Things take the time they take: Lessons on measurement error from high-frequency time use surveys*

Tamara J. McGavock,[a] Ellen B. McCullough,[b] and Thomas Assefa[c]

August 2025

Preliminary—please do not circulate widely—see latest version here

Tracking labor and productivity of self-employed workers almost universally relies on workers' recollections of events and perceptions of their duration. While the frontier of neuroscience investigates the inter-dependence of memory and time perception to one's activities or context, errors in labor and time use measurement are not well understood by economists and often ignored altogether. We provide novel experimental evidence on economically and statistically significant recall errors in time allocation to economic activities, household chores, and other activities in a sample of self-employed women in rural Ethiopia. Relative to traditional day reconstruction recall surveys (the control), high-frequency, randomly-timed phone surveys about activities at the time of the call (the treatment) identify short episodes of work that go unreported in recall surveys. However, the more economically active women are, the larger their errors are in over-estimating the duration of their work. As a result, respondents' implied labor force participation and productivity are both under-estimated, with important implications for measuring labor supply elasticities.

JEL codes: D1, J1, J2, J24, O1

keywords: time use, measurement error, labor productivity, self employment, phone surveys, recall error, methodological experiment, household production

*Corresponding author: Tamara McGavock, *mcgavock@grinnell.edu*. [a]Grinnell College, Department of Economics. [b]University of Georgia, Agricultural & Applied Economics Department. [c]International Food Policy Research Institute.

We thank Girum Abebe, Chris Barrett, Lori Beaman, Andrew Dillon, Jessica Goldberg, Daniel Hamermesh, James Heckman, John Hoddinott, Joseph Hotz, Chris Udry, and Nicholas P. Magnan, as well as participants at the IPA/Northwestern Methods & Measurement Conference 2021, American Economic Association meetings 2023, CEPR-STEG NYU-Abu Dhabi workshop 2024, and numerous other seminars and conferences. To our intrepid survey staff at EconInsight Center for Development Research, we are indebted for your outstanding execution of a highly complex study protocol, and during unprecedented times —we especially thank Dr. Tigabu Getahun, Abiy Samuel, Yordanos Tenaw, and our incredible field and call center enumerators. We were well-supported in the field by Libse Sumete, Asnake Tsegaye, and other staff at Organization for Rehabilitation and Development in Amhara (ORDA). For financial support, we thank Livelihoods for Resilience Activity (LfR), a CARE Ethiopia project; Strengthen PSNP4 Institutions and Resilience Development Food Security Activity (SPIR), a World Vision project; Grinnell College, especially for the Jack ('39) and Lucile Hanson ('40) Harris Faculty Fellowship; and the University of Georgia. We are grateful to many at LfR, SPIR, and CARE, especially John Meyer (USAID, formerly LfR), Michael Mulford (SPIR), Elisabeth Farmer (LfR), and Vidhya Sriram (CARE International) for thought partnership and support of the project's vision. We thank Feteh Demmelash, Ruth Sleshi, and the gender teams at LfR and SPIR for invaluable insights in planning and piloting. It brought us joy to work with excellent student research assistants at Grinnell College: Saketan Anand, Hyein Cho, Samuel Heller, Amna Kapur, Hang Li, Tate McClean, Shloke Meresh, Anshul Tambay, Simon Taye, and Maximilian Worth; and at the University of Georgia: Imani Matos.

"Time past and time future / What might have been and what has been / Point to one end, which is always present." -T.S. Eliot

1 Introduction

Recall data generate measures of labor supply, wages, productivity, women's economic empowerment, and other key economic variables tracked by governments everywhere and used in a majority of applied economics work. Data on hours worked (the most commonly collected time use data) typically rely on survey respondents' recall of labor performed over days, weeks, and seasons. Errors in memory or in estimating the passage of time—topics at the cutting edge of neuroscience—will generate measurement error in any variable that uses time use data as an input (e.g., wages) and can impact estimation of key parameters like labor supply elasticities. Recall errors may be especially acute among low-income households, given high participation rates in sectors and jobs where hours worked are unpredictable, occasional, and/or erratic (e.g., self-employment, seasonal farm labor, and casual work). Firms and salaried workers may report 40 hours per week even when hours vary substantially. Errors in recall may particularly arise for anyone who spends a substantial share of their time multi-tasking, especially when caring for dependents or when engaged in household production (Becker, 1965; Grossbard-Shechtman, 1984), which happens often with remote work. In low-income countries, limited numeracy and low ownership of timekeeping devices likely exacerbate errors in recalling time use, and administrative records are not available or comprehensive as an alternative source of time use data. Even when administrative data are available, they rarely include hours worked. Despite its likely prevalence, non-classical measurement error in time use data—and therefore of hourly productivity or wages, too—receives surprisingly little attention in the economics literature.[1]

Mis-measurement of labor supply and time allocation could partially explain major open questions of concern for economic policy, such as large cross-sector productivity gaps between agriculture and other sectors (Gollin et al., 2014; Adamopoulos & Restuccia, 2021; McCullough, 2017). Billions of dollars—about 2% of each country's GDP, but less than committed under the African Union—are spent annually on efforts to improve productivity in African agriculture, with mixed evidence on their impacts. Similarly, mis-measurement of labor inputs on both the intensive and the extensive margins of labor supply could bias estimates of the treatment effects of other expensive anti-poverty programming (e.g., the

[1]For example, only one of the 10 most cited articles in the *American Economic Review* from 2021 about "wages" or "productivity" mentions measurement error in hours, ever, despite the fact that recall data were used to calculate hours (and therefore wages) in six of them. Based on a Google Scholar search from October 12, 2022.

Earned Income Tax Credit, EITC, Wilson, 2020), labor supply elasticities, and trends in gender/racial/ethnic wage gaps (Borjas & Hamermesh, 2024) if the data are not of high quality (Lachowska et al., 2022). Since measurement errors in recall data are probably correlated with other characteristics that affect these outcomes, this non-classical measurement error implies that causal estimates may be biased in any direction (Millimet & Wang, 2011; Bound et al., 2001; Heckman, 1993) and cannot be corrected with instrumental variable methods. Similarly, errors in classifying labor force participation can bias estimated effects of program impact.

We provide novel experimental evidence on economically and statistically significant recall errors even in gold standard methods for measuring time allocation to economic activities ("labor supply"),[2] unpaid household work including childcare and chores, and other activities in a sample of women in rural Ethiopia. Our call center conducted brief, randomly timed interviews by mobile phone during which respondents reported on their activities at the time of the call (ATOC) five times daily for seven to twelve days, a method we adapted from psychology's experience sampling method (Csikszentmihalyi, 2014). Importantly, our primary measure of time allocation for women in this ATOC survey group does *not* rely on *any* estimate of duration of their activities—rather, for each woman we simply identify the share of calls during which she reported being engaged in a given activity category at the moment when the phone rang.[3]

During the same period, respondents in the same village who were not randomly assigned to these high-frequency phone surveys reported on their activities during one traditional "day reconstruction" (DR) interview about the previous day (Kahneman et al., 2004), which was conducted in person. This method, sometimes referred to as a 24-hour recall survey, is not only the status quo in research on time use across the developing and the developed world, but is considered by many to be a "gold standard," and thus our control methodology. We construct daytime[4] activity shares for these respondents as the share of hours—that overlapped with the hours when ATOC surveys were conducted with the other

[2]We use the International Labor Organization's definition to characterize activities as economic if they relate to paid or unpaid employment. This could include wage employment, or self-employment on a household farm or non-farm enterprise, even if items are produced for barter or crops are grown for self-consumption (Hussmanns et al., 1990).

[3]Of course, to the extent that women misreport or misrepresent their activities and/or that some calls were not answered at their randomly appointed time, these measures of time allocation would be biased. We provide substantial empirical evidence in Section **??** suggesting that such errors are limited in this setting.

[4]Phone calls were placed between 7am and 6pm only, which captured the majority of our respondents' waking hours given limited access to electricity (only 9 of 46 villages had any access) and a consistent 12 hours of sunlight around the year (due to Ethiopia's position close to the Equator). We also collected information about early morning, late evening, and overnight activities from ATOC respondents through recall but our primary estimates do not rely on these data (see Appendix **??** for details).

respondents—during which she reported performing a given category of activity. Comparing activity shares across women assigned to the ATOC and the DR groups, we identify substantial differences in time allocation measures elicited by the two different survey methods.[5]

We find that average reported labor supply to "economic activities" is twice as large among the DR (control) women, who recalled the previous day during a single interview, relative to the ATOC (treatment) women who received frequent, randomly timed phone calls and reported on the activities they were performing at those times. That is, as a result, constructed measures relying on recall data can significantly understate women's labor productivity (or wages) across the entire distribution (see Figure 7). At the same time, our results suggest that the unpaid household work performed by women in our study population is even more time-consuming than traditional recall data show (e.g., firewood and water collection, slow cooking over open flames), and we identify instances of childcare that go unreported. Together, our results suggest that women's total time allocation to paid and unpaid labor is under-estimated in the day reconstruction surveys by 12 percent.[6]

Put differently, our results imply that the "double burden" faced by women—rooted in deep cultural norms about gendered divisions of unpaid labor—is possibly even larger than previously understood. This burden continues to pose large constraints to women's economic participation. Reducing this burden, and identifying best practices for measuring it, should thus be a key target of Sustainable Development Goal (# 5) that motivates a large share of anti-poverty interventions to increase women's labor supply. During the COVID-19 pandemic, the implications of this burden for women's economic opportunities became more apparent than ever in developing and developed countries alike, and it can significantly harm women's mental health; progress on gender pay and wealth gaps; and the quality of investment in children's human capital (Hotz & Wiswall, 2019).

Using unique features of our experimental design, we provide empirical evidence strongly suggesting that the high-frequency phone surveys which document activities at the time of the call (ATOC) capture time allocation more accurately than day reconstruction (DR) recall surveys (see Section **??**). For example, selection effects from unanswered or dropped calls—less than 10% of all calls—*cannot* explain the differences we observe between the ATOC and DR respondents' activity shares. We confirm this first with a bounding exercise

[5]We use individual-level activity shares for our main analysis for ease of interpretation and because it was the main outcome variable defined in our pre-analysis plan, AEARCTR-0005961. Readers may also be interested in results using the disaggregated data, allowing us to control for and estimate heterogeneity in survey mode effects by the time of day or enumerators. See Section 5.5 for this analysis.

