OC care	-0.025	-0.024	-2.478	-1.999	0.001	0.002
	(0.021)	(0.021)	(5.571)	(5.592)	(0.001)	(0.002)
Covariates	No	Yes	No	Yes	No	Yes
Num. Covariates	–	3	–	3	–	4
Obs.	4491	4491	4491	4491	4491	4491

Note. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered by village, shown in parentheses. All regressions report the treatment effect for adult males using the adapted model ofspecification in Equation 1 (see, Footnote 30). All columns include village-level fixed effects, endline interview time-of-day fixed effects and the baseline value of the outcome variable. For each outcome, column 1 uses no additional covariates, while column 2 reports results from a regression that also includes a vector of covariates selected using PDS Lasso from the list of baseline characteristics in Table A3. Across all of the results reported in this table, the covariates selected include: PMs main activity is unpaid housework in last 7 days; PFs main activity is daily laborer in last 7 days; PFs main activity is none/other non-market activity in last 7 days; No. of daughters.

Table A19: Treatment effects on performing chores by AF and chore type

	Any chore		**Total person-chore minutes**		**Share of person-chore minutes**	
	(1)	(2)	(1)	(2)	(1)	(2)
Treat X AF X Fuelwood	0.041**	0.039**	1.250	0.758	0.002	0.002
	(0.016)	(0.016)	(2.692)	(2.791)	(0.002)	(0.002)
Treat X AF X Water	0.127	0.126	7.942	7.482	0.006	0.006
	(0.143)	(0.141)	(8.256)	(7.919)	(0.010)	(0.010)
Treat X AF X Cooking	0.355**	0.353**	35.977*	35.465*	0.048**	0.048**
	(0.147)	(0.147)	(19.888)	(19.636)	(0.024)	(0.023)
Treat X AF X Laundry	0.181	0.179	26.484	25.992	0.025	0.025
	(0.121)	(0.120)	(20.878)	(20.523)	(0.021)	(0.021)
Treat X AF X Grains	0.041**	0.039**	1.250	0.758	0.002	0.002
	(0.016)	(0.016)	(2.692)	(2.791)	(0.002)	(0.002)
Treat X AF X Cleaning	0.240*	0.239*	12.434*	11.969*	0.013	0.013
	(0.140)	(0.138)	(6.871)	(6.501)	(0.008)	(0.008)
Treat X AF X Dishes	0.358**	0.357**	23.345*	22.888*	0.023*	0.023*
	(0.155)	(0.154)	(12.912)	(13.321)	(0.012)	(0.013)
Treat X AF X IC care	0.388*	0.387*	16.400	15.903	-0.000	0.000
	(0.198)	(0.200)	(36.312)	(36.856)	(0.069)	(0.070)
Treat X AF X OC care	0.154	0.153	17.213	16.737	0.016	0.016
	(0.131)	(0.130)	(14.448)	(14.101)	(0.015)	(0.015)
Covariates	No	Yes	No	Yes	No	Yes
Num. Covariates	–	3	–	3	–	4
Obs.	4491	4491	4491	4491	4491	4491

Note. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered by village, shown in parentheses. All regressions report the treatment effect for adult females using the adapted model ofspecification in Equation 1 (see, Footnote 30). All columns include village-level fixed effects, endline interview time-of-day fixed effects and the baseline value of the outcome variable. For each outcome, column 1 uses no additional covariates, while column 2 reports results from a regression that also includes a vector of covariates selected using PDS Lasso from the list of baseline characteristics in Table A3. Across all of the results reported in this table, the covariates selected include: PMs main activity is unpaid housework in last 7 days; PFs main activity is daily laborer in last 7 days; PFs main activity is none/other non-market activity in last 7 days; No. of daughters.

Table A20: Treatment Effects on distribution of person chores and chore minutes by person type

	Domestic Work:				**Childcare:**			
	Person-chore Share		**Minutes Share**		**Person-chore Share**		**Minutes Share**	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Treat X PM	0.071**	0.070**	0.066**	0.065**	0.036	0.036	0.035	0.035
	(0.029)	(0.028)	(0.028)	(0.028)	(0.044)	(0.044)	(0.047)	(0.047)
Treat X PF	-	-	-0.097**	-0.098**	0.019	0.018	0.018	0.018
	0.097***	0.098***						
	(0.035)	(0.036)	(0.037)	(0.038)	(0.057)	(0.057)	(0.061)	(0.061)
Treat X CM	0.054*	0.054*	0.051*	0.050*	0.001	0.002	0.004	0.005
	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.029)	(0.020)	(0.021)
Treat X CF	-0.027	-0.027	-0.021	-0.021	-0.082	-0.082	-0.080	-0.080
	(0.034)	(0.034)	(0.036)	(0.036)	(0.067)	(0.067)	(0.069)	(0.070)
Treat X AM	-0.018	-0.018	-0.011	-0.011	0.013	0.013	0.013	0.013
	(0.040)	(0.040)	(0.026)	(0.026)	(0.012)	(0.012)	(0.012)	(0.012)
Treat X AF	0.158*	0.158*	0.212***	0.212***	0.163	0.163	0.136	0.135
	(0.088)	(0.089)	(0.076)	(0.076)	(0.189)	(0.190)	(0.197)	(0.198)
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
Num. Covar.	–	3	–	3	–	3	–	3
Obs.	499	499	499	499	499	499	499	499
Control Means								
PM	0.11	0.11	0.095	0.095	0.29	0.29	0.27	0.27
PF	0.79	0.79	0.83	0.83	0.64	0.64	0.67	0.67
CM	0.047	0.047	0.050	0.050	0.0081	0.0081	0.012	0.012
CF	0.14	0.14	0.10	0.10	0.081	0.081	0.079	0.079
AM	0.042	0.042	0.036	0.036	0	0	0	0
AF	0.050	0.050	0.065	0.065	0.10	0.10	0.10	0.10
RI P-value								
PM	0.010	0.010	0.030	0.030	0.44	0.45	0.44	0.45
PF	0	0	0	0	0.75	0.76	0.77	0.77
CM	0.050	0.060	0.050	0.050	0.97	0.95	0.84	0.80
CF	0.51	0.51	0.62	0.62	0.20	0.20	0.23	0.23
AM	0.78	0.77	0.78	0.78	0.25	0.25	0.27	0.27
AF	0.22	0.22	0.12	0.12	0.55	0.56	0.62	0.62

Note. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered by village, shown in parentheses. All regressions report the treatment effect for each person type using the specification in Equation 1. All columns include village-level fixed effects, endline interview time-of-day fixed effects and the baseline value of the outcome variable. For each outcome, column 1 uses no additional covariates, while column 2 reports results from a regression that also includes a vector of covariates selected using PDS Lasso from the list of baseline characteristics in Table A3. Across all of the results reported in this table, the covariates selected include: PMs main activity is unpaid housework in last 7 days; PFs main activity is daily laborer in last 7 days; PFs main activity is none/other non-market activity in last 7 days. We also report the baseline period control group means for each outcome and the randomization inference p-value for a t-test that each treatment effect equals zero, as explained in Section 4.

B Additional Intervention Details

In this section we summarize the talking points for the interventions. The full scripts used by enumerators are included below. Table B1 summarizes the objectives, content, and style of each intervention call. Below, we provide the full "scripts" as they appeared in the user interface for enumerators. The placebo call scripts appear at the end of this section. Enumerators were trained to conduct intervention/placebo calls in a manner that resembled a natural conversation, not a survey with questions, but the enumerator used the program interface as a checklist to ensure consistent and complete coverage of the intended content. See Section 3.2 for a description of the project timeline.

Table B1: Summary of Intervention Call Scripts

	Calls 1 & 4	**Calls 2 & 5**	**Calls 3 & 6**	**Call 7**
Objectives	Better appreciate women's time burden; reflect on unequal distribution of chores	Reflect on his upbringing, influence of his male authority figures, and fatherhood	Discuss the social construction of gender-specificity of tasks; see that men can do the same things	Review previous discussions
Content				
Daily tasks of PF	✓			✓
Daily tasks of PM	✓			✓
Distribution of tasks	✓			
Parenting/fatherhood		✓		✓
Gendered chores			✓	✓
Equity/gender roles		✓	✓	
Style				
Direct question	✓	✓	✓	✓
Open-ended qns	✓	✓	✓	✓
Thought exercise		✓	✓	✓
Conveying info			✓	
Agree/disagree			✓	
New assignment	✓	✓	✓	✓

Data from intervention calls. We note that the data recorded by enumerators during these calls was minimal. As the scripts detail, calls sometimes presented each PM with scenarios (e.g., a hungry child, an unclean house while expecting a guest, or a tired spouse)

and asked him to describe his responses (e.g, he would cook, clean, both, or take no action). The invitation to commit to trying a *new* task ("that you usually never do") or behavior was connected to this content in the conversation's flow.

While dynamic analysis of these data for treatment effect mechanisms is not possible, descriptive analysis of the data provides confirmation of men's engagement in the commitment-setting process. Using data from the beginning of intervention calls, we see clear evidence that many PMs can correctly recall two weeks later their commitment to attempt a task not commonly performed: about 65% said they remembered the task they had agreed to, and among those who remembered, 60% were able to correctly name the task they had previously agreed to. This suggests that a substantial share of men not only engaged with the prompting during a given intervention call but also retained the information and recalled it accurately when asked two weeks later.

Follow-through on these tasks not commonly performed is more modest, but still meaningful: about one third of treatment PMs (23 men) reported that they actually performed that same task at the next call. We view the non-completion rate implied by this as a lower bound, however. That is, our theory of change leaves open the possibility that men chose to perform a different task, perhaps after further reflection and/or communication with their spouse about where their contributions might be most helpful. Regrettably, we did not program our intervention scripts to allow enumerators to record this information explicitly.

Our overall results throughout the manuscript are consistent with the idea that men's performance of UDCW increased in ways that are specific to their households rather than on any single activity. Indeed, we do find evidence of continued engagement: fifteen men (nearly 20% of the treatment group) later went on to name a second "new task" in subsequent calls that he would perform—that is, some participants' responses suggest that their set of existing tasks expands because they add additional tasks when asked to commit to a task not performed.

INTERVENTION CALLS

CALL NUMBER 1 OR 4 SCRIPT

Objective: Showing the respondent, the time women spent on housework as compared to men.