[6]This calculation comes from adding the coefficients in Table 2 for economic activities (-3.34 percentage points) and household work (9.04 percentage points) and dividing by the total share of time spent on these two activities as measured in the control group whose time allocation was elicited during a day reconstruction interview (0.48, or 48% of their time between 7am and 6pm).

that calculates the minimum and maximum possible activity shares for each respondent and each category. Additionally, data from our detailed "callback" protocol—triggered in the event that an intended call was unsuccessful at the randomly appointed time—show that women's reports of their activities at the time of a missed call are not significantly different from those calculated using the calls that they answered, with the exception that travel is more likely to be reported during a callback about a missed call than during successful calls. This is consistent with the previous evidence that shorter recall periods improve accuracy.

We also provide evidence that other sources of bias, including Hawthorne, survey burden, phone endowment, or enumerator effects, are unlikely to have affected our ATOC activity shares. Therefore, we conclude that the differences we observe across survey types are attributable to recall errors arising in the DR data. That is, the status quo methodology likely generates flawed estimates of time allocation across activities.

We explore three mechanisms that can explain these errors in the DR data: 1) under-estimating the occurrence of short-duration activities; 2) mis-perceiving the length of longer-duration activities, conditional on reporting their occurrence correctly; and 3) differentially reporting secondary activities, such as passive childcare, which often take place alongside other activities that may be more salient to the respondent. Regarding the first mechanism, women in the ATOC group are more likely to report the occurrence of any economic activity, implying that 24-hour recall surveys may under-estimate women's labor force participation at the extensive margin among populations like ours. Importantly, this effect is not completely driven by the minimum allowable activity duration of 30 minutes in the DR recall data. Regarding the second mechanism, we find suggestive evidence that women over-estimate the duration of their activities that take up more of their time. Among those women who report any economic activity, the most economically active women in each survey group —those who report the highest shares of economic activities within each method —have the largest gap between their labor supply estimates (economic activity shares). That is, the error in measuring duration of economic activities appears to be systematically related to the duration itself, which implies non-classical measurement error in estimates of labor supply on the intensive margin. Regarding the third mechanism, women in the ATOC group report a secondary activity—mostly childcare—only 29.5% of the time compared to 35.0% of the time in DR interviews, suggesting that time shares that incorporate both primary and secondary activities may overstate time allocated to secondary activities. However, using detailed data on *children's* activities[7] and/or women's companions at the time of the call as reported by the female respondents, it does appear that women are often responsible for

[7]We also find statistically significant differences in children's reported activities in the ATOC data compared to the DR data. Specifically, girls are more likely to be playing or doing school-related activities.

children's care even when they do not report it as an activity. The routine simultaneity of childcare and other tasks points to the complexity of women's schedules. Since passive supervision of children often generates interruptions to tasks being actively performed, our results suggest that childcare or supervision that goes unreported can affect the *quality* of labor supplied to economic activities on the one hand and investments in children's nutrition and education on the other.

We contribute to a small but important literature on measurement of time use specifically and, more generally, on non-classical measurement errors in microdata, which pose major, yet often underappreciated, challenges to impact evaluation (Millimet & Wang, 2011; Abay et al., 2019b,a; Bound et al., 2001; Dinkelman & Ngai, 2022). Surprisingly little is known about the nature of measurement errors that afflict time use data in particular. Indeed, the user guides for the American Time Use Survey (ATUS) —the survey instrument used by the U.S. Current Population Survey (CPS) for measuring time use— and the Multinational Time Use Study (MTUS) never use the word "bias" apart from possible sampling errors due to survey non-response.[8] Previous work on measurement error in time use in both high-income and low- and middle-income countries has found, like we do, evidence that during recall surveys respondents may significantly over-estimate hours worked (Borjas, 1980; Bound et al., 2001; Borjas & Hamermesh, 2024). Moreover, the recall errors in labor supply increase with the length of the reference period and the time elapsed between the reference period and the interview (Arthi et al., 2018; Gaddis et al., 2021; Frazis & Stewart, 2010; Heath et al., 2021; Barrett et al., 2019).[9] By shortening the recall period to a single instant, our findings show that this recall error still appears in the gold standard 24-hour day reconstruction surveys.

These problems in estimating time use are particularly problematic for those engaged in seasonal or irregular work, where hours worked can vary significantly across days, weeks, months, and seasons.[10] In Sub-Saharan Africa, smallholder agriculture accounts for more than 56% of labor force participation and, together with off-farm self-employed or informal labor (an additional 30% or more), up to 65% of GDP (Davis et al., 2017). Issues with standard recall data therefore affect estimates of labor productivity for a large share of the continent's population. Moreover, these biases are likely to be systematically related to occupation, job types, hours worked, income/wealth, and education, generating biases of

[8] See ATUS User Guide June 2022 and MTUS User Guide October 2020.

[9] Specifically, Arthi et al. (2018) compare weekly interviews (both phone and in-person) to seasonal in-person interviews, and find that respondents report about four times more labor inputs per person/plot when interviewed seasonally. Gaddis et al. (2021) find that seasonally reported per person/plot farm labor inputs are 10% higher than weekly reported labor inputs.

[10] In the developing world, the general practice is to ask respondents to recall a "typical" day, week, month, or season (Gaddis et al., 2021), and Frazis & Stewart (2010) find that conditioning on participation in a given week can help.

unknown sign from non-classical measurement error (NCME).

We also contribute to the literature on women's economic and social empowerment by providing new evidence on women's time burden, possible corrections for errors in measuring it, and flexible guidelines for collecting panel data on time use without incurring the costs of repeated field visits to remote communities. Cross-culturally, women consistently perform the vast majority of unpaid household tasks and care work —cooking, cleaning, and childcare, of course, but also even more physically demanding and dangerous tasks such as collecting and carrying firewood and water on their backs (or shoulders or heads) over long distances (Blackden & Wodon, 2006; Jayachandran, 2015; Dinkelman & Ngai, 2022; Blau et al., 2017; Goldin, 2014; Folbre, 2006). Poor women—who cannot afford to outsource unpaid household work and have limited access to labor-saving technology, especially in the developing world (Dinkelman & Ngai, 2022) —are frequently targeted by policies and programs that seek to increase women's participation in economic activities, from the Earned Income Tax Credit (EITC) in the United States to the increasing number of "graduation-from-poverty" programs with training and work requirements in the developing world.[11,12] While some programs have been shown to increase women's time spent in economic activities, as measured through study participants' recall time use data (Banerjee et al., 2015; Bandiera et al., 2017),[13] our results suggest that previous estimates of hours worked may be over-stated *and* that ultra-poor women's time spent on household chores may be understated. Little is understood about how graduation-from-poverty programs affect women's allocation of time to unpaid household work and possible re-allocation of household chores to other household members, in part due to the difficulty and expense of collecting high-quality time use data. Panel data on women's time allocation are almost non-existent. Our methodology is easily adaptable for collecting not only women's time use data but also the activities of their spouses, children, and other family members.[14] These data would be far too time-intensive to collect during a traditional in-person or phone-based recall interview covering time periods, but in the moment they are relatively easy to report on.

[11]For example, the Productive Safety Net Program (PSNP) in Ethiopia requires most participants to contribute labor to public works programs, effectively paying wages for their work on roads, bridges, and other public works projects during the agricultural off-season. Often, this labor is contributed by women.

[12]Safety net programs frequently target women because of strong evidence that the welfare gains from transfers are larger when they are received by women, e.g., Duflo (2003); Qian (2008).

[13]See also an extensive literature in labor economics on the EITC.

[14]In follow-up work to the experiment described in this article, we have collected high-frequency phone survey data from husband and wife pairs in a similar population—about their own activities and their (knowledge of their) spouse's activities—and find substantially different reports on time spent on economic activities, household/care work, and leisure/personal/social time. Specifically, men underestimate women's participation in economic activities while women over-estimate their husbands'. Both spouses significantly under-estimate the time that their partners report spending on leisure activities, but the magnitude is largest when wives under-report on their husbands' leisure time relative to his own reports.

Finally, to our knowledge, ours is the first study to adapt experience sampling practices for the purposes of understanding recall errors in time use data in a developing country setting (Seymour et al., 2020) and one of only a few studies to test time use data collection methods in a randomized survey experiment with a large enough sample to detect statistically meaningful differences (Field et al., 2022; Kilic et al., 2024). We join a growing list of studies that use high-frequency phone surveys to improve data collection in developing countries (Zezza et al., 2022). The spread of phone ownership, high-quality network coverage, and experience of COVID and natural disasters have made phone surveys a particularly attractive tool for remote monitoring and proactive crisis response, as well as for generating high-dimensional panel data to track changes in outcomes over time. For the developing world this is especially important, as high field survey costs, low population density, and frequent migration can limit the size and generalizability of many panel datasets.

Full time allocation data beyond labor supply are rarely collected by economists and policymakers, in part because the challenges and costs of doing so remain high. We contribute to a small set of papers developing innovative methods to improve the accuracy and lower the costs of collecting time use data (Field et al., 2022; Kilic et al., 2024; Bardasi et al., 2011). The supplementary appendices and website accompanying this article will provide survey instruments and guides for implementation that others can use in implementing experience sampling methods for measuring time use and other outcomes in the developing world.[15] We also provide information about the differences in costs between high-frequency mobile phone surveys and traditional in-person surveys.

In Section 2, we describe our sample and experimental design. Section 3 provides summary statistics on time use shares, survey compliance rates, and balance across treatment and control groups. Section 4 presents experimental results on the impact of survey method on measurement of our primary set of women's time use outcomes. In Section 5 we provide evidence showing that these differences are not driven by errors in the ATOC method and are thus attributable to issues with the DR method, which we characterize in Section 5.5. Section 6 describes the implication of our findings for measuring key development outcomes on both the extensive and intensive margins, including simulating the implications of our results on recall bias for measured impacts of a randomized cash transfer program on labor supply outcomes (Banerjee et al., 2021). Section 7 concludes by discussing the implications of our results not only for future work—including cost tradeoffs related to survey design—but also for understanding potential biases in existing measures of time use, labor supply outcomes, and treatment effects.

[15] Forthcoming.

2 Sampling and Study Design

We recruited a sample of women, conducted in-person baseline interviews to obtain basic information about the households, and then conducted interviews about time use via two different survey modes during a seven-day validation period. Before explaining the treatment groups and study design, we describe our sample of respondents.

2.1 Sampling and Field Survey Logistics

Our study population is women whose households participate in graduation-from-poverty programming in rural Amhara, Ethiopia. These programs, sponsored by two international NGOs in partnership with a single local NGO, provide supplementary support for beneficiaries of the government's flagship public safety net program—the Productive Safety Net Program (PSNP). Well in advance of our study period, the NGOs recruited all PSNP-eligible households into "Village Economic and Savings Associations" (hereafter, VESA groups), which comprise the main delivery infrastructure for the NGOs' graduation-from-poverty livelihoods interventions (see Appendix B for more details).

Our sample from this population was formed in two stages. First, we randomly selected sixty VESA groups from the full list of VESA groups that were formed by the NGOs with whom we cooperated. Second, from the full roster of members participating in each sampled VESA group, we recruited every unique household to participate in our study.[16] Within each household, the female VESA member (if applicable) and/or the female spouse of the VESA member became the respondent in our study. We note that in order to ensure that participation in our sample was not selected on the household's (or respondent's) prior phone ownership, we provided each respondent with a new, basic phone.[17]

Each VESA group was randomly assigned to one of eight survey waves (see Figure 1A),[18] stratifying on the estimated distance from the VESA to the nearest market town, to help in smoothing field logistics and to ensure that the remoteness of VESAs was balanced across seasons. Of our full intended (pre-registered) sample of 60 VESAs and about 800 households, we completed data collection for only 75% (6 waves) before we had to cease operations due

[16] VESA groups often contained more than one member of a given household, most commonly both the male household head and his female spouse.

[17] See more on phone provision and tests we perform to ensure that being provided a phone did not affect women's time use in Section 5.4.

[18] Due to the desire to observe and control for seasonality when measuring time allocation, VESAs were assigned to one of eight "waves," each a period of about forty-five days of continuous fieldwork. This design feature also offered staffing benefits, allowing us 1) to provide field enumerators periodic breaks from the grueling field work protocol, and 2) to staff our study with a small group of high-quality enumerators over the full (intended) 11-month study.

to the spread of civil conflict to our study area. More details on the sampling frame appear in Appendix B. Our completed sample includes women respondents, representing 646 unique households across 46 VESA groups.

Within each VESA, we collected data for nine to twelve days. First, every household's primary female respondent participated in an in-person baseline interview to report demographic information as well as labor participation in the previous 7 days, household asset ownership, and an abbreviated women's empowerment index. The number of unique households in the VESA group and field team capacity determined the length of the baseline interview period in each village (two to four days). After all baseline interviews were complete in a VESA, one enumerator remained behind in that VESA for a seven-day "validation period" (see Figure 1B), while the other enumerators re-located and conducted baseline interviews in the next VESA in the wave.

During the seven-day validation period, each respondent participated in a research protocol that corresponded with her pre-assigned treatment (balanced within each VESA group), as described in Sections 2.2 and 2.3.

All assigned protocols called for one in-person "validation interview", which occurred on a randomly assigned day during the validation period, stratified on treatment group. However, the content of these in-person interviews varied according to treatment assignment. One-third of the sample was assigned to the ATOC group. These women received high-frequency phone calls during the validation period, in which they reported their current activities. In contrast, their in-person validation interviews focused on topics —such as diet diversity, sanitation practices, and financial transactions —which we refer to as "placebo" content. The remaining two-thirds of the sample followed a traditional day-reconstruction (DR) protocol, in which participants reported on time use during their in-person validation interviews. Within this DR group, half also answered high-frequency phone calls about the placebo topics, unrelated to time use; we refer to these women —who make up one-third of the total sample —as the DR phone placebo group. They serve as a direct comparison for the ATOC participants. The final one-third of sample received *no* phone calls at all; their in-person validation interviews covered both time use and placebo topics via traditional recall methods. We refer to these women as the DR control group, as they serve as the baseline for comparisons with the DR phone placebo group in analyses described in Section **??**). Basic phones were provided to women assigned to the two high-frequency phone survey groups (the ATOC and the DR phone placebo groups) at the end of their baseline interview, while women in the DR control group were given a phone at the end of the second in-person interview at the request of the NGO to ensure fairness across women in the VESA (see

details in Appendix B).[19] For each household, all survey activities were completed within 12 days.

2.2 In-Person Day Reconstruction Recall Surveys—the Control

The day-reconstruction (DR) groups received a single in-person interview on a randomly selected day during the seven day validation period following completion of baseline interviews. The respondent was not informed in advance which day this visit would occur. During the interview, the enumerator used a traditional 24-hour time-use "day reconstruction" recall module, a commonly used survey instrument for collecting time use data (Frazis & Stewart, 2010). Respondents were first asked to describe verbally their activities during the previous day, in the order in which they occurred, while the enumerator made hand-written notes. Then, referring to his notes and asking for verification or additional input as needed from the respondent, the enumerator "reconstructed" the respondent's day by mapping her primary and secondary activities into 30-minute periods covering the day. The data were recorded digitally using a CAPI program, with 36 standard activity codes.[20] We categorized these activities into five groups: economic activities, social (or religious) activities, household chores, care of others, and personal/travel.

Our phone-based time use surveys, described in the next section, covered more than just the activities at the time of the call, so we also included analogous questions in the in-person survey. Some of these are commonly part of time use surveys (e.g., ATUS) but some are not. When women indicated passive or active care of others (children and/or elderly and sick individuals), enumerators recorded the roster ID(s) of those being cared for so that we could link the woman's care activity to demographic information from the baseline interviews. If the recipient of care was not a household member, we recorded information about the respondent's relationship to the person they were caring for. Enumerators also asked the respondent to name all other companions who were with them during each time period within the reconstructed day and their general location during each time period (e.g., home, a neighbor's house, church, the market). For time spent on non-agricultural self-employment activities, we built an enterprise roster to link the activity back to the respondent's baseline information about that business. After reconstructing the respondent's previous day, we collected some additional information that is not usually standard in a time use survey. For children under 5, we asked who was responsible for supervising the child (if not the

[19] No phone calls made to these respondents are reported on in this article. However, we did call some of them on these phones up to two times after their participation in the time use measurement experiment was complete, for reasons reported in separate articles.

[20] All survey instruments for in-person and phone interviews are publicly available on our accompanying website.

respondent) in each time period, and for children ages 5 to 12 we asked the respondent to report (or guess) their activities and whereabouts for each time period.

For reasons that we explain in detail in Section **??**, the control group was actually comprised of two equally sized subgroups, both receiving the same in-person day reconstruction (DR) interview module on a random day. One of these DR control groups (hereafter the DR phone placebo group) received a phone at baseline and multiple calls each day about placebo content *unrelated* to time use (diet diversity, sanitation practices, and financial transactions). The other DR group (hereafter DR control group) did not respond to any phone surveys but answered questions in person about diet diversity, sanitation practices, and financial transactions. For equity purposes, respondents in this DR control group were provided a phone after completing all study activities. Together, these DR groups comprised two-thirds of the sampled households in each VESA. When we refer simply to the DR group throughout this article, without referencing either sub-group specifically, we include pooled data from both of the DR sub-groups, since our estimates show that their time allocations are not statistically significantly different from each other (Figure C3).

2.3 High-frequency Phone-Based Surveys—the Treatment

Respondents assigned to the treatment group received five randomly timed calls per day from highly trained and experienced enumerators based in our call center in Addis Ababa. During each phone interview, respondents reported on their current activities ("what are you doing right now?"), whereabouts, and companions in what might be called a "momentary recall" period *at the time of the call* (ATOC). Additionally, we asked about the current activities of each child aged 5-12 as well as who was currently providing care for each child under 5. Enumerators also asked these questions about any relevant "bounded recall" periods—"what have you been doing since...?" with clearly defined bounds (e.g., "we last spoke with you" or "you woke up").

The call center operated from 7am to 6pm, covering eleven of the twelve hours that are consistently daytime in Ethiopia, given its position close to the Equator. Calls were randomly assigned on a pre-programmed schedule to each household within five strata of equal duration (two hours and fifteen minutes long—see Figure 1C). Within each stratum, calls to respondents were assigned a random order, without regard for how long it had been since the respondent was last contacted. Enumerators were instructed to call as close as possible to the appointed time, but were given some buffer time to use when necessary. If the respondent did not answer, enumerators immediately re-dialed the respondent's number, making up to three additional attempts to reach the respondent. If the respondent did not

answer after all four dials, the enumerator assigned the respondent to the next available callback time from a list and then proceeded with other calls as scheduled. Each phone interview took approximately five to eight minutes to complete. To incentivize respondents to answer calls and to remain in the sample, we provided a small amount of phone credit for each call that they completed, but not until after the final call at the end of the validation period. We reminded respondents of their accumulated balance of these credits occasionally.[21] For more logistical details, see Appendix B.

2.4 Constructing the Outcome: Activity Shares

Our pre-specified outcome variable of interest is an activity share: for each respondent, we measure the share of time for which the respondent reported performing a given activity during time periods when the call center was operating (7am to 6pm). For the ATOC group, we construct an activity share y_{iva} as the fraction of answered calls during which respondent i in village v reports performing activity a.[22] Note that this measure does *not* use any information from the respondent about the duration of activities. We consider, and dismiss, concerns about possible bias in this measure arising from unanswered calls in Section **??**. For the DR group—who "reconstructed" the previous day with the assistance of an enumerator by recording activities performed during each 30-minute reference block—we generate DR activity shares by dividing the total number of 30-minute periods when the respondent reported doing activity a within the call center's operating hours (7am to 6pm) by the total number of hours covered (11).[23] Since randomization was at the individual level, comparing activity shares relies on comparisons across individuals.

[21] Receiving our calls did not cost them anything, so these phone credits could be used for them to make other calls. To ensure that these incentives did not affect their time use or cause them to drain their batteries, we made the transfers after the final call with the respondent, at which point we reminded them that the phone was theirs to keep. Initially our protocol was to tell each respondent her new balance at the end of each call. In piloting, however, respondents told us that they found this off-putting and too transactional, that they were happy to receive our calls and help with the research, and that occasional reminders were sufficient. In practice, women seemed to enjoy participating in the survey. In fact, to our delight and amusement, by the end of their survey period, it was not uncommon for respondents to end phone calls with the enumerators by saying something to the effect of "I love you and we will talk soon."