Script content:

- **Check if he did the assignment from the previous call (Call 4 only)**
- Ask the respondent to imagine a typical day for man and woman, from the time they wake up to the time they wake up again tomorrow (24-hours day).
- Ask the respondent to think the amount of time he spends on daily tasks: work, sleep, eating, recreation, leisure, housework, etc. Also, ask him to identify the time spent on tasks of child rearing such as education and playing.
- Next, ask the respondent to reflect on the perspective of the *mother* of his children. In other words, how does he think the mother divides her time in a 24-hour day?
    - Then ask the following questions: how do you feel about the current distribution of tasks that you have with your partner? Do you feel this distribution is fair? Why or why not? How do you think your partner may feel about the current time distribution? Which one do you do at home?
    - Who generally performs these activities? Why?
    - What benefit do you gain from being more involved in caring for your children?
    - What about from being more involved in domestic work like cooking (more than just stirring wot), cleaning the house, washing dishes…etc.?
    - What value can redistribution of between you and your partner can have for women and the whole family.
- **Next week's assignment**: perform at least one domestic activity that he usually never does, which household chore will you perform?

CALL NUMBER 2 AND 5 SCRIPT

Objective: reflect upon the influence that fathers or other male authority figures have had on him, discuss how he can take the positive aspects of his fathers' influence, as well address the negative impacts so as not to repeat harmful patterns, encourage him be involved in childcare.

Script content:

- **Check if he did the assignment from the previous call**
  Ask about his experience with fatherhood
    - What kind of parent was your father or father figure? What was your relationship with your father or father figure?
    - How would you like to be a different kind of parent from your father or father figure?
    - What are the challenges of being a father? How can these challenges be addressed?
    - What are the benefits for a child who has a father active in his or her life?
- Read this information about parenthood to the respondent:
    - Parenthood can be a very stressful time for many couples. Babies cry and require constant attention and love from both mother and father. It is important to remember that the best thing to do is to respond to the baby with affection, try to figure out what the baby needs.
    - The world is changing. Before, parental roles were not flexible, men went to work and women took care of domestic affairs, but now, the only thing men cannot do is breastfeed. The acceptance of men as involved caregivers is growing.
    - Gender equality includes sharing domestic responsibilities including childcare.
- Present scenario to respondent
    - Imagine this: it is 6 o'clock in the morning. Your child is crying, and you are not sure why. Your partner is exhausted because she has been breastfeeding all day and night, cooking foods

and engaged in other domestic chores. You just get up to go to farm field. As a father, what can you do?

- *[During the conversation make sure you state that caring for a child can be exciting, but also exhausting and stressful periods in mothers' and fathers' lives. A baby cannot express himself or herself with words, so they cry. Some babies cry a lot, other babies cry less. The most important thing a parent can do is provide warmth through physical affection (e.g. hugging, cradling and rocking the child), feeding, and try to figure out what the baby needs. Men should spend quality time every day with his child.]*

- *[Remember that men are capable of caring for babies and can satisfy all their babies' needs (except Breastfeeding). The roles of motherhood and fatherhood are social constructs, i.e. they are formed by society; with enough practice, any man can be a competent caretaker.]*

- **Next week's assignment:** practice a new way to care for your baby (or older children), what new way of caring for your baby will you do?

CALL NUMBER 3 AND 6 SCRIPT

Objective: show men can do what are thought to be women's activity and it is just the socially constructed gender role that is hindering man from doing the activities, correct the man's thinking about some of the myths hindering men involvement in household chores and childcaring.

Script content:

- **Check if he did the assignment from the previous call**
- Ask the respondent to categorize which of the household chores he thinks are Women's and Men's activities listed in Table A1.
- Ask the respondent to categorize which of the household chores he can and cannot do
- Present scenarios to the respondent and ask him what he would do in these scenarios:
  - *"You just arrived home after an early morning farming activity. It is a market day and your spouse has gone for the market and she has not returned yet. Your kids were playing outside and just followed you when you come into the house. It is almost lunch time; the kids are hungry and there is nothing cooked for lunch. The kids don't know how to cook, and you know how to cook something. What would you do?"*
  - *"It is another day. You and your spouse just arrived home after half day long farming activity. She was weeding while you were doing land preparation on the other part of your land. You know some guests will arrive at your home soon. You are both exhausted but your spouse rushes to cook something for you and the coming guests. As you were seating while waiting, you noticed the house is not clean and presentable to guests and the dishes have not been washed and cleaned. What would you do to fix things?"*
- Read the following statements and ask whether he agrees
  - Women are equal with men.
  - A woman's most important role is to take care of the children and the home.
  - Because a man works hard to provide for the family, he should be given priority at mealtimes.
  - It is a shame for a man to engage in domestic chores.
  - Women are better at communicating with their children.
  - Men are not better at communicating with their children.
  - Growing children with the best care possible from both the mother and the father is good for the future of the country
- **Next week's assignment**: observe how his spouse does one of the domestic chores listed above that he thinks cannot do and try to do it.

CALL NUMBER 7 - SUMMARY SCRIPT

Objective: summarize contents in all of the previous calls.

Script content:

- **Check if he did the assignment from the previous call**
- Summarize content from previous calls

PLACEBO CALLS

Objectives: ask the man what he has been eating and how many times he eats in the last 24 hours.

Script content:

- How many times did you eat in the last 24 hours?
- What did you eat during each time?

C Additional Sampling and Data Protocol Details

C.1 Selection of *kebeles* and *gotts*

The 16 *kebeles* in Amhara were chosen to allow for cost-sharing during baseline fieldwork with a separate evaluation study led by our funding partner, the International Food Policy Research Institute (IFPRI) of Strengthening PSNP Institutions for Resilience II (SPIR II), a multi-faceted graduation program implemented by World Vision supporting PSNP-eligible households with additional livelihoods, nutrition, WASH, and other trainings. The IFPRI evaluation selected three *gotts* (villages) per *kebele* at random.

C.2 Identifying eligible households

Enumerators visited households listed on a census of all PSNP-participating households in a sampled *gott* to identify those eligible for the IFPRI evaluation (with at least one pregnant woman or a child under age 2) or for ours (not eligible for IFPRI and meets other criteria as in Section 3.1). In a few cases, we were not able to find four eligible households in a single *gott*, so we oversampled from a nearby *gott*.

C.3 SPIR II villages and VESA group formation

All in-person baseline surveys for both our study and IFPRI's occurred before any SPIR II activities, of which the first stage was to form VESA groups through which training programming was delivered. All VESAs were formed by the time our phone-based baseline data collection occurred. Since *gotts* are nested within *kebeles*, our main specifications, which include *gott* fixed effects, absorb any *kebele*-level SPIR II treatment status (and other *kebele*-wide programming differences).

Importantly, any SPIR II recruitment/participation for households in village-level savings and loan associations (VSLAs) began before our BCC treatment window but after our baseline data were collected, so VSLA membership is entirely predetermined with respect to our household randomization. Even if VSLA participation involves household-level self-

selection, random assignment for our BCC intervention implies balance within each *gott*, and thus cannot bias our treatment effect estimates reported for this experiment. With respect to external validity, our study population is representative of households in the sampled geographies whose household composition matched our sampling criteria and is not limited households who have previously self-selected into a livelihoods intensification program.

C.4 Timing and staffing of interviews

After each household's in-person interview was complete, survey staff in our call center in Addis Ababa placed weekly calls to the mobile phone that had been provided to each household to maintain contact with them until their randomly assigned start date for the next phase of data collection.

All phone interviews (and intervention / placebo calls) were completed by female enumerators. We chose to keep enumerator gender constant to avoid confounding enumerator gender effects with the intervention and outcome measurement. In our context, there was a very high quality of experienced female enumerators available to hire for the project. Furthermore, we believed it would be less likely to arouse suspicion in our respondent households if frequent calls from outsiders were placed by women rather than men.

To spread the survey and intervention activities across time so that we could continuously employ a small group of highly qualified enumerators, each cluster was assigned to one of eight temporal study "waves" (6 clusters per wave), which determined the start of the second phase of data collection. By randomly assigning clusters to waves, we also varied the timing of the intervention with respect to the seasonality of labor demand (both for economic activities and household chores) in this predominantly agricultural sample.

We separated the PM and PF baseline interviews by a few days, rather than holding them on the same day, to keep phone interviews as short as possible, to simplify the coordination process, and to eliminate possible cross-contamination from one respondent over-hearing the other's interview. A tradeoff with this approach is that the 24 hour reference periods for PF and PMs interviewed within the same household do not overlap. Because the interview order

was randomly assigned orthogonally to treatment status, our staggered timing here poses no issue for identification. Similarly, we randomized the order in which the two primary respondents completed their 7-day high frequency time use interview periods. Since these weeks required the target respondent to carry the phone at all times and for the phone to be well-charged, and because not all households had electricity at home, we placed reminder calls in the days prior to each HFTU week and also included an "off" week between the two HFTU weeks so that the household would have a chance to recharge the phone.

C.5 Survey Instruments

The survey instrument modules used in the "chores interviews" conducted during the *baseline* and *endline* phases are available at the following link.

D Additional Analysis Details

This appendix provides additional detail on attrition, balance, and robustness checks about (i) time-of-day effects in phone-based mental health modules, (ii) potential enumerator effects and enumerator–respondent rapport, and (iii) alternative inference procedures, including randomization inference and sharpened q-values for multiple hypothesis testing.

D.1 Additional Details on Attrition and Balance

Based on a 5% critical value for rejecting the null hypothesis of an equal outcome between the treatment and control group, the outcomes that are not balanced at baseline include the PHQ-9 score being borderline moderate or worse for PFs only (the treatment group is higher), the PM's share in childcare performance, chore minutes, and share in total chore minutes (the control group is larger), and the PFs share in childcare chore minutes (the treatment group is higher). At the person-type chore level, PFs in the treatment group are more likely to participate in cooking, and PMs in the control group perform more minutes of index child care. Note that, in these few cases of imbalance in pre-treatment outcomes, the treatment group PF's outcomes are consistent with a higher chores burden and the control group PM's outcome are consistent with worse mental health and a higher burden. Thus, we would expect imbalances from these differences to attenuate our estimates of the treatment effect.

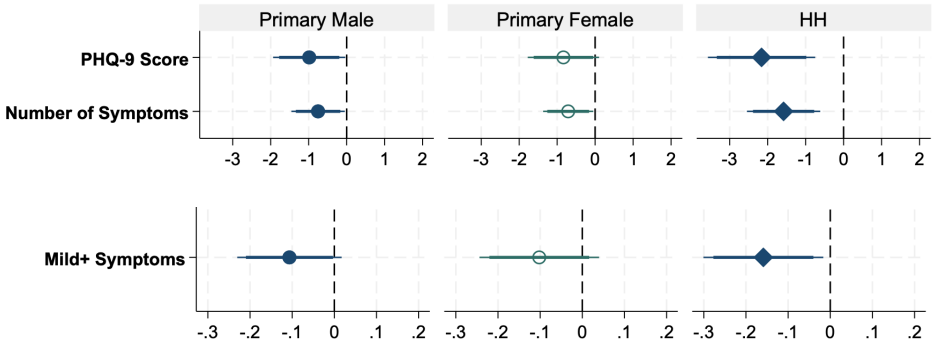
We find that the treatment and control groups are balanced by baseline characteristics for the subset sample of 147 households where PMs also complete baseline and endline interviews, as explained in Section 3.3 (Table D1). We also find that our core results for the PF-reported outcomes are consistent if we restrict to this subsample (Figures D1 and D2).

Table D1: Baseline characteristics of study sample, by treatment group

	Treatment arm (N=76)	Control arm (N=71)	P-value
Primary male characteristics			
Age in years	33.97	33.42	0.59
	(8.21)	(7.13)	
Completed: any school	0.37	0.31	0.48
primary school	0.20	0.31	0.11
secondary school	0.09	0.17	0.15
Main activity in last 7 days is: ag	0.51	0.48	0.58
non-ag	0.09	0.14	0.30
daily laborer	0.17	0.14	0.59
unpaid housework	0.01	0.00	0.32
none/other non-market activity in last 7 days	0.21	0.24	0.66
Primary female characteristics	0.00	0.00	
Age in years	28.33	27.85	0.50
	(5.95)	(4.95)	
Completed: any school	0.34	0.39	0.54
primary school	0.16	0.11	0.49
secondary school	0.07	0.03	0.35
ag	0.04	0.01	0.35
non-ag	0.05	0.06	0.91
daily laborer	0.00	0.00	
unpaid housework	0.91	0.92	0.87
none/other non-market activity	0.00	0.01	0.33
Household characteristics	0.00	0.00	
Age gap between PM and PF	5.64	5.58	0.93
	(6.90)	(5.39)	
No. of young children (under 5): males	0.67	0.56	0.16
	(0.50)	(0.53)	
females	0.49	0.52	0.71
	(0.62)	(0.53)	
No. of older children (5-18): males	0.66	0.68	0.87
	(0.87)	(0.82)	
females	0.61	0.58	0.76
	(0.77)	(0.73)	
No. of adults ($\geq$ 18): males apart from PF	0.11	0.04	0.19
	(0.35)	(0.20)	
females apart from PF	0.12	0.10	0.72
Total value (USD, PPP) of: productive assets	56.72	62.79	0.16
	(30.68)	(35.96)	
consumptive assets	176.29	161.17	0.58
	(198.58)	(163.73)	
livestock assets	615.50	581.64	0.68
	(653.98)	(644.29)	
PM/PF are spouses	0.93	0.97	0.21
PM/PF are parents of index child	0.99	0.99	0.64
Primary female is sole hh head	0.07	0.06	0.80
HH owned phone before study	0.71	0.69	0.79

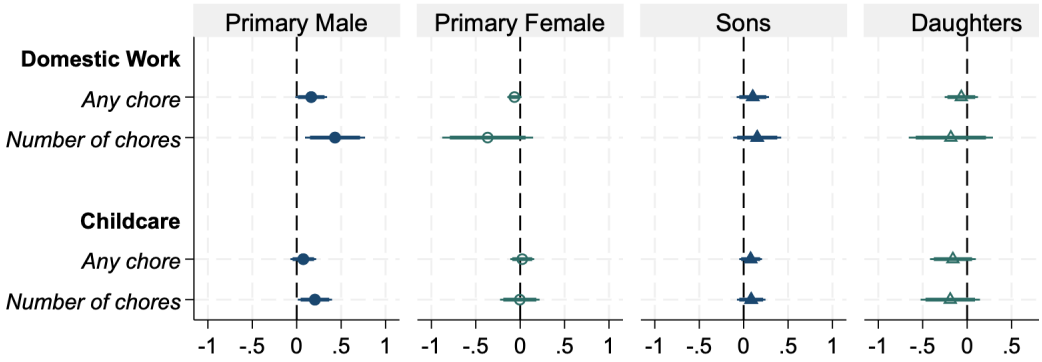
Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The table is identical to Table A3 but uses only the subsample of 147 households with complete PM data. For additional details on this subsample and attrition, see Section 3.3 and Table A2. See the note to Table A3 for additional information on this table.

Figure D1: Treatment effects on presence of depressive symptoms (subsample robustness check)

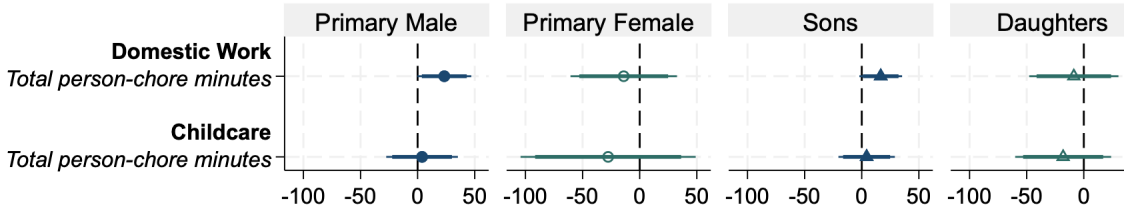


Note: The figure is identical to Figure 3a but uses only the subsample of 147 households with complete PM data. For additional details on this subsample and attrition, see Section 3.3 and Table A2. See the note to Figure 3a for additional information on the outcomes and specification.

Figure D2: Treatment effects on chores performance by person type (subsample robustness check)



(a) Performance of any chore and number of chores performed



(b) Minutes performing chores

Note: The figure is identical to Figure 4 but uses only the subsample of 147 households with complete PM data. For additional details on this subsample and attrition, see Section 3.3 and Table A2. See the note to Figure 4 for additional information on the outcomes and specification.

D.2 Panel Estimation Details

Our main sample consists of the 163 households for whom the PF-reported outcomes are reported at endline. For specifications at the person type (or person-by-chore type), we use probability weights so that estimates represent the average household rather than more heavily weighing households with more total person types present. That is, an observation for a given household receives a weight equivalent to the inverse of the number of person types in the household. For the mental health outcomes which are measured only for the PM and PF, we are missing endline observations from 16 PMs who were not interviewed at endline (see above on attrition) and apply weights in a similar manner.

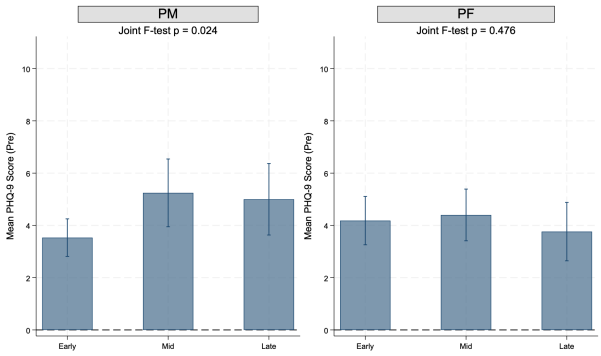
D.3 Measurement: possible time of day effects

Interview timing may affect responses, especially for self-reported mental health outcomes. Thus we examine whether PHQ-9 outcomes vary with the time-of-day at which the module is administered. We create tertiles of interview start time for endline PHQ-9 administration: Early (8:00-12:59), Mid (1:00-3:59), and Late (4:00-8:59). Figure D3 plots baseline and endline PHQ-9 scores by tertile. While average levels differ somewhat across tertiles, all confidence intervals overlap. Joint tests of tertile fixed effects indicate that time-of-day is not an important predictor of endline PHQ-9 scores for PF, whereas for PMs the relationship is not monotonic (mid-day PHQ-9 scores are higher) and we reject the null that time-of-day controls are jointly unimportant. Accordingly, we include time-of-day fixed effects in all estimating equations. For additional details on measurement using the high-frequency time use (HFTU) data, see Appendix E.

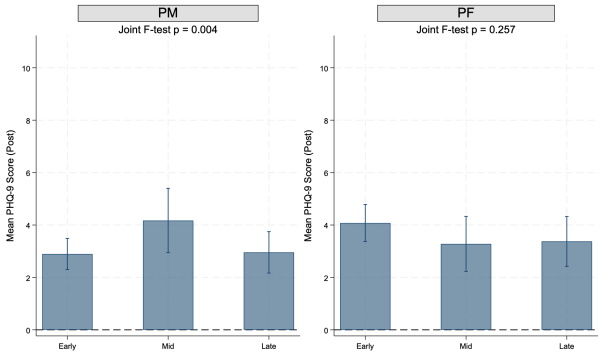
D.4 Measurement: possible enumerator effects

Enumerators can also influence data from interviews, especially for subjective self-reported measures related to mental health outcomes. Additionally, since our intervention phase involved repeated contact between enumerators and respondents (in particular the PMs) and discussions of a personal nature, enumerators and respondents might develop a rapport

Figure D3: Baseline and Endline Depression scores (PHQ-9) by interview time of day



(a) Baseline



(b) Endline

Note: Panels (a) and (b) show mean PHQ-9 scores by respondent and time-of-day tercile for baseline and endline interviews, respectively. Whiskers show 95% confidence intervals. Reported p-values are from joint significance F-tests, based on regressions of the PHQ-9 score on time-of-day fixed effects and village fixed effects, with standard errors clustered at the village level. The time-of-day terciles are morning to early afternoon (8:00 AM-12:59 PM), mid-afternoon (1:00 PM-3:59 PM), and late afternoon to early evening (4:00 PM-8:59 PM).

by the endline interviews that would also affect measurement of PHQ-9 related outcomes. These issues could bias treatment effects if enumerator-respondent rapport develops differently between treatment and control households and/or if rapport develops differently between enumerators and PMs compared to enumerators and PFs.

Mitigating features of study design. We first note that several features of the study design mitigate concerns about rapport being established differently by treatment assignment. First, since intervention phase calls were implemented for all PMs (treatment and placebo) with similar frequency and duration, we maintained comparable levels of contact between survey staff and the PMs. Second, no data were collected during intervention calls other than a stated goal task at the end of treatment calls, limiting the scope for enumerators to establish memories of specific men in a way that would affect their PHQ-9 interviews months later. Indeed from the enumerator side it is unlikely that specifics were retained in their memory for a given respondent since they were conducting intervention phase calls to men from some households at the same time as PHQ-9 surveys for other households due to our staggered wave design. Thus we can be reasonably confident that to the extent that rapport affected measurement of depressive symptoms or any other outcomes, it would arise from the respondent's side. However, since only the PMs were speaking with enumeration staff during the intervention phase, a reduction in measurement error between baseline and endline would likely affect the PMs but not the PFs.