[22] We group activities into five broad categories: economic activities (including self employment in agriculture and livestock cultivation, self employment in a non-farm business, and wage employment), social activities (including church and community events), household work, care-taking (for children, elders, and/or the sick), and other activities (e.g., personal care, leisure, travel, sleep).

[23] Note that 15- or 30-minute blocks are standard for developing country surveys on time use, but the ATUS allows respondents to report shorter durations and does not require rounding to the nearest 15-, 30-, or 60-minute increment.

3 Summary Statistics and Survey Compliance

3.1 Baseline Characteristics

Table 1 reports baseline characteristics for the sample. Our respondents range in age from 20 to 75, with an average age of 38. The majority of respondents are between 30-39 years old. 96% of households have at least one child member (under 12), and more than half of households have at least three child members. The average respondent household has 5 members and 2.86 children. More than 80% of our respondents are members of a VESA group themselves, while 18% of them are not but their spouse is. Nearly 90% of women state they *always* attend their VESA group meetings. As shown in Table 1, 28.2% of the respondents have a child under 24 months of age and 27.7% of them are breastfeeding a child. Almost all (98.8%) of the households participate in farming, and 70% of them report owning the land they farm. The average total value of livestock assets owned by a household is 19,487 ETB ($US 373.37).[24] Only 31.1% of households participated in wage labor during the previous year (apart from their contributions to PSNP public works projects). Wage workers report an average daily wage of 106.38 ETB (2.04 USD) for days spent on work not part of the PSNP.

3.2 Balance on baseline characteristics

Table 1 shows differences between average characteristics for the primary female respondents across the ATOC and DR groups, with p-values for a test of differences in means in Column 4. Figure A3 uses coefficient plots to show that our sample is balanced on observable baseline characteristics related to age, family composition, time literacy, and participation in economic activities. The only variables where a difference seems apparent are the education variables (both an indicator for whether the respondent has any school and the number of years of formal education), which appear slightly larger for the DR phone placebo group relative to the ATOC group and the DR control group. The difference for the binary formal education variable has a p-value of 0.178 thus it is not statistically significant. The difference for the number of years of formal education (unconditional on receiving any) is significant at the 95% level. This difference can be attributed to one respondent in the DR placebo group who had completed high school (13 years), whereas over 90% of the sample had completed fewer than 4 years of schooling. Our results are robust to dropping this respondent from the analysis.

[24] we convert current ETB into constant $USD PPP using the exchange rate from 2020.

3.3 Call Success Rates and Metadata

Figure A1 shows call success rates across the day for both the treatment group (ATOC, who received phone-based time use interviews about activities at the time of the call) and the placebo phone group (DR phone placebo, who received calls about other content). The graph shows that over 90% of calls were answered on the first try. The small lavender regions in Figure A1 show that immediately attempting to call the respondent three more times was rarely successful. When a respondent could not be reached, the enumerator recorded their perceived reason for a callback. Figure A1 shows that poor network coverage accounted for 50% of the instances when a callback was required, and phone battery issues generated another 10% of callbacks. We further analyze missed calls and the robustness of our treatment effects to assumptions about these missed calls in Appendix C (section C.1.1).

3.4 Summary statistics of time use activity shares

Figure 2 shows primary and secondary activity shares by activity category across the entire sample, pooling the DR activity shares and the ATOC activity shares (from the time of the call only). These data imply that women in the sample spend more than 40% of their time on unpaid household chores, the most frequently reported primary activity. Care of others is the least likely activity to be reported as a primary activity but it is the most frequently reported *secondary* activity. Social activities—which include eating with family, drinking coffee with neighbors, worshipping at church, and attending community meetings—are the second most reported primary activity, covering about 25% of women's time in the pooled sample. Women's time spent on paid work or self-employment activities is less than 10%. Breaking each activity category down further, Figure A4 shows primary and secondary activity shares by activity subcategory. Cooking takes by far the most time within household chores, and eating or spending time with others are the most frequently reported social activities. The vast majority of the time that women report as care for others is for children, not adults, and this time is split relatively evenly across active and passive childcare.

4 Results: Activity Shares Differ by Survey Type

In this section, we present results comparing activity shares, our primary outcome of interest, from women's reports at the time of the call (ATOC) to those from the day reconstruction recall surveys (DR). We find that the differences across these reporting methods are statistically significant and meaningful in magnitude.

Figures 3A and 3B show summary statistics for respondents' primary and secondary activity shares by treatment group, respectively. The activity shares for the treatment group—the ATOC time-use phone survey group (right, red)—suggest a treatment effect across several activity categories when compared with the in-person recall activity shares from the placebo phone group (DR). The ATOC activity shares, based on reports at the time of the call, are lower for economic activities and social activities than those from the in-person recall group. In other words, women report performing economic activities less often when they are asked in-the-moment during a phone call than they do when recalling the previous day. On the other hand, respondents in the ATOC group report performing household chores and travel/personal activities more often relative to those in the DR group. The difference in the "travel/personal" category is mostly driven by an increased likelihood of reporting travel (e.g., walking to their fields or the market) for the ATOC group, but we also observed personal care and rest more often for the ATOC group than for the DR group. For secondary activities, ATOC respondents are slightly more likely to report household chores but are significantly less likely to report caring for children or others when compared with the DR group. Figure 4 shows that these differences in secondary activity shares result in differences in reports of multi-tasking (defined as reporting any secondary activity), with respondents in the control group reporting multi-tasking 30% of the time during their in-person recall surveys while ATOC respondents report multi-tasking just over 20% of the time. As shown in Figure 2, the reported secondary activity is usually caring for others, mostly children (Figure A4). This is true across survey methods.

We formally test the null hypothesis that the average activity share of respondents in the ATOC group ($T_{iv} = 1$) is equivalent to the average activity share of respondents in the DR group ($T = 0$). For each activity category a, we estimate the following OLS model:

$$y_{iva} = \alpha + \beta_a T_{iv} + X'_{iv}\gamma + \nu_v + \epsilon_{iva}, \tag{1}$$

where T_{iv} is the treatment assignment dummy for individual i in VESA group v, X_{iv} is a vector of individual- and household-level controls from the baseline interview (which we omit from our main results), and ϵ_{iva} is a normally distributed error term reflecting a combination of individual-, VESA-, and activity-level unobservables. We include VESA-level fixed effects ν_v to account for characteristics common across individuals in the same VESA, such as remoteness of the village, the presence of electricity or a water tap, accessibility via a paved road, or impacts of commonly received program interventions, which could affect not only the average time spent on various tasks across women whose households are in the same VESA group but might also be related to their recall errors. Standard errors are clustered at the VESA level, and there are 46 clusters. Since treatment was randomly assigned and

stratified on VESA, we estimate within-VESA treatment effects. For simplicity, we refer to VESAs as "villages" hereafter, since each group is located in a unique village, unless we are discussing specific activities related to VESA group participation.

The coefficient β_a captures the difference in the average shares of time spent on activity a between those assigned to the ATOC group and those assigned to the DR groups, which is visually represented as the difference in the height of the bars in Figure 3. Figure 5 reports the β_a coefficients for each activity category (these results are also displayed in Table 2). The differences by treatment group in primary activities shown in Figure 3 are statistically significant for all primary activities except childcare. Specifically, the blue coefficient furthest to the left in Figure 5 shows that women in the ATOC group are 3.3 percentage points less likely to report performing an economic activity (i.e., the coefficient β_a is -0.033). Since the average share of time spent on economic activities reported during recall surveys by the control group is about 8.3 percent, this coefficient implies a decrease of 40 percent compared to the control group.[25] The coefficient is statistically significant at the 99% level. Activity shares for social activities are also smaller for ATOC respondents (by 13 percentage points, a 45 percent decrease). On the other hand, ATOC respondents report more household work (9.0 percentage points, a 23 percent increase) and more time spent on travel and personal care (also 7.8 percentage points, a 47 percent increase).

5 Interpreting time use differences: evidence on potential mechanisms

Our analysis, which includes a bounding exercise to account for missed calls, strongly suggests that most of the differences in activity shares between treatments can be attributed to recall errors. In this section, we present a comprehensive list of possible explanations for differences in time use measured between the two treatments and address them here.

5.1 Measurement error framework

Survey methods affect measurement of all measures collected from human respondents (Bound et al., 2001). Generally, collecting survey data from respondents relies on multiple processes, including respondents' ability to understand a question, to retrieve information from memory, to ensure that the retrieved information corresponds with the requested information, and to communicate a response to the interviewer (Tourangeau, 1987, 2018). The

[25] Readers may wonder about the extensive margin (reporting any work) as well, which we take up in Section 6.

enumerator's encoding of answers can also introduce another layer of survey measurement error. With respect to continuous measures of time use, many possible errors identified in the survey methodology literature are potentially relevant for either the ATOC method, the DR method, or both. These errors can generally be classified into four categories. The first category is recall errors, which generally relate to the respondent's ability to retrieve information about events that happened in the past. The second category relates to strategic misreporting of information by respondents. The third category of measurement errors relate to sampling errors. The final category of methodology-related measurement error relates to a change in true activities of the respondent induced by the survey methodology.

When we compare the two measures of time use devised between the two method, both the ATOC and the DR methods include all errors in measurement that apply. In the remainder of this section, we discuss each of these four measurement error mechanisms in more detail. We report additional analysis that uses features of our research design to identify the error types that could explain the differences between the ATOC and DR activity shares that we observe. We first discuss strategic misreporting, sampling errors, and survey-induced changes in time use. We find no evidence for strategic misreporting or survey induced changes in true activities. We can rule out sampling errors as a driver of differences, with the exception that fitting activities into 30 minute blocks with DR method appears to lower the probability of reporting short-duration activities. We focus the remaining discussion on recall errors, finding that differences in reporting are related to the respondents' time literacy, and that experimentally induced lengthening of bounded recall periods lowers the probability that short-duration activities are reported by respondents. These findings are consistent with recall errors being largely responsible for the differences in time use measured between the ATOC and DR methods. We report additional details about the analysis we use to explore measurement error mechanisms in Appendix C.