It is unclear whether becoming more accurate would increase or decrease the sample distribution of PHQ-9 scores. Thus a measurement difference would plausibly act as a reduction in classical measurement error in PHQ-9 outcomes between baseline and endline. Since the presence of classical measurement error in an outcome variable attenuates coefficients on treatment variables, the reduction in measurement error at endline for PMs would imply that treatment effects are indeed accurate. On the other hand, since PFs did not have the opportunity to build a rapport with enumerators during the intervention phase, we might still estimate attenuated treatment effects arising from such measurement error, which could explain the slightly smaller effect sizes for PFs and the fact that under the most conservative estimation strategies with baseline covariates we are underpowered for two-way hypothesis

testing (see Section 5.1).

We nonetheless conduct a set of direct diagnostic tests and also demonstrate that our results are robust to including controls for possible enumerator influence.

Descriptive evidence on enumerator influence at baseline. We begin by testing whether baseline PHQ-9 scores differ systematically by enumerator identity, conditional on village fixed effects and time-of-day fixed effects. Specifically, we regress baseline PHQ-9 scores on village fixed effects and time-of-day fixed effects, then add baseline-PHQ-9 enumerator fixed effects and test their joint significance. As reported in the top portion of Table D2, we fail to reject the null hypothesis that baseline enumerator fixed effects are jointly zero, suggesting that baseline enumerator identity does not meaningfully predict PHQ-9 levels after accounting for village and time-of-day for baseline interviews.

Table D2: Enumerator diagnostics for endline PHQ-9

Sample	Enum FE type	Joint tests (p-values) Village FE	Time FE	Enum FE	Key coefficients (θ and κ) Scoring tendency	Familiarity share
Pooled sample at baseline						
		0.001	0.043			
	Baseline	0.004	0.061	0.242		
Control at endline						
(A)		0.527	0.940			
(B)		0.597	0.949		1.546 (0.735)	
(C)		0.621	0.948		1.557 (0.752)	-0.109 (1.367)
(D)	Endline	0.541	0.867	0.020	2.343 (1.407)	-0.562 (1.350)
Treatment at endline						
(A)		0.108	0.237			
(B)		0.206	0.462		1.334 (0.535)	
(D)	Endline	0.263	0.509	0.655	2.187 (1.098)	0.037 (1.195)

Notes: The table reports joint tests from regressions of PHQ-9 on village and time-of-day fixed effects with PHQ-9 enumerator fixed effects, enumerator scoring tendency, and enumerator familiarity share added incrementally for the pooled sample at baseline, and the control and treatment households at endline as described for equations (A) thru (D) below.

Descriptive evidence on enumerator influence at endline. For endline interviews, the core question of rapport is about dynamics: repeated interactions could build enumerator-

respondent familiarity that might shift how respondents answer the PHQ-9 at endline. Simply including endline enumerator fixed effects does not isolate such familiarity/rapport mechanisms because they are specific to an enumerator-respondent matching. We can also use the baseline data to capture an enumerator's tendencies directly rather than relying on fixed effects. We therefore construct the following variables:

1. *Enumerator scoring tendency.* For each enumerator, we calculate the enumerator's average tendency *at baseline* to have higher PHQ-9 scores after controlling for village and time-of-day fixed effects. That is, we predict residuals from a regression of PHQ-9 scores at baseline on village and time-of-day fixed effects and then calculate each enumerator's average residual. To examine the influence of this on endline PHQ-9, we then assign each respondent the scoring-tendency value of the enumerator conducting their endline PHQ-9 interview.

2. *Enumerator-household familiarity.* For each enumerator-PM pair, we compute the share of the PM's intervention phase calls conducted by that enumerator. Taking into account that the PHQ-9 may be administered to the PM and PF by different enumerators—they occurred on different randomly assigned days—we match the correct enumerator familiarity value for each respondent separately based on their household's PM.

Using these measures, we estimate a sequence of regressions for endline PHQ-9 scores, separating the sample by treatment status to avoid confounding with treatment effects. Variables are as described in Equation 1 except for what we have added:

$$\begin{aligned}
\text{PHQ-9}_{ij,post} &= \alpha + \beta\text{PHQ-9}_{ij,pre} + \gamma_v + \tau_d + \varepsilon_{ij} && \text{(A)}\\
&= \text{(A)} + \theta \cdot \text{Scoring Tendency}_{ij} && \text{(B)}\\
&= \text{(B)} + \kappa \cdot \text{Familiarity Share}_{ij} && \text{(C)}\\
&= \text{(C)} + \eta_l && \text{(D)}
\end{aligned}$$

where person type j in household i is interviewed at endline by enumerator l with endline

enumerator fixed effects given by η_l.

Table D2 summarizes key results from these specifications. Across both treatment and control subsamples: (i) village and time-of-day fixed effects are not jointly significant predictors of endline PHQ-9 conditional on baseline PHQ-9; (ii) enumerator scoring tendency is positively associated with endline PHQ-9 (consistent with persistent enumerator-level differences in elicitation/recording); (iii) the familiarity share coefficient is close to zero and not statistically significant; and (iv) endline enumerator fixed effects are not jointly significant in the treatment subsample; in the control subsample they are statistically significant but with a relatively small F-statistic.

As an additional check verifying that enumerator fixed effects are not required, we report Shapley decompositions of the regression R^2 for the control and treatment subsamples using the full set of covariates from (D). Table D3 shows that baseline PHQ-9 explains the largest share of variation; enumerator scoring tendency explains a non-trivial share; and the remaining contributions of village fixed effects, time-of-day fixed effects, familiarity share, and enumerator fixed effects are negligible.

Table D3: Shapley decomposition of R^2 for endline PHQ-9 regressions

Covariate group	Control Shapley value	Treatment Shapley value
Baseline PHQ-9	0.0958	0.1551
Village fixed effects	0.0020	0.0019
Time-of-day fixed effects	0.0100	0.0012
Enumerator scoring tendency	0.0281	0.0768
Enumerator familiarity share	0.0034	0.0072
Endline enumerator fixed effects	0.0046	0.0080
Total R^2	0.1439	0.2501

Notes: Shapley values decompose the regression R^2 into additive contributions by covariate group. The control and treatment regressions use the covariates from specification (D).

As a final check, we re-estimate our main PHQ-9 results from Figure 3 and Table A10 controlling for the enumerator scoring tendency and familiarity shares. Table D4 shows these do not substantively affect our results.

Table D4: Treatment effects on PHQ-9 scores controlling for possible enumerator effects

					With Spouse PHQ-9	
	(A)	(B)	(C)	(D)	(A)	(C)
Treat X PM	-0.910*	-0.780	-0.779	-0.771	-0.950*	-0.821*
	(0.480)	(0.470)	(0.472)	(0.467)	(0.491)	(0.482)
Treat X PF	-0.805*	-0.703	-0.703	-0.723	-0.837*	-0.738*
	(0.457)	(0.452)	(0.452)	(0.466)	(0.443)	(0.437)
Enumerator scoring tendency		1.389***	1.370***			1.385***
		(0.416)	(0.457)			(0.455)
Enumerator familiarity share			0.117	-0.275		0.120
			(0.804)	(0.751)		(0.800)
Enumerator scoring tendency	No	Yes	Yes	Yes	No	Yes
Enumerator familiarity share	No	No	Yes	Yes	No	Yes
Endline enumerator fixed effects	No	No	No	Yes	No	No
F-statistic				4.708		
P-value				0.014		
One-sided hypothesis tests						
PM P-value	0.032	0.052	0.053	0.053	0.030	0.048
PF P-value	0.042	0.063	0.063	0.064	0.033	0.049

Notes: The table reports treatment effect estimates for PHQ-9 scores, with the first column here (A) showing results from our main specification (Figure 3 and Table A10 Column 2) for the pooled sample of 310 PMs and PFs with PHQ-9 scores at both baseline and endline. Columns (B), (C), and (D) add controls for enumerator influence as described above. All other estimation details (e.g., control for baseline score, covariates, clustering, sample) are identical across columns. The final two columns again show results for specifications (A) and (C), respectively, adding the respondent's spouse's baseline PHQ-9 score. For PFs with no PM with complete data, the value has been imputed using the average across all other PFs.

D.5 Alternative inferences

In addition to conventional (two-sided) p-values reported in the main tables, we report three complementary inference approaches to show robustness to specification choice and multiple hypothesis testing.

D.5.1 Pre-specified one-sided tests

For outcomes and directions pre-specified in the pre-analysis plan, we also report one-sided p-values. Several estimates that are marginally significant at the 10% level using two-

sided tests are significant at the 5% level using the corresponding pre-specified one-sided test.

D.5.2 Randomization inference (RI)

We report randomization-inference p-values computed by re-randomizing treatment assignment in a manner consistent with the study's randomization procedure and recomputing the test statistic under each re-randomization. RI p-values are reported alongside conventional p-values in all primary tables.

D.5.3 Multiple hypothesis testing: Anderson (2008) sharpened q-values

To address multiple-testing concerns—for families of related mental health outcomes and chore outcomes—we compute sharpened q-values following Anderson (2008). Table D5 reports sharpened q-values by outcome family, indicator, and household member type. Consistent with the main results on mental health outcomes based on one-sided tests, we continue to find statistically significant improvements in PM's mental health (q-val < 0.05) after adjusting for multiple hypothesis testing, while PF's mental health effects are marginally significant (q-val < 0.1). For chore outcomes, the pattern of results is unchanged with PM's exhibiting a robust increase in DW contributions (q-val < 0.05); while there is no change in PFs' overall chore participation, they experience a significant reduction in their relative DW burden (q-val < 0.05). Overall, adjusting for multiple hypothesis testing does not affect our main results and conclusions.

Table D5: Sharpened q-values by outcome family, indicator, and person type

Outcome family	Indicator	PM	PF	Sons	Daughters
Mental health					
PHQ-9 family	PHQ-9 score	0.05	0.07		
	Number of symptoms	0.05	0.07		
	Mild or worse symptoms (binary indicator)	0.05	0.07		
Domestic work					
DW family	Total number of chores	0.032	0.17	0.13	1
	Total minutes performing chores	0.032	0.40	0.13	1
	Share of person-chores	0.032	0.03	0.13	1
	Share of chore minutes	0.032	0.03	0.13	1
Childcare					
Care family	Total number of chores	0.61	1	1	0.79
	Total minutes performing chores	1	1	1	0.79
	Share of person-chores	0.85	1	1	0.79
	Share of chore minutes	0.85	1	1	0.79

Notes: Sharpened q-values are computed following Anderson (2008). Sharpened q-values for mental health outcomes are based on one-sided tests, reflecting the pre-specified hypothesis, while all chore outcomes are based on two-sided tests.