5.2 Strategic misreporting of outcomes

In the context of time use reporting, respondents might misreport their activities intentionally because they want to over-report socially desirable activities or under-report socially undesirable activities. Most often associated with desirability bias, the respondent might report untrue outcomes if she believes that misreporting could help her to avoid social sanction or could affect future payoffs. The potential for social desirability to affect respondents' reporting of time use depends on their activities and their beliefs about the extent to which some activities might be perceived as desirable or undesirable by enumerators (Bound et al., 2001). Strategic mis-reporting of time use could occur with either method of data collec-

tion. With the DR method, respondents might have additional opportunity to shade their estimates of yesterdays' activities, emphasizing desirable activities while de-emphasizing undesirable activities. DR respondents could also "fudge" their reporting of a current activity, instead favoring a very recent activity that they might believe is more desirable.

We made design choices to minimize the influence of desirability bias. Specifically, communicated clearly to all respondents and enumerators that this study's only purpose was to learn the truth about respondents' lives, and that their responses to questions would not affect any aspects of their VSLA program participation (see Section 2). We might still be concerned that survey mode biases might affect respondents differently based on survey mode (phone versus in person) or enumerator gender (male enumerators in person versus female enumerators in person). In a separate study, we show that respondents did not report sensitive gender norms differently when enumerator gender or survey mode were experimentally assigned (Kadam et al., 2025), suggesting that desirability bias is unlikely to explain activity share differences between ATOC and DR groups.

5.3 Sampling errors

Sampling-related measurement errors relate to when and how respondents are asked about their time and who conducts the interview. The DR method, for example, requires the respondent to describe her previous day in a way that enumerators then map onto 30-minute blocks. The respondent might make errors in fitting her activities on these blocks or might fail to report short-duration activities that never comprise the majority of any 30-minute block. The respondent's ability to sit for an interview and take the time to carefully construct her prior day might relate to her activities at the time the interview occurs. With regards to the ATOC method, a respondent might be less likely to answer a call depending on her activities or her location at the time the call is placed. Furthermore, her activities at times outside of call center operation (early morning, late evening and overnight) are never subject to experience sampling.

We ensure that our results are not driven by the limited hours of call center operation by restricting the DR activity shares to the same hours that our call center was also in operation (7am to 6pm).

We next examine missed calls with for the ATOC treatment group (Section C.1.1). We show that missed calls are not very common and appear to be driven by network outages rather than respondent's activities at the time of missed calls. We bound our experimental estimates of the effect of ATOC on time shares by recoding missed calls and show that the research conclusions are robust to this recoding. While most respondents answered all or all

but a few of their calls, a small subset missed a large share of calls. We show that our results are not driven by this subset of respondents who missed a large share of calls by comparing activity shares between ATOC respondents who missed a large share of ATOC calls and DR-PP respondents who missed a large share of their placebo content calls.

We do find evidence that respondents surveyed using the DR method are less likely to mention short-duration activities than respondents surveyed with the ATOC method. A plausible explanation for this is that respondents are asked to mention the main activity that took place within each 30-minute block, and so activities that might last shorter than 15 minutes, even if they occur multiple times per day, are unlikely to be reported.

5.4 Survey-induced changes in true outcomes

A respondent might change her true activities because she knows she is being observed (Hawthorne effect). Because the ATOC method involves repeatedly asking a respondent about her time use, we might be concerned that she engages in different activities than she would otherwise because she knows she will be asked about her activities. With the DR method, respondents report their time use during one single interview, so the Hawthorne effect is not a concern. With both survey methods, we might be concerned that a respondent might engage in different activities because she was given a mobile device, an asset with some value that could shift her consumption of leisure (income effect) or could cause her to substitute her activities because of a change in her ability to receive or transmit information. We collectively refer to income- or substitution- effects on activities that result from our giving phones to a respondent as an endowment effect.

In Section C.1.2, we test for an endowment effect by comparing DR-reported time use between respondents in the DR-PP group (who were given phones) and those in the DR-NP group (who were not given phones). Also in Section C.1.2, we test for a Hawthorne effect by looking for changes over time in time use reports (or difference between the first call(s) and subsequent calls) and show no evidence that activities change over the reference period. We find no difference between both DR sub-groups.

5.5 Recall error mechanisms

Respondents tend to organize their memories hierarchically, with life domains such as history of residence or employment at the top, extended periods within each domain below that, memories of individual events below that, and details of individual events at the lowest (Tourangeau, 2018). Respondents might use different strategies to recall their time use, such as constructing a response from actual activities (episodic enumeration) estimating a

response based on patterns (Bound et al., 2001). When respondents use episodic reporting, they often make telescoping errors, whereby they place additional events inside the reference period that actually occurred outside of it or they fail to include events that occurred in the reference period that did occur within it. Recall errors do not occur in clear patterns, though are generally thought to be larger when the reference period is longer, which creates more room for telescoping errors, and when the activities reported are not particularly salient to the enumerator.

When we restrict our activity shares comparison to the ATOC group and the DR-PP group, we come closest to isolating recall errors as responsible for the differences observed. Both of these groups are similar in that they were both given a phone several days before the study period, received four to five calls per day during the study period, and were set to receive a small compliance incentive in the form of mobile phone credit at the end of the study period. Table A2 reports that the phone survey effect is very similar when we include only the DR-PP group in the comparison group.

In Section C.1.3, we explore in detail the possible recall mechanisms that could explain the experimental differences between the ATOC and DR survey groups. We find that recall error could be related to the respondent's time literacy. Respondents who more accurately guessed the length of their in-person baseline interview reported more similar economic activity shares between the DR versus ATOC methods than did respondents who over- or under-estimated the duration of their baseline interview by more than 15 minutes (Figure C5A). We also show that, within the DR group (but not the ATOC group), the total reported time share for economic activities increases as the error in estimating the duration of the in-person baseline interview grows (Figure C7).

We find that respondents surveyed using the DR method are less likely to report brief-duration activities on the extensive margin (whether they occurred) than respondents surveyed with the ATOC method (Figure C8A). We also construct a robustness to see whether the respondents report their activities differently based on whether they are in the presence of others and conclude that presence of others is unlikely to change the activitity shares reported using either the DR or the ATOC method (Figure C9).

6 Applications to Labor Productivity and Lessons for Future Studies

In this section, we discuss how our findings can shape interpretation of past studies and provide guidance for those in the future, regardless of survey method used. We first demon-

strate the implications of our findings on recall errors for measuring labor participation and for labor productivity. Then, we discuss how our results can be used to assist interpretation of existing studies using time use data for measuring key outcomes, including a discussion of the extent to which our results might be generalizable to populations beyond the rural poor in Ethiopia. Finally, we provide a description of the survey costs of the experience sampling methodology (ATOC) and offer implementation considerations that future researchers may consider.

6.1 What Do Recall Errors Imply for Labor Force Participation and Labor Productivity Measurements?

Our results presented thus far show that on average women's likelihood of reporting time spent on economic activities (the extensive margin) is higher in the ATOC group (phone interviews) than the DR (in-person interviews) group, by a factor of two (see Section 4). These are very large average effects. In Section 5.5, we discussed one possible recall error mechanism is that when respondents are constrained to report on their activities in 30-minute increments, this forces them to omit brief episodes of an activity, even if the total time spent on that activity across all episodes might be large in a given day.

Our baseline "labor force participation rate" (LFP), which relied on the respondents' reports of involvement in any economic activities during the previous week, was 80%, and balanced on treatment group. By contrast, the DR method returns a LFP rate (non-zero economic activities on the day prior to the interview) of 30%, markedly lower.[26] On the other hand, the ATOC method returns a 7-day participation rate of 62%—still nearly 20 percentage points lower than the baseline but much closer than the DR. Figure 6A shows this graphically, plotting activity shares for economic activities (y-axis values) from lowest to highest. These are plotted separately by treatment group and for women who were economically active according to their baseline survey (left panel) versus those who were not economically active (right panel).

There are several interesting patterns to observe in these figures. First, about 25% of the women who were economically active at baseline did not report any economic activities during their interviews by phone (for ATOC) or in person (DR), since both curves in the left panel are zero for the first quarter of the sample. However, this is matched by an equal proportion on the right panel: about 25% of the respondents from the group who was *not*

[26]Of course, women may not be economically active every day, but since we randomly assigned women to a single day for their in-person recall interview, the average across the group should be representative. Indeed, using data from the DR group in each village to represent a pseudo-panel of the 7 days yields an identical estimate of participation in economic activities.

economically active at baseline reported nonzero shares for economic activity during the experimental period, regardless of the survey methodology used for asking about their time use during this period. The implied labor force participation rates across the two survey types *are* different, however, because a sizable 80% of women who were economically active at baseline and were assigned to report on their activities at high frequency by phone reported economic activities during their phone interviews, while only 40% of the DR respondents did, since the DR curve remains at zero until the x-axis shows 0.6 when it becomes nonzero. As a result of these patterns, only 37% the DR respondents would be recorded in the data based only on their in-person day reconstruction interviews as being economically active, or members of the "labor force."[27] By contrast, 69% of the respondents in the ATOC group would be recorded as being economically active based on their activity shares (recall that at baseline 80% of the women were economically active in the previous seven days, and this is balanced across treatment groups). The implication is that traditional recall surveys may understate women's attachment to the labor force.

What about the intensive margin? Figure 6 also provides two important implications for measures of the intensive margin of labor supply. Whenever the ATOC curve lies above the DR curve, the duration estimated by the ATOC is larger, and vice versa. Two patterns emerge. First, the majority of women who are economically active based on the ATOC but would not be if they had been surveyed using a traditional DR mdoule are working for very short periods of time. This makes sense. In the aggregate, mapping this pattern onto a population, this would serve to decrease the average hours worked. This is similar to other issues related to part-time work. Additionally, the figures show that for nearly all women who are identified as being involved in the labor market regardless of survey type, the duration is larger using the DR method than using the ATOC method. In some cases, it is larger by more than double (e.g., the highest activity share for the DR women is 0.6 and the highest for the ATOC women is 25%). Again, this would serve to reduce the average estimate of labor supply on the intensive margin across these groups. Additionally, the steep incline for the DR group and the relatively flat incline for the ATOC group suggests the possibility of heterogeneous treatment effects, as the gap between the ATOC and the DR group at the highest percentiles in the sample gets larger.

[27] Given the baseline labor force participation rate of 80%, which determines the fraction of the sample represented by the left panel (participants) versus the right panel (non-participants, 20%), this is calculated as $0.8 * 0.4 + 0.2 * 0.25$—i.e., the weighted sum of 40% participation (left panel) and 25% participation (right panel).

6.2 What Can We Learn About Existing Time Use Data?

Given the errors in recall survey data we report in this article, readers may wonder what is to be done about existing studies, or in scenarios where funds are not available to collect high-frequency phone survey. We provide some insights here.

First, we explore the implications of non-classical measurement error for the findings of a recent study examining the long-run implications of a large transfer program to the ultra-poor in India—a setting and population not unlike ours in terms of education and numeracy, reliance on subsistence agriculture, low participation rates among women in economic activities, and a high burden of household chores. Specifically, we work with the replication data from Banerjee et al. (2021), simulate non-classical measurement error in the sample, and re-estimate treatment effects.

Modelling a scenario similar to what is implied by our Figure 6A, we assume that everyone over-estimates their labor supply (minutes) by 1 percent but their over-estimation increases by 2 percentage points for every additional m minutes that they report working. We simulate the implied labor measure when m is allowed to vary between 10 and 60 minutes. For $m = 30$, this implies that someone who reported working 90 minutes over-estimated this by 7 percent and worked only 83.7 percent. Figure 8A shows the distribution of reported time and adjusted ("corrected") time, showing that the errors increase as time spent working increases. Figure 8B shows the implied treatment effects (in minutes) estimated using these data for the treatment group relative to the control, 10 years after households in the treatment group was provided with a large one-time transfer. As the coefficients show, modeling individuals who over-estimate their working time produces smaller and smaller treatment effects as we increase the parameter heterogeneous over-estimation tendencies (that are an increasing function of time reported working).

6.3 What Should Future Researchers Know for Implementing High-Frequency Phone Surveys about Current Activities?

Phone ownership rates at baseline were 44 percent, but only about 8% of respondents reported being the primary user of the phone (and only 30% used it at all), so we had to provide phones to participants in order to ensure that they could participate in the study. This was a substantial driver of survey costs—about $20 per phone in 2019, or $13,000 total for the entire sample). However, if certain types of respondents are more or less likely to participate in high-frequency phone surveys, sample selection issues would limit the external validity of estimates. Even if researchers provide phones to respondents to enable their participation, the provision of a phone might induce an endowment effect, with implications

for time use. For example, respondents might use the phone for business transactions, for entertainment, or to light their home at night, all of which might drive changes in time use. They might also become more aware of time if they use the phones as a timekeeping device. We were also concerned that women's husbands would take over the phone after some time or phones might be lost, stolen, or broken—these concerns partly motivated our design of completing all survey activities with a given household within 12 days.[28]

7 Discussion

We provide experimental evidence of substantial errors in recall data about time allocation using the traditional methodology that is considered to be best practice. To our knowledge, this is the first study to successfully implement a series of high-frequency mobile phone surveys using an "experience sampling method" to provide this analysis. Our study population consists of women in poor households in rural Ethiopia who are participating in multi-faceted graduation-from-poverty programs. Specifically, we find that the traditional day reconstruction (DR, or "24 hour recall") surveys *over-estimate* women's time on economic and social activities and *under-estimate* women's time on household chores, travel, and personal care. The recall surveys also under-estimate...

While we do not observe women's true time shares and thus cannot say definitively that our experience sampling method is more accurate than the traditional 24-hour time use diary, we are able to rule out some sources of error that could bias our high-frequency mobile phone measures of time use. Because a very large share of our phone calls are answered by respondents at the time the call is placed, we are confident that any activity-related reasons that affect our respondents' probability of reporting their time use are not meaningfully altering their time shares. We confirm this by bounding our time share estimates for each activity by the highest (and lowest) possible shares if a respondent were to have reported doing the activity for all (none) of her missed calls. We also rule out that our methodology itself induced changes in women's time allocation, either through a phone endowment effect or a Hawthorne affect arising from repeated questioning about time use during the study period. We rule out these explanations by comparing time shares reported using the 24-hour time use diary between two groups —the phone placebo group, which was given a phone and received high-frequency phone calls about a placebo topic unrelated to time use during the same period, and the control group, which was not given a phone until after the study was

[28] We are currently implementing a new project in a similar population via phone surveys over the course of 6 months, after providing respondents with phones, and are having no trouble with either of these concerns, however.

completed. We do not detect differences in time shares between these two groups, suggesting that the phone endowment did not induce changes in time use.

Low time literacy among our study population could explain differences between time shares reported through our experience sampling method and the 24-hour recall method. Our respondents have low levels of educational attainment, numeracy, and time literacy. Only 31% of our respondents were able to correctly report the number of minutes in an hour and only 44% could correctly report the current time within 30 minutes. Our respondents show greater ability to estimate the passing of time, but many of them mis-estimate the duration of their baseline interview.

Accurately measuring time use is important for understanding labor productivity. It is also essential for understanding investments in human capital formation, which can affect outcomes across generations. Time use can also be an important determinant of a range of physical or psychological health outcomes of which our understanding is limited. Using simulations from replication data of a successful study, for example, we show that using an upwardly-biased measure of economic labor inputs over-estimates the impact of the intervention on labor supply (minutes) but also under-estimates labor productivity.

Similarly, if a program induces women to supply more economic labor, but her household work is underestimated (as our results show is true in the study sample), researchers evaluating the program could mis-estimate the impact of the program on women's economic empowerment.

Measuring women's time use accurately is also important for developing a holistic perspective on the possibility for long-term improvement in women's empowerment. While many graduation programs target women's economic activity as a key outcome of interest, to our knowledge, no existing evaluation of graduation programs to date provides evidence on whether these changes in women's labor supply to economic activities are offset by any accompanying changes in other activities. There is evidence to suggest that the gains observed may be short-lived or, sometimes, occur at the expense of other aspects of women's lives (Banerjee et al., 2015). If women merely add more to their workloads as a result of these programs, they risk burnout, depression, and other physical and mental health problems, which pose significant barriers to fighting poverty in the long run (Schilbach et al., 2016). Moreover, if unpaid household chores are re-allocated to (girl) children at the expense of time spent studying or playing, or if women who increase labor supply due to participating in these programs in turn reduce the quality of their investments in children's nutritional, emotional, and cognitive development, inter-generational welfare losses could arise. Our methodology for measuring time use could be used for tracking program outcomes over time

without requiring multiple field visits,[29] to understand unintended consequences of targeted programming on women's labor supply.

The phone-based experience sampling method we utilized offers a promising tool for tracking outcomes like labor supply, productivity, and empowerment. Whether it is also a cost-saving tool depends on the researchers' objectives, study design, and local context. If multiple time-use measures are required from respondents who are likely to own a phone already and if transportation costs are high but enumerator labor costs are not, the method can absolutely be cost effective.

Our research raises a number of questions that can be explored in future research. First, it would be interesting to test whether this phone methodology affects the evaluation of a program's impacts. By nesting a similar experiment to ours into an experimental impact evaluation of a livelihoods improvement program, one could explore the program's impacts on time use in more detail and also assess whether the time use measurement method affects findings about the program's effectiveness or unintended consequences. It also would be interesting to generalize validation of this methodology in a population where prior phone ownership is common and phones are not provided to respondents.

References

Abay, K. A., Abate, G. T., Barrett, C. B., & Bernard, T. (2019a). Correlated non-classical measurement errors, 'Second best' policy inference, and the inverse size-productivity relationship in agriculture. *Journal of Development Economics*, *139*, 171–184.

Abay, K. A., Bevis, L., & Barrett, C. B. (2019b). Measurement error mechanisms matter: Agricultural intensification with farmer misperceptions and misreporting. *National Bureau of Economic Research Working Paper Series: 26066*.

Adamopoulos, T., & Restuccia, D. (2021). Geography and agricultural productivity: Cross-country evidence from micro plot-level data. *The Review of Economic Studies*, (September), 1–25.

Arthi, V., Beegle, K., De Weerdt, J., & Palacios-López, A. (2018). Not your average job: Measuring farm labor in Tanzania. *Journal of Development Economics*, *130*(July), 160–172.

Bandiera, O., Burgess, R., Das, N., Gulesci, S., Rasul, I., & Sulaiman, M. (2017). Labor markets and poverty in village economies. *Quarterly Journal of Economics*, (November), 811–870.

[29] We currently have a project in the field doing this for a new study funded by the International Food Policy Research Institute (IFPRI) with a sample that is very similar to the sample for this project.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., & Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, *348*(6236).

Banerjee, A., Duflo, E., & Sharma, G. (2021). Long-term effects of the targeting the ultra poor program. *American Economic Review: Insights*, *3*(4), 471–86.

Bardasi, E., Beegle, K., Dillon, A., & Serneels, P. (2011). Do labor statistics depend on how and to whom the questions are asked? Results from a survey experiment in Tanzania. *World Bank Economic Review*, *25*(3), 418–447.

Barrett, G. F., Hamermesh, D. S., Barrett, G. F., & Hamermesh, D. S. (2019). Labor supply elasticities : Overcoming nonclassical measurement error using more accurate hours data. *54*(1), 255–265.

Becker, G. S. (1965). A Theory of the Allocation of Time. *The Economic Journal*, *75*(299), 493.
URL http://www.jstor.org/stable/10.2307/2228949?origin=crossref

Blackden, C. M., & Wodon, Q. (2006). *Gender, Time Use, and Poverty in Sub-Saharan Africa*. World Bank.

Blau, F. D., Kahn, L. M., Blau, F. D., & Kahn, L. M. (2017). Changes in the labor supply behavior of married women: 1980-2000. *Labor Economics*, *25*(3), 393–438.

Borjas, G. J. (1980). The relationship between wages and weekly hours of work: The role of division bias. *Journal of Human Resources*, *15*, 409–23.

Borjas, G. J., & Hamermesh, D. S. (2024). The mismeasurement of work time: Implications for wage discrimination and inequality. *National Bureau of Economic Research Working Paper Series: 32025*.

Bound, J., Brown, C., & Mathiowetz, N. (2001). Chapter 59: Measurement error in survey data, ch. 59. In *Handbook of econometrics*, vol. 5, (pp. 3705–3843). Elsevier.