D.5.4 Multiple hypothesis testing: Outcome indexes

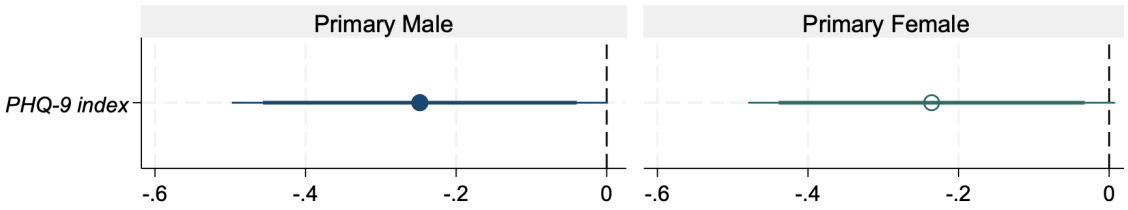
To address concerns about multiple hypothesis testing while preserving the underlying patterns in the data, we aggregate related outcomes into standardized indices for both mental health and chore performance.

For mental health, we construct an index that combines the PHQ-9 score and the total number of depressive symptoms. For chore performance, we construct two indices—one for domestic work (DW) and one for childcare. Each chore index incorporates the outcomes you mention—the total number of chores performed, total minutes performing chores, and the share of person-chores, as well as the share of chore minutes—into equal-weight standardized indices. Each outcome is standardized using the control-group baseline mean and standard deviation within person type and the index is computed as the row mean of the standardized components.

Figure D4 presents results for the mental health index, which is consistent with the main findings from the individual mental health outcomes, showing improvements for both PMs and PFs. Figure D5 reports results for the DW and childcare indices. These results preserve the core PM pattern—a statistically significant increase in DW contributions without a

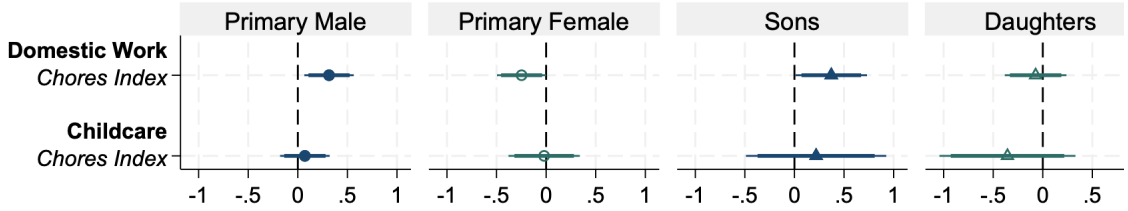
corresponding increase in childcare—while strengthening the PF results, showing a reduction in PFs' DW burden and no change in childcare contributions. The DW index also indicates increased contributions by sons and no change for daughters. Together, these index results provide additional reassurance that our main findings are not driven by false detection across multiple related outcomes.

Figure D4: Treatment effects on mental health index by person type



Note: The figure reports the impact of the intervention on a mental health index at the person type level. The bars around each coefficient depict its 90% and 95% confidence intervals.

Figure D5: Treatment effects on chores performance index by person type



Note: The figure reports the impact of the intervention on a chores performance index at the person type level. The bars around each coefficient depict its 90% and 95% confidence intervals.

D.6 Causal pathways analysis

D.6.1 Total chores performance

We confirm that the household level performance results are consistent with those reported in Table A12 by calculating the weighted sum of the person-type treatment effects, with weights proportional to the number of observations for each person type in the unbalanced household panel (163 PMs, 163 PFs, 76 sons, 75 daughters, 10 adult males, and 12 adult females). The sum total effect is an increase in number of DW chores by 0.59 and an

increase in DW chore minutes by 8.5 minutes per household. The number of childcare chores increases by 0.59 and childcare minutes decreases by 9 minutes per household.

D.6.2 Household chores redistribution outcome

The five household level redistribution categories are formed by first categorizing whether each PF's and PM's chores increased, remained the same, or decreased. Out of the nine possible joint outcome possibilities, we then collapse these into five mutually exclusive groups as follows: (1) Redistribution to PM: $\Delta\text{PF} \leq 0$, $\Delta\text{PM} > 0$; (2) Redistribution to PF: $\Delta\text{PF} > 0$, $\Delta\text{PM} \leq 0$; (3) Both increase: $\Delta\text{PF} > 0$, $\Delta\text{PM} > 0$; (4) PF load lightens: $\Delta\text{PF} < 0$, $\Delta\text{PM} \leq 0$; and (5) No change: $\Delta\text{PF} = 0$, $\Delta\text{PM} = 0$.

D.6.3 Time allocation and desirability bias

We note that these HFTU-based measures of DW and childcare performance give additional assurance that the results reported in Section 5.2 reflect a true response and not desirability bias on the part of the respondent. We pre-specified PF reporting of chores as our main outcome because we believed the PF would be most knowledgeable about performance of UDCW and we noted that reports from the PM could have been affected by a desirability effect which might be affected by the intervention. While we designed the chores module to minimize possible bias in reporting, it remained a possibility that a PF's desire to report the PM's participation in chores could have been affected by the PM's treatment. On the other hand, the HFTU data do not rely on any respondent's ability to recall or estimate duration of activities from the previous day at all, and thus are not affected by the many forms of recall bias that can affect reporting (e.g., telescoping). Our measurement approach is designed to limit social desirability bias relative to standard recall-based surveys, rather than to eliminate it entirely.

E High-frequency Time Use (HFTU) Survey Details

This appendix section provides additional details on the high-frequency time use (HFTU) data collected which we use for analyses in Section 6. We include summary statistics, attrition information, and details on the empirical methodology, including robustness tests.

E.1 Methodology Protocol for HFTU surveys

The HFTU surveys aimed to collect time allocation data from the "target respondent" (self-report) and about the other primary respondent (proxy report) using an experience sampling method (ESM), a psychological research method pioneered to collect data on affect and emotion (Csikszentmihalyi, 2014). Our research team previously implemented this method in a different but very similar sample in Ethiopia (2020-2022) alongside traditional 24-hour recall surveys like the World Bank's LSMS and the American Time Use Survey (Krueger et al., 2009), which is reported on in McGavock et al. (2025) as a measurement experiment. There we show that the HFTU method is likely more accurate than traditional 24-hour day reconstruction surveys for this population. We note that a similar comparison was outlined by Krueger (2007), but it was never completed.

E.1.1 Sampled Weeks

For the BCC intervention experiment reported on in this article, HFTU surveys were conducted during four different weeks for each household, with each primary respondent surveyed one week each at baseline and endline. At baseline and endline, the order within a household was randomly assigned (e.g., whether the PM's or the PF's HFTU week occurred before the other person's during baseline/endline) such that all possible combinations were covered within each four-household village cluster (see Footnote 18 in Section 3). These surveys took place after the respondents' interviews covering chores completion and PHQ-9 modules as described in Section 3.

E.1.2 Phone Logistics

All households received a new basic phone during their in-person baseline interview to ensure that phone access would not vary across or within households due to prior ownership or use patterns. Due to the need for the target respondent to carry a phone throughout their HFTU survey week and limited access to electricity for some households, we placed reminder calls to the phone before the survey period and scheduled an "off" week between the two target respondents' HFTU survey weeks so they could begin the week with a fully charged phone (e.g., by sending the phone with someone to the market for charging).

E.1.3 Daily Randomized Sampling

Once a respondent's 7-day HFTU survey period began, enumerators dialed the phone number according to a pre-programmed schedule. In total, each respondent was assigned five calls per day. Of the respondent's 24-hour day, 11 hours are subject to sampling using our HFTU methodology (7am-6pm, the call center's operating hours). Call times were randomized within five strata covering the day such that all respondents have balanced coverage of these hours.

E.1.4 Call attempts and callback protocol

At the assigned time, the enumerator initiated an instance of the survey, recorded the assigned time, and then dialed the phone up to three times. If the phone was answered, the enumerator identified herself and verified that she was speaking with the respondent. In some cases, someone else answered the phone. If we failed to speak with the respondent at the intended time, the enumerator recorded the reason as they perceived it (e.g., respondent not available, no answer, busy signal, phone off, network error message). She marked the call for a callback, which involved making one additional attempt within two hours at a time set aside for callbacks within the same stratum on that day's randomized schedule. If the respondent did not pick up the phone during the callback attempt, that phone call (e.g., call number 2 of that day) was marked as a missing observation and the enumerator moved

onto the respondent's next randomized time in the schedule. Section E.2 shows summary statistics on call outcomes and the resulting usable sample.

E.1.5 Survey Content

Each completed HFTU call lasted around five minutes, during which the enumerator asked about the respondent's current activity, their spouse's activity (or best guess), current companions and location, and their current happiness (on a 5-point Likert scale). The brief surveys minimized respondent burden while collecting data on subjective well-being, secondary activities, and the reported activities of the spouse, helping us explore labor and household dynamics. We also collected information on children's activities (ages 5-12) and the supervisor(s) of any children under 5 by inquiring about children listed on the household roster. We do not report on children's activities here, however, in order to keep the analysis focused on primary outcomes for this article.

E.2 Call Compliance and Attrition

E.2.1 Sample restriction

Before addressing call completion, we restrict the HFTU data to the set of all calls exported by the survey team for the 163 households in our primary analysis sample (see Section 3 for details). As assigned this would be 22,820 calls (35 for four weeks total per household). Due to some data entry errors, however, about three observations were lost per household on average. In total, we have 22,352 calls.

E.2.2 Call outcomes

For PM and PF, at baseline and endline about 65% of calls were answered on the first try, with 67% usable since some were recovered using our callback procedure or by asking the person answering to connect us with the target respondent.

Enumerators reported more issues with phones being turned off or network error messages at endline than baseline, consistent with what we heard from respondents and our field team

regarding power supply, mobility, and general social concerns about the unfolding political situation. Callbacks were not often successful, but they were on occasion. If a call was answered when intended, the respondent could be located by passing the phone or calling a different phone, or the callback was successful and we have complete information for the respondent's activity at the time of the call (the first substantive question of the survey)[45] we refer to the call as being "usable."

E.2.3 Attrition based on compliance thresholds

Figure E1 shows the share of calls that are usable by matching PMs (y-axis) and PFs (x-axis) from the same household, by treatment and control groups. Within a household, these are positively correlated but not one to one. The elasticity of call usability for PMs with respect to PF call usability is less than one: that is, the PFs with the highest compliance rates are more compliant than their PMs (below the 45 degree line), while in households with PFs with the lowest compliance rates, the PMs are more compliant than PFs. There are no major differences in this relationship across treatment (blue) and control households (green) since there are no significant portions where the confidence intervals fail to overlap. However, we note that the mean for control households is above the confidence interval for treatment households beteween 20 and 40 percent usability for the PFs. This makes 40% a reasonable threshold for defining minimum compliance for including a household in the HFTU analysis. We thus restrict our HFTU analysis to these households, which generates a sample of 87 households who have reasonably complete HFTU data sufficient to generate a meaningful summary of activity shares for them based on the available calls. Table E1b shows that this restriction generates a sample in which both the PMs and PFs have reasonably high compliance rates on average, with the mean for calls answered on the first try exceeding 70% for both PMs and PFs at both baseline and endline. Table E2 demonstrates that this approach results in about 56% of the households being retained (87 total households), and

[45] Sometimes respondents told enumerators that they were too busy to talk longer, so during piloting we established a protocol whereby respondents were instructed that if at all possible we wanted them to pick up the phone at the time and simply answer the first question and if they did so they would receive the participant incentive for that call (a tiny token amount of airtime credit that accrued during the week and was transferred to the survey at the end of the baseline or endline phase).