Csikszentmihalyi, M. (2014). *Flow and the foundations of positive psychology*, chap. Validity and reliability of the experience-sampling method, (pp. 35–54). Springer.

Davis, B., Di Giuseppe, S., & Zezza, A. (2017). Are african households (not) leaving agriculture? patterns of households' income sources in rural sub-saharan africa. *Food policy*, *67*, 153–174.

Dinkelman, T., & Ngai, L. R. (2022). Time use and gender in africa in times of structural transformation. *Journal of Economic Perspectives*, *36*(1), 57–80.

Duflo, E. (2003). Grandmothers and granddaughters: Old-age pensions and intrahousehold allocation in South Africa. *The World Bank Economic Review*, *17*(1), 1–25.

Field, E., Pande, R., Rigol, N., Schaner, S., Stacy, E., & Moore, C. (2022). Understanding rural households' time use in a developing setting: Validating a low-cost time use module. *SSRN Electronic Journal*.

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

Frazis, H., & Stewart, J. (2010). How to think about time-use data: What inferences can we make about long- and short-run time use from time diaries? *Institute of Labor Economics (IZA) Discussion Paper: 5306*, *November*.

Gaddis, I., Oseni, G., Palacios-Lopez, A., & Pieters, J. (2021). Measuring farm labor: Survey experimental evidence from Ghana. *The World Bank Economic Review*, *35*(3), 604–634.

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

Gollin, D., Lagakos, D., & Waugh, M. E. (2014). The agricultural productivity gap. *Quarterly Journal of Economics*, *129*(2), 939–993.

Grossbard-Shechtman, A. (1984). A theory of allocation of time in markets for labour and marriage. *The Economic Journal*, *94*, 863–882.

Heath, R., Mansuri, G., Rijkers, B., Seitz, W., & Sharma, D. (2021). Measuring employment: Experimental evidence from urban Ghana. *The World Bank Economic Review*, *35*(3), 635–651.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–61.

Heckman, J. J. (1993). What has been learned about labor supply in the past twenty years? *AEA Papers and Proceedings*, *83*, 116–21.

Hotz, V. J., & Wiswall, M. (2019). Child care and child care policy: Existing policies, their effects, and reforms. *Annals of the American Academy of Political and Social Sciences*, *686*, 116–21.

Hussmanns, R., Mehran, F., Varmā, V., & Office, I. L. (1990). *Surveys of Economically Active Population, Employment, Unemployment, and Underemployment: An ILO Manual on Concepts and Methods*. International Labour Office.

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

Kadam, A., McCullough, E., McGavock, T., & Magnan, N. (2025). Who's Asking And How? Effects of survey mode and enumerator gender and survey mode on measuring women's life experience. *World Development*.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). 'A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method'. *Science*, *306*(December), 1776–1780.

Kilic, T., Koolwal, G., Vundru, W. D., Daum, T., Buchwald, H., Seymour, G., Mvula, P., Munthali, A., & Kachinjika, M. (2024). Recording the time divide: A comparative study of smartphone- and recall-based approaches to time use measurement. *The World Bank, Policy Research Working Paper Series: 10695*.

Lachowska, M., Mas, A., & Woodbury, S. A. (2022). How reliable are administrative reports of paid work hours? *Labour Economics*, *75*.

Malapit, H., Quisumbing, A., Meinzen-Dick, R., Seymour, G., Martinez, E. M., Heckert, J., Rubin, D., Vaz, A., & Yount, K. M. (2019). Development of the project-level women's empowerment in agriculture index (pro-weai). *World Development*, *122*, 675–692.

McCullough, E. B. (2017). Labor productivity and employment gaps in Sub-Saharan Africa. *Food Policy*, *67*, 133–152.

Millimet, D. L., & Wang, L. (2011). Is the quantity-quality trade-off a trade-off for all, none, or some? *Economic Development and Cultural Change*, *60*(4078), 155–195.

Qian, N. (2008). Missing women and the price of tea in China: The effect of sex-specific earnings on sex imbalance. *Quarterly Journal of Economics*, *123*(3), 1251–1285.

Schilbach, F., Schofield, H., & Mullainathan, S. (2016). The psychological lives of the poor. *American Economic Review*, *106*(5), 435–440.

Seymour, G., Malapit, H., Quisumbing, A., Seymour, G., Malapit, H., & Measuring, A. Q. (2020). Measuring time use in developing country agriculture: Evidence from Bangladesh and Uganda. *Feminist Economics*, *0*(0), 1–31.

Tourangeau, R. (1987). Attitude measurement: A cognitive perspective. In *Social information processing and survey methodology*, (pp. 149–162). Springer.

Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education*, *26*, 169–181.
URL https://doi.org/10.1108/QAE-06-2017-0034

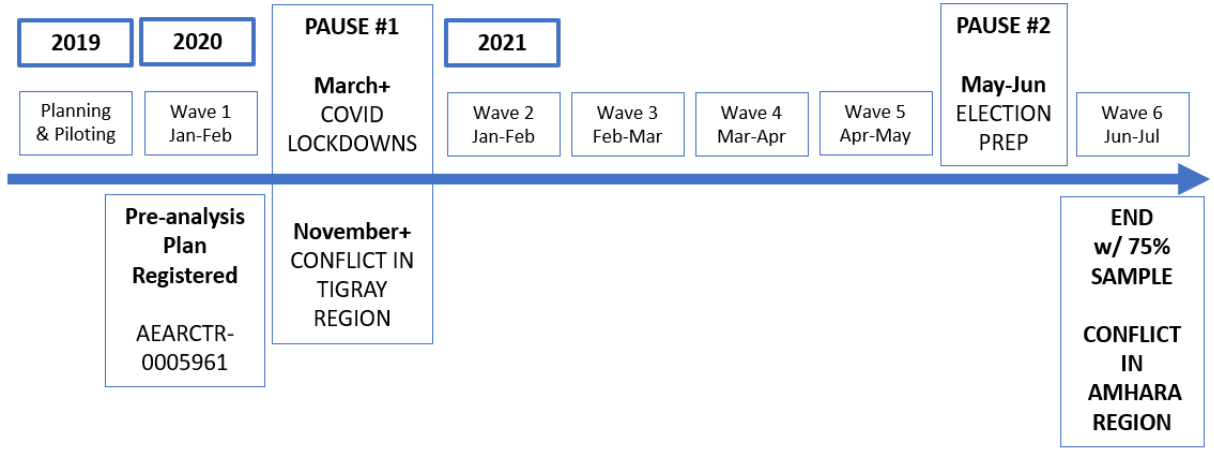
Wilson, R. (2020). The EITC and employment transitions: Labor force attachment and annual exit. *National Tax Journal*, *73*, 11–46.

Zezza, A., Mcgee, K., Wollburg, P., Assefa, T., & Gourlay, S. (2022). From necessity to opportunity: Lessons for integrating phone and in-person data collection for agricultural statistics in a post-pandemic world. *The World Bank, Policy Research Working Paper Series: 10168*.

Figures

Figure 1: Data Collection Timelines

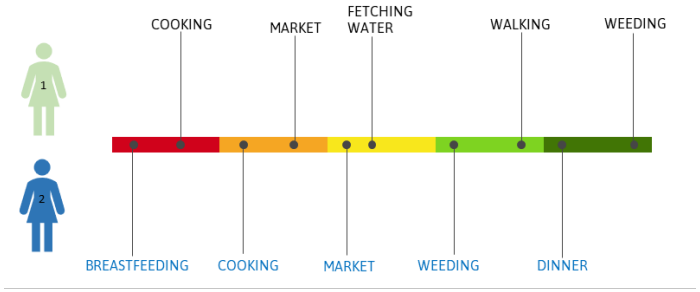
(A) Eight Survey Waves Planned, Six Completed



(B) VESAs Overlap Within Waves

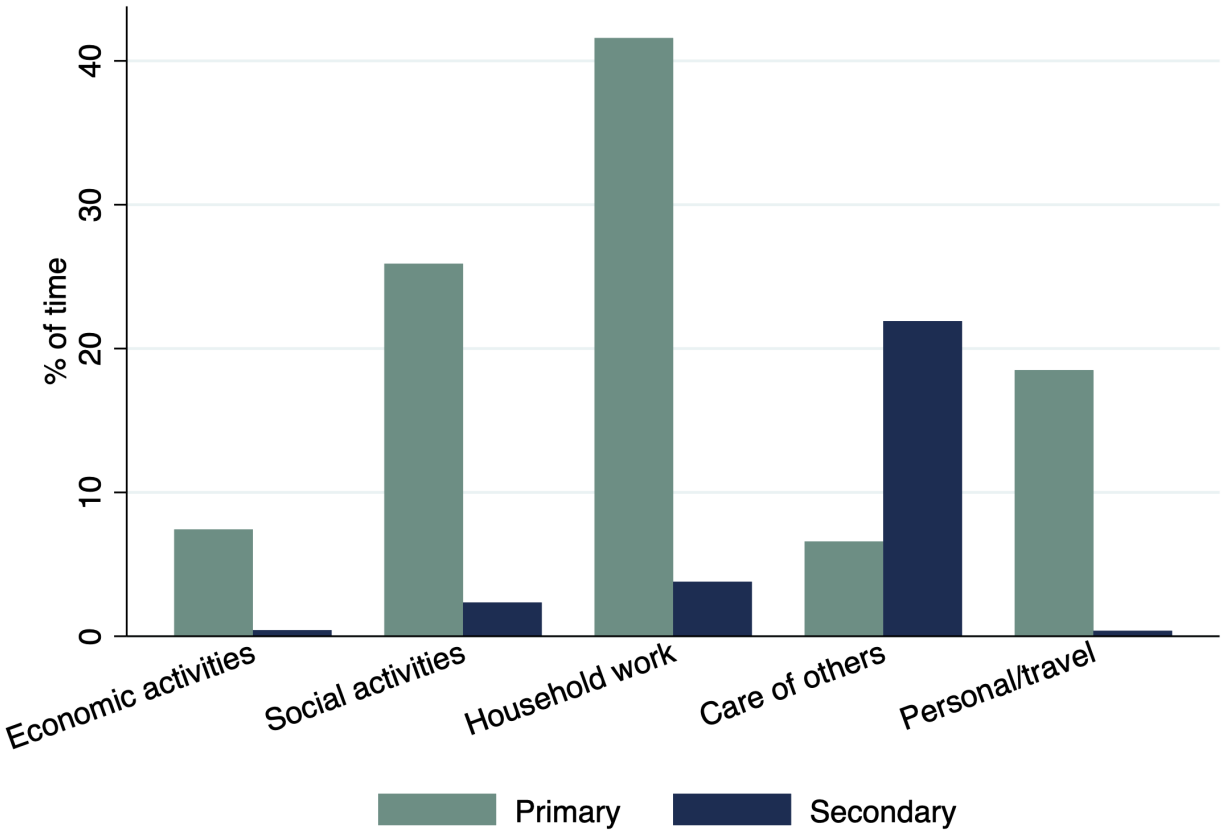


(C) Calls Within Strata



Note. See Section 2 and Appendix B for more details on logistics of data collection.