Table E1: Call outcomes for full sample ($N = 163$) and HFTU sub-sample with $\geq 40\%$ usable calls

(a) Full sample

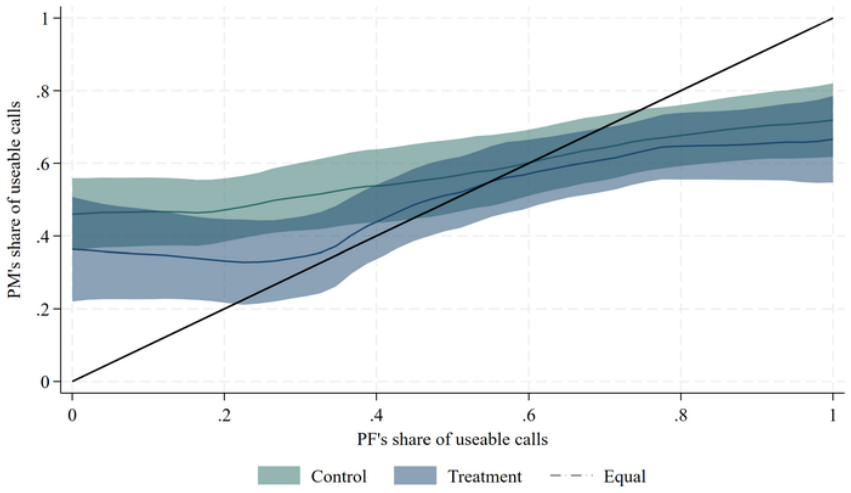
	Primary Male			**Primary Female**		
Baseline	Share	T-C	P-value	Share	T-C	P-value
Answered first try	0.651	-0.050	0.219	0.606	-0.012	0.793
Callback successful	0.018	-0.005	0.183	0.018	-0.004	0.311
Phone with someone else	0.050	0.011	0.411	0.025	-0.019	0.003
Phone off	0.219	0.049	0.138	0.275	0.050	0.209
Network issue	0.012	-0.003	0.532	0.022	-0.004	0.532
Other	0.050	-0.002	0.858	0.054	-0.011	0.244
Usable call share	0.675	-0.047	0.000	0.620	-0.015	0.236
N calls	5604	.	.	5738	.	.
N households	163	.	.	162	.	.
	Primary Male			**Primary Female**		
Endline	Share	T-C	P-value	Share	T-C	P-value
Answered first try	0.533	-0.074	0.105	0.516	0.012	0.806
Callback successful	0.007	0.002	0.454	0.010	-0.001	0.738
Phone with someone else	0.037	0.019	0.030	0.032	-0.009	0.306
Phone off	0.349	0.060	0.165	0.377	0.009	0.854
Network issue	0.022	-0.011	0.052	0.018	-0.007	0.223
Other	0.052	0.003	0.755	0.048	-0.005	0.612
Usable call share	0.541	-0.075	0.000	0.524	0.003	0.812
N calls	5467	.	.	5543	.	.
N households	161	.	.	163	.	.

(b) HFTU sub-sample ($\geq 40\%$ usable calls)

	Primary Male			**Primary Female**		
Baseline	Share	T-C	P-value	Share	T-C	P-value
Answered first try	0.740	-0.017	0.730	0.727	0.033	0.533
Callback successful	0.015	-0.004	0.308	0.019	-0.009	0.112
Phone with someone else	0.047	0.006	0.769	0.018	-0.027	0.001
Phone off	0.153	0.015	0.703	0.179	-0.004	0.925
Network issue	0.010	-0.002	0.449	0.017	-0.005	0.449
Other	0.036	0.003	0.730	0.041	0.013	0.280
Usable call share	0.756	-0.021	0.180	0.738	0.036	0.021
N calls	3036	.	.	3094	.	.
N households	87	.	.	87	.	.
	Primary Male			**Primary Female**		
Endline	Share	T-C	P-value	Share	T-C	P-value
Answered first try	0.724	-0.022	0.527	0.716	0.000	0.993
Callback successful	0.007	0.004	0.146	0.011	-0.003	0.515
Phone with someone else	0.028	0.019	0.017	0.026	-0.008	0.357
Phone off	0.185	0.016	0.606	0.185	0.008	0.801
Network issue	0.018	-0.019	0.001	0.013	0.000	0.967
Other	0.038	0.003	0.813	0.049	0.002	0.909
Usable call share	0.729	-0.021	0.190	0.726	-0.005	0.769
N calls	2957	.	.	2980	.	.
N households	87	.	.	87	.	.

we find no significant difference in the attrition rates between treatment and control groups based on this HFTU threshold.

Figure E1: Treatment and control usable call shares for PMs vs PFs



Note: The figure reports the relationship between usable call shares for PMs and PFs by treatment assignment.

We show that analyses using the HFTU data are robust to restricting the sample with even higher thresholds (see below). Additionally, household fixed effects are always used in analyses, allowing us to use the panel nature of the dataset to control for unobservable characteristics that might be time invariant but related to compliance. We also show that results are robust to a bounding exercise for selection effects arising from calls that are unanswered selectively based on the respondent's activity and/or location. Finally, we investigate re-weighing observations using the call usability for respondents. In all cases, we find qualitatively similar results.

Table E2: Share of households with usable call shares at endline above a given threshold

	Share of Respondents		**Share of Households**	**Household Differential Attrition Test (T vs. C)**
Usable Calls Threshold	PM	PF	PM & PF	P-value
20	0.850	0.784	0.680	0.907
40	0.752	0.686	0.556	0.943
60	0.523	0.471	0.340	0.692
80	0.196	0.301	0.118	0.598

E.3 Construction of HFTU Outcomes

E.3.1 Primary outcome from HFTU: activity shares

We use responses from the HFTU surveys at baseline and endline to estimate a set of treatment effects for the PM and PF summarized in Table E3. Our primary outcome of interest can be conceptualized as the share of time that the PM and PF devoted to activities by category over the 7-day survey period, based on the fraction of answered calls during which they reported activities that were coded by the enumerator. We aggregate these to eight mutually exclusive categories and one "total work" category that is the combination of economic activities, domestic work, and childcare. Economic activities include unpaid subsistence agriculture and livestock rearing; social activities include eating or drinking coffee with family, visiting neighbors, attending community events or religious services; unpaid domestic work includes all of the domestic work chores reported on in other parts of the manuscript as well as some that were not separated (e.g., sewing/repairing clothes, repairing home construction); care of children covers all activities reported by the respondent as primary activities, such as bathing a child (but does not include any supervision as a secondary activity); and leisure activities include entertainment; travel includes movement between activities (e.g., walking to their fields); and other includes visiting the market and personal hygiene.

In practice, our estimation occurs at the call level, so that we can control for call-specific information like time of day or day of week that would be associated with activities or non-response. This means our outcome variable in these estimations is a binary indicator for whether the respondent's activity at the time of the call was in that category.

E.3.2 Measurement concerns

Because the HFTU data are reported in the moment with no reliance on recall and our outcomes are shares constructed across many such calls, we are substantially less concerned that these data are subject to desirability bias than a 24-hour recall survey would be. Supporting this point, we find relatively large shares of time spent on leisure for both the PF

Table E3: Baseline Self- and Proxy-reported Time Allocation Shares

	Baseline mean activity shares (%)		Baseline Difference	
	Self Report	**Proxy Report**	**Self - Proxy**	**P-value**
PM's Activities				
Work: Sum of economic, chores, & childcare	0.365	0.392	-0.027	0.017**
economic activities	0.269	0.296	-0.027	0.010***
household chores	0.065	0.044	0.021	0.000***
childcare	0.030	0.051	-0.021	0.000***
Social activities	0.202	0.208	-0.006	0.519
Other, excluding leisure	0.178	0.021	0.157	0.000***
Leisure	0.177	0.134	0.043	0.000***
PF's Activities				
Work: Sum of economic, chores, & childcare	0.630	0.471	0.159	0.000***
economic activities	0.050	0.032	0.018	0.000***
household chores	0.493	0.319	0.174	0.000***
childcare	0.086	0.119	-0.033	0.000***
Social activities	0.119	0.070	0.049	0.000***
Other, excluding leisure	0.091	0.028	0.063	0.000***
Leisure	0.121	0.046	0.075	0.000***
Combined Share: Sum of PF and PM Shares				
	PF's report	PM's report		
Work: Sum of economic, chores, & childcare	0.501	0.403	0.098	0.000***
economic activities	0.167	0.145	0.022	0.051*
household chores	0.269	0.187	0.082	0.000***
childcare	0.066	0.072	-0.006	0.449
Leisure	0.123	0.109	0.014	0.163
Relative Share: PF share / Combined Share				
	PF's report	PM's report		
Work: Sum of economic, chores, & childcare	0.641	0.558	0.083	0.000***
economic activities	0.143	0.100	0.043	0.071*
household chores	0.927	0.842	0.085	0.000***
childcare	0.620	0.770	-0.150	0.001***
Leisure	0.490	0.203	0.287	0.000***

Notes: This table presents baseline time allocation shares (%) based on self- or proxy-reported activities at the random times when a respondent was reached. The difference column represents the self-reported value minus the proxy-reported value, so negative values indicate that the person's spouse reported the activity more often, as a share of all calls to them, than the person did themselves. p-values are from paired t-tests testing for differences between self and proxy reports.

and PM, and both respondents report a leisure activity *more* often for themselves than their partner does when proxy-reporting. Further, our estimates suggest mostly null treatment effects on time allocations across major activity categories.

E.3.3 Subjective well-being scores

During each call immediately after asking for their primary and secondary activity, we recorded the respondent's answer to the question "How happy are you right now?" according to a 5-point Likert scale, where higher scores indicate more happiness. For each respondent type, we normalized scores at baseline using the mean and standard deviation from all subjective well-being scores across all activities.

E.4 HFTU Estimation strategy

Using the HFTU data, we estimate the effect of our BCC intervention on self-reported activity shares and other outcomes. Following McGavock et al. (2025), our preferred estimates use the data at the call level to allow for controls like call time that would affect the outcomes, as well as household fixed effects. Specifically, we estimate the equivalent of a difference-in-differences specification, with household fixed effects, by pooling all calls into a single sample. The equation is:

$$Y_{kh\tau} = \omega_1 T_h * \text{Post}_\tau + \omega_2 Post_\tau + \nu \text{Call number}_\tau + \delta_\tau \eta_h + u_{kh\tau}, \tag{5}$$

where $Y_{kh\tau}$ is an outcome variable measured at randomly assigned time τ for individual k in household h. The variable T_h is an indicator that equals one if household h is assigned to the treatment group, meaning that the PM received the intervention calls, and zero if the PM received the control calls with placebo content. Post_τ is an indicator that equals one if the call occurred during the post-intervention period. The coefficient ω_1 on the interaction of the treatment and post dummies is our estimate of interest. The control Call number_τ takes values from 1 to 35 for the call number in the week, to control for any time trend or day of week effects. We control for time of day fixed effects (δ_τ) using the five strata for the call

center schedule, and household fixed effects (η_h), which absorbs all household-specific time-invariant characteristics (including village fixed effects). The error term for person type k in household h at time τ is denoted by $u_{kh\tau}$. We compute robust standard errors clustered at the household level to correct for within-household correlations, including serial correlation.

Since the PM and PF both reported on their own activities at the time of a given call and on their spouse's activities, we estimate Equation 5 separately for own- and self-reports, for the PM and PF.

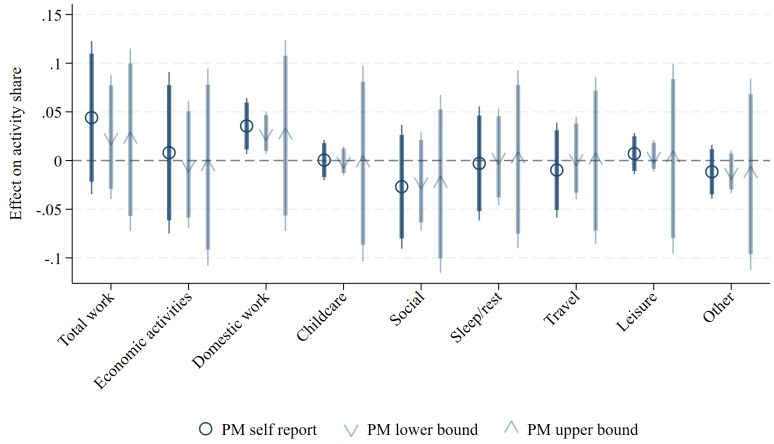
E.5 Robustness of HFTU results

We report on robustness checks for the HFTU analyses in this section.

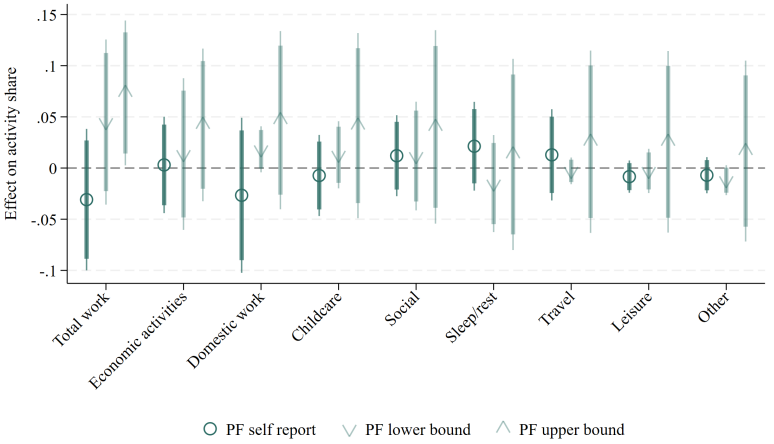
E.5.1 Bounding activity shares for selective non-compliance

The HFTU activity shares (average share of time performing an activity) and treatment effects estimated for them rely on calls that are answered. To the extent that calls are not answered, activity shares may not represent the true activity shares and therefore treatment effects may be biased if there are any systematic reasons that certain activities are more/less likely to be performed (i.e., not reported on) at the time of calls that are not answered. To evaluate whether this causes concern for our estimates reported in Figure 5, we re-estimate the results for PM and PF activity shares using alternative definitions of the outcome where we assume that an activity was always (never) performed at the time of an intended (dialed) call that did not get answered (i.e., was not usable). Figure E2 shows that the treatment effects are broadly consistent if we make these changes. In particular, Panel (a) shows that the PM's increased domestic work is robust to the assumption that at times we failed to reach the PM he was *never* performing domestic work (the lower bound).

Figure E2: Figure 5 robustness - possible selection effects due to unanswered calls



(a) PM lower and upper bounds



(b) PF lower and upper bounds

Note: The figure repeats the coefficients presented in Figure 5 but uses the entire sample of all calls to establish bounds (upper and lower) on treatment effects under the assumption that a call that was not picked up (i.e., not usable) would have recorded a respondent as always (or never) performing the activity category.

E.5.2 Robustness of main results to alternative compliance thresholds

We also re-estimate treatment effects for Figure 5 by retaining households based on different thresholds for defining attrition in the HFTU sample. That is, Figure E3 shows that results are largely unchanged for the HFTU data if we use a 20% or 60% cutoff for compliance rates to determine the set of households with complete HFTU data for analysis purposes. When we use 80% as the threshold, there are a few changes in treatment effects (e.g., the economic activities treatmetn effects are positive and statistically significant for the PM and for the total HH share) but we believe this to be a combination of treatment effects and selection effects based on who is able to maintain an 80% HFTU call compliance share, rather than a pure treatment effect (recall this is only 12% of the total sample as seen in Table E2).

Table E4: Balance tests for call usability

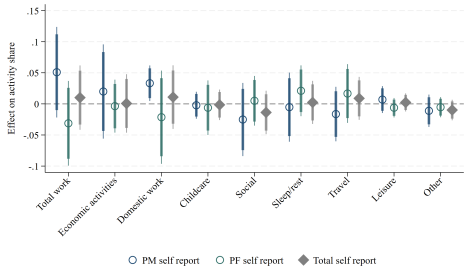
	(1)	(2)	(3)	(4)
Panel A: All Households				
Treat	-0.011	-0.005	-0.004	-0.006
	(0.013)	(0.013)	(0.013)	(0.013)
PM	0.069***	0.073***	0.073***	0.071***
	(0.013)	(0.013)	(0.013)	(0.013)
Post	0.027	0.049*	0.072**	0.052*
	(0.017)	(0.028)	(0.028)	(0.029)
Treat x PM	-0.030	-0.038**	-0.038**	-0.038**
	(0.018)	(0.018)	(0.018)	(0.018)
Treat x Post	0.020	0.007	0.006	0.009
	(0.018)	(0.018)	(0.018)	(0.018)
Post x PM	-0.011	-0.025	-0.024	-0.022
	(0.018)	(0.019)	(0.019)	(0.019)
Treat x Post x PM	-0.050*	-0.022	-0.023	-0.022
	(0.026)	(0.026)	(0.026)	(0.026)
Calendar Week	-0.008***			
	(0.001)			
Call Number in Week			0.002***	
			(0.000)	
Calendar Week F-stat		11.867	12.396	11.819
Calendar Week P-value		0.000	0.000	0.000
Call Number in Week F-stat				4.723
Call Number in Week P-value				0.000
N calls	22352	22352	22352	22352
N households	163	163	163	163
Panel B: Usable Calls Greater Than 40				
Treat	0.031**	0.031*	0.033**	0.024
	(0.016)	(0.016)	(0.016)	(0.016)
PM	0.043***	0.047***	0.048***	0.042***
	(0.016)	(0.016)	(0.016)	(0.016)
Post	0.098***	0.080**	0.119***	0.051
	(0.021)	(0.040)	(0.040)	(0.042)
Treat x PM	-0.053**	-0.053**	-0.054**	-0.051**
	(0.022)	(0.023)	(0.023)	(0.023)
Treat x Post	-0.038*	-0.040*	-0.041*	-0.031
	(0.023)	(0.023)	(0.023)	(0.023)
Post x PM	-0.033	-0.035	-0.034	-0.028
	(0.022)	(0.023)	(0.023)	(0.023)
Treat x Post x PM	0.038	0.033	0.032	0.028
	(0.032)	(0.033)	(0.033)	(0.033)
Calendar Week	-0.006***			
	(0.001)			
Call Number in Week			0.002***	
			(0.000)	
Calendar Week F-stat		7.380	7.471	7.307
Calendar Week P-value		0.000	0.000	0.000
Call Number in Week F-stat				4.017
Call Number in Week P-value				0.000
N calls	12067	12067	12067	12067
N households	87	87	87	87

Table E5: Activities results

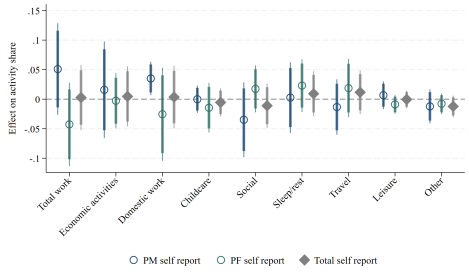
Variable	PM	PF	Total
Total work	0.051	-0.043	0.003
	(0.039)	(0.036)	(0.028)
Economic activities	0.016	-0.003	0.005
	(0.041)	(0.023)	(0.026)
Domestic work	0.035**	-0.025	0.004
	(0.014)	(0.040)	(0.027)
Childcare	-0.000	-0.014	-0.005
	(0.012)	(0.021)	(0.012)
Social	-0.035	0.017	-0.011
	(0.032)	(0.020)	(0.019)
Sleep/rest	0.003	0.023	0.009
	(0.030)	(0.022)	(0.019)
Travel	-0.013	0.019	0.012
	(0.024)	(0.025)	(0.019)
Leisure	0.007	-0.009	-0.000
	(0.012)	(0.008)	(0.008)
Other	-0.012	-0.008	-0.012
	(0.014)	(0.009)	(0.009)

Figure E3: Figure 5 robustness - alternative compliance thresholds

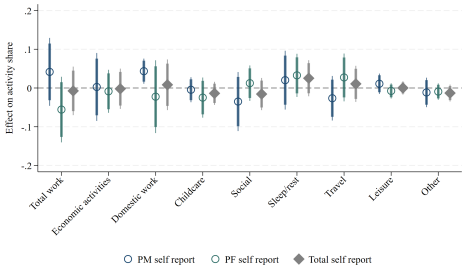
(a) 20%



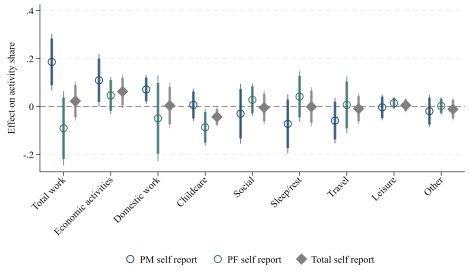
(b) 40% (as appears in main figure)



(c) 60%



(d) 80%



Note: These figures repeat the coefficients presented in Figure 5 but vary the sample of households by restricting to households where compliance rates for calls is at least above a given threshold for both the PF and the PM, as indicated in the captions. Table E2 shows the size of the sample retained and that attrition defined in this manner is balanced.

F Pre-specified Analysis Plan

Prior to conducting this study, we pre-specified our intervention and analysis plan (Assefa et al., 2023). The full plan is available here and was filed in the AEA REC Registry (ID: AEARCTR-0011907). Below is a summary of the plan and an explanation of key deviations.

Summary of the Pre-Analysis Plan

Our pre-analysis plan pre-specified the experimental design, key outcomes, and estimation strategy for evaluating a 13-week phone-based behavior change communication (BCC) intervention that encouraged men in ultra-poor Ethiopian households to increase participation in unpaid domestic and care work.

Design. The plan detailed a household-level randomized controlled trial with 192 households sampled from 48 clusters. Primary male respondents in treatment households received seven scripted, bi-weekly phone calls engaging them in personalized interactive discussions about gender norms and chores. Primary male respondents in control households received placebo calls about food consumption. Baseline and endline data were collected by phone from both primary male and primary female respondents.

Primary outcomes. Five main hypothesis groups were pre-registered. (1) Performance of household chores (extensive and intensive margins) by men, women, and other household members; (2) Time allocation to economic, household, childcare, and leisure activities from high-frequency time-use surveys; (3) Perceptions about and satisfaction with the household division of labor and leisure time; (4) Concordance between men's and women's reports and beliefs about time use and chores; (5) Mental health outcomes for both spouses, measured using the standard PHQ-9 score and an indicator for moderate-to-severe depressive symptoms; and (6) A sixth, conceptually distinct, hypothesis group concerning a methodological experiment on labor-supply measurement was also pre-specified.

Estimation. For the core groups except (2) and (6), the PAP specified ANCOVA models controlling for baseline outcomes without additional controls and also including baseline controls selected using a machine learning algorithm. Due to the research design and number of clusters, we specified that we would cluster standard errors at the village (*gott*) level, performing either a wild bootstrapping inference procedure or randomization inference (Cameron & Trivedi, 2005). The key outcomes were enumerated *ex ante*, along with subgroup analyses (e.g., by age of the primary male and task visibility to others in the community).

Hypotheses. We anticipated that we would be underpowered after attrition for some two-sided hypotheses and were constrained on resources for sample size, so we pre-specified some one-sided hypothesis tests. For mental health, the PAP predicted that the PF's mental health would improve but were agnostic about the direction of change for the PM's mental health. For chores participation, we hypothesized a decrease in women's participation measures (number, minutes, share) and an increase for at least some member types apart from the PF. For time use, we hypothesized that men's participation in chores would increase, his participation in "other" activities (leisure, travel and social) would decrease, and his participation in economic activities would not change. For the PFs, we specified a decrease in chores, an increase in economic activities, and an increase in "other" activities. We hypothesized improved concordance between PMs and PFs in their reporting of each others' activities. And we hypothesized that PFs' life satisfaction measures would improve.

Deviations from the Pre-Analysis Plan

The analyses presented in this manuscript closely align with the pre-analysis plan. Here we list and explain the deviations.

1. **Outcomes not reported.** We did not report on hypothesis group (3) (perceptions and satisfactions with the household divisions of labor) because these questions were inadvertently dropped from the questionnaire due to programming errors by our survey staff. We do not report on the conceptually distinct hypothesis group (6) because it uses a distinct experimental design and addresses a distinct set of questions. For

the present article we have limited our reporting of hypothesis group (4) addressing concordance between PMs' and PFs' reporting of chores performance due to space limitations. Because these hypotheses require a different type of conceptual framework to motivate and test, there was not room in the current manuscript to properly report on concordance, thus we plan to explore this in future research. We report only on perception gaps in time use, not in the chores module reporting.

2. **Additional outcomes not pre-specified.** The PAP pre-registered mental health outcomes separately for both spouses. In this article, we also incorporated a household-level mental health outcome that is the sum of both spouse scores, allowing us to extend the analysis and explore whether the improvements were complementary within households.

3. **Ordering of outcomes.** We initially prioritized relative measures of chores performance (shares of chore minutes) over absolute measures (number of chore minutes) due to concerns about PFs' ability to estimate time duration. After reflecting more on the interdisciplinary literature addressing mental health and UDCW, which separately emphasizes pathways by which absolute burden and relative burden can affect mental health, and seeing our results on mental health gains for both spouses in the same households, we decided to explore the absolute mechanism separately from the relative burden mechanism. Thus in this article we have placed a higher emphasis on the total chores minutes than initially prespecified. We showed robustness to recall bias by comparing the time use estimates to separate time measured using the high frequency time use data. Furthermore, we address the mental health outcome first for storyline purposes in our introduction and conceptual framework, to prioritize what our results suggest is the main policy implication and contribution.

4. **Pooling public and private chores.** During our more extensive review of the interdisciplinary literature connecting mental health and UDCW in preparation of the manuscript, we found a strong distinction between childcare and domestic work (DW) in terms of their mental health impacts. Therefore, we have separately reported on

DW and childcare. By pooling specific chores into these groups this also assists with concerns about multiple hypothesis tests. Adding a public versus private distinction on top of this introduced unnecessary complexity to the storyline, so we dropped it.

5. **Collapsing of household member types.** We prespecified that, in addition to the PM and the PF, we would test performance of chores by other adult women, other adult men, male youth, female youth, male children, female children, male elderly and female elderly. However, our sampled households have very few adults apart from the PM and the PF. Therefore we collapsed the member type categories into sons, daughters, other adult men, and other adult women.

6. **Omission of mechanism based on predicted chore.** We initially said we would complete heterogeneity analysis based on the PMs' and PFs' predictions about chores they or their spouse would perform more (or less) of if they performed more (or fewer) chores. We did not pursue this heterogeneity analysis after reviewing the data at baseline because in order to streamline the storyline.

7. **Inference.** The pre-analysis plan specified a clustered wild bootstrap procedure or randomization inference (RI) for statistical inference. The wild cluster bootstrap was proposed as a finite-sample refinement to cluster-robust inference but later deemed unnecessary. In the final analysis, we report clustered standard errors instead, which is appropriate given the clustering structure of the data and number of clusters (48). Consistent with the pre-analysis plan, we also report RI p-values for the main results. Inference is qualitatively unchanged across methods.

8. **Heterogeneity analysis.** Initially, we reported that we would explore heterogeneity by the PM's age. We did not explore age heterogeneity in detail for this manuscript as our sample was more homogeneous in age than we anticipated and thus it did not connect with the storyline. Instead, we expanded our heterogeneity analysis to consider baseline participation by PMs and baseline UDCW burden by PFs to fit with our expanded theory of change.

G Theoretical Bounds on PF Effort Reductions

This appendix section details the back-of-the-envelope calculation we use to demonstrate that any effort reductions for PFs afforded by the intervention's impacts on chores participation by other household members cannot be very large. First, define a household's total production of unpaid domestic and care work (UDCW) as follows:

$$Y_t = \sum_p \mu_p e_{pt} h_{pt} \tag{1}$$

where t indexes the time period and p indexes a person in the household; $h_{pt} \geq 0$ is the time spent on UDCW by person p at time t. Time produces UDCW for the household based on 1) a productivity parameter that is intrinsic, person-specific, and fixed over time ($\mu_p > 0$); and 2) a level of effort $e_{pt} \in [0, 1]$ chosen by each person in each period. For simplicity, we will assume two members of the household, f and m, to indicate the primary female and primary male members. This also fits our sample well: since the PFs perform the vast majority of UDCW, it is without loss of generality a way to capture UDCW performed by her or by other people (including the PM).

We make a few normalizations and reasonable assumptions when considering pre- and post-intervention outcomes ($t = 0$ and $t = 1$, respectively):

i. Assume that production is constant before and after, i.e., $Y_0 = Y_1 = \bar{Y}$ so that we can pin down a change in effort. This is consistent with the results we present on UDCW chores coverage in Section 6.1.

ii. Assume that the PF exerts maximum effort at baseline, i.e., $e_{f0} = 1$, to maximize the allowable reduction in her effort.

iii. The PF's time spent is the same across time, i.e., $\bar{h}_f = h_{f0} = h_{f1}$, to be consistent with our results in Section 5 showing no significant change in the PF's time spent on all UDCW.

iv. Define $\hat{\mu} = \frac{\mu_f}{\mu_m}$ as the relative productivity of the PF compared to the PM in UDCW

specifically. This is equivalent to normalizing $\mu_m = 1$. Consistent with the literature on UDCW, we will assume that the PF is more productive than the PM at UDCW, i.e., $\hat{\mu} > 1$.

By assumption i., we know that:

$$\mu_f e_{f0} h_{f0} + \mu_m e_{m0} h_{m0} = \mu_f e_{f1} h_{f1} + \mu_m e_{m1} h_{m1} \tag{2}$$

Making substitutions from ii. and iii., we have:

$$\mu_f \bar{h}_f + \mu_m e_{m0} h_{m0} = \mu_f e_{f1} \bar{h}_f + \mu_m e_{m1} h_{m1} \tag{3}$$

Rearranging and incorporating iv., we know:

$$\hat{\mu} = \frac{e_{m1} h_{m1} - e_{m0} h_{m0}}{\bar{h}_f (1 - e_{f1})} > 1 \tag{4}$$

This means that we can define an upper bound for the effort saved for the PF as follows:

$$\frac{e_{m1} h_{m1} - e_{m0} h_{m0}}{\bar{h}_f} > (1 - e_{f1}) = \text{PF's effort reduction} \tag{5}$$

Note this is greater than or equal to zero since e_{f1} is bounded above by 1.

Using approximate values for the baseline data on minutes spent by person and the treatment effects we estimate, we can calculate the maximum effort reduction afforded to the PF using the expression on the left hand side in (5), under different scenarios, i.e., combinations of PM's baseline and endline effort:

Table H1: Effort reduction scenarios

Scenario	h_{m0}	h_{m1}	e_{m0}	e_{m1}	$1 - e_{f1}$
PM no time change, full effort	25	25	1	1	0
PM time ↑ at full effort	25	50	1	1	0.083
PM effort ↑ only	25	25	0.5	1	0.042
PM both time and effort ↑	25	50	0.5	1	0.125
PM time ↑, effort ↓	25	50	1	0.5	0

Note: PM's baseline and endline time (h) and effort (e) are assumed values for illustration purposes. We set $\bar{h} = 300$.