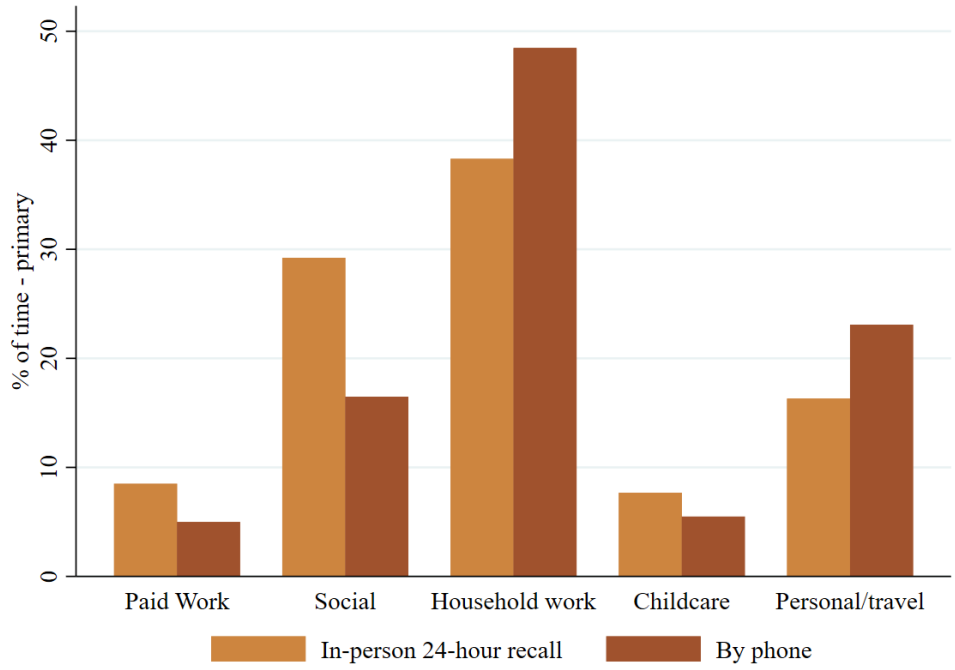
Figure 2: Primary and Secondary Activity Shares - Pooled Data



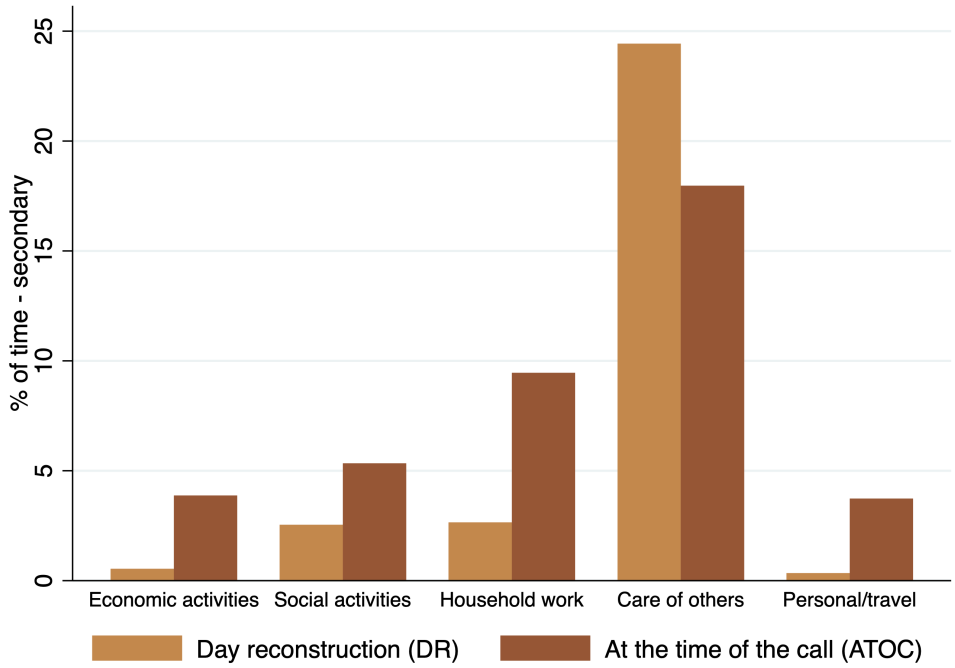
Note. The figure shows average activity shares for women in the sample across the five categories, pooled across all 646 women in the sample. The calculation of activity shares depends on randomized assignment to a survey mode (see Section 2).

Figure 3: Activity Shares by Treatment Group

(A) Primary Activities

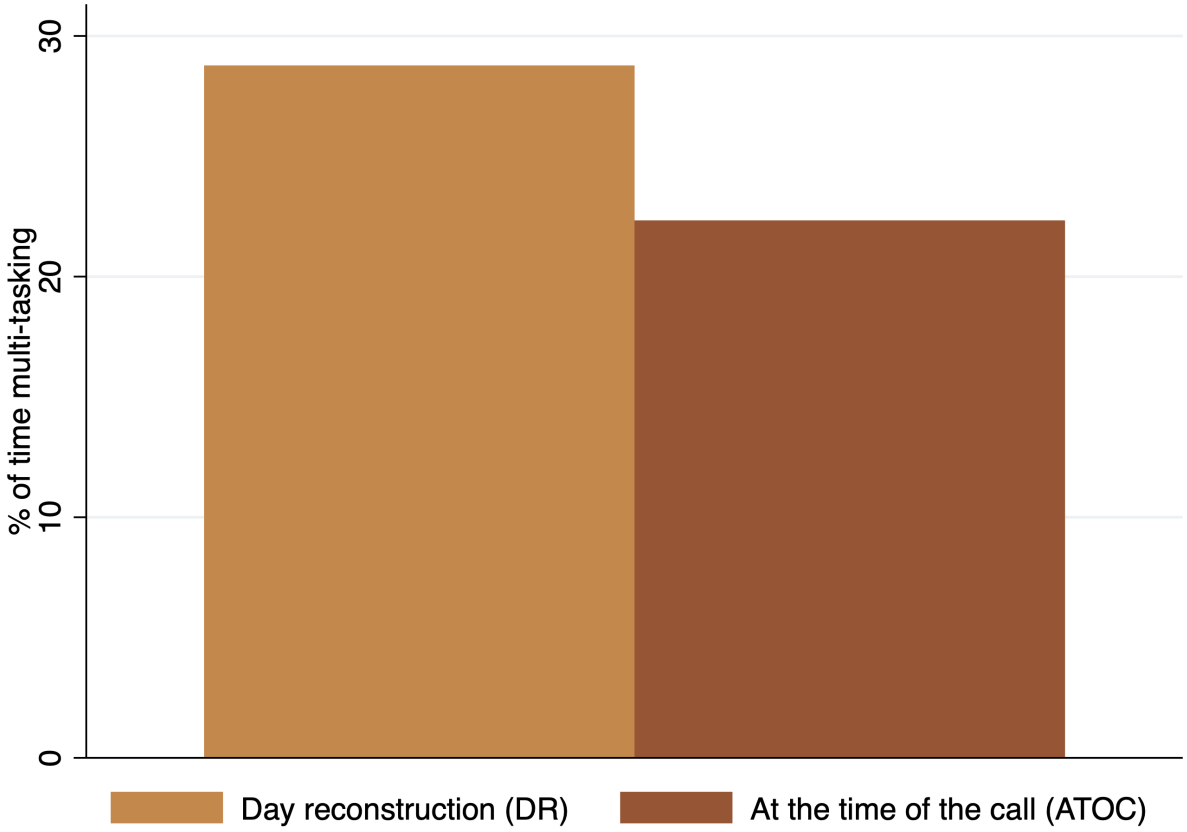


(B) Secondary Activities



Note. This figure shows the average activity share for respondents in the sample across the five categories, separately by survey method.

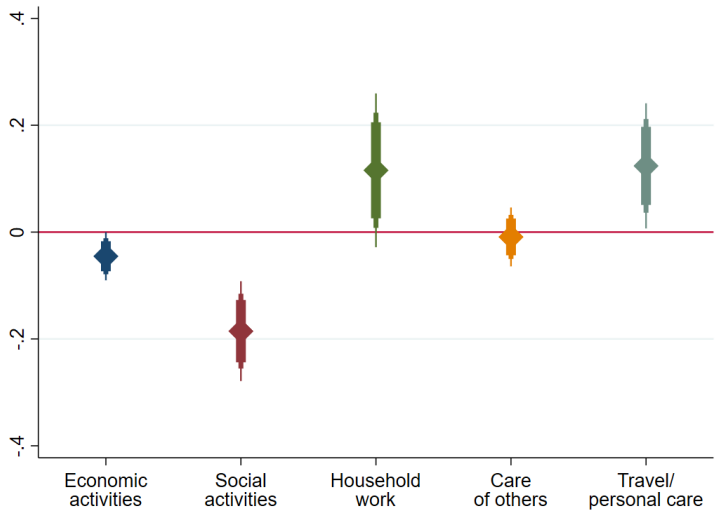
Figure 4: Multitasking by Treatment Group



Note. This figure shows the average share of time that respondents were "multi-tasking" based on treatment groups, where multi-tasking is defined as reporting any secondary activity. Importantly, the DR surveys may record both a primary and secondary activity that were performed sequentially rather than simultaneously for the entire 30-minute block of time, so this may overstate true multi-tasking in the DR data.

Figure 5: Survey Mode Treatment Effects: Difference between Activity Shares for High-Frequency Phone Respondents (ATOC, the treatment) and In-person Day Reconstruction Respondents (DR, the control)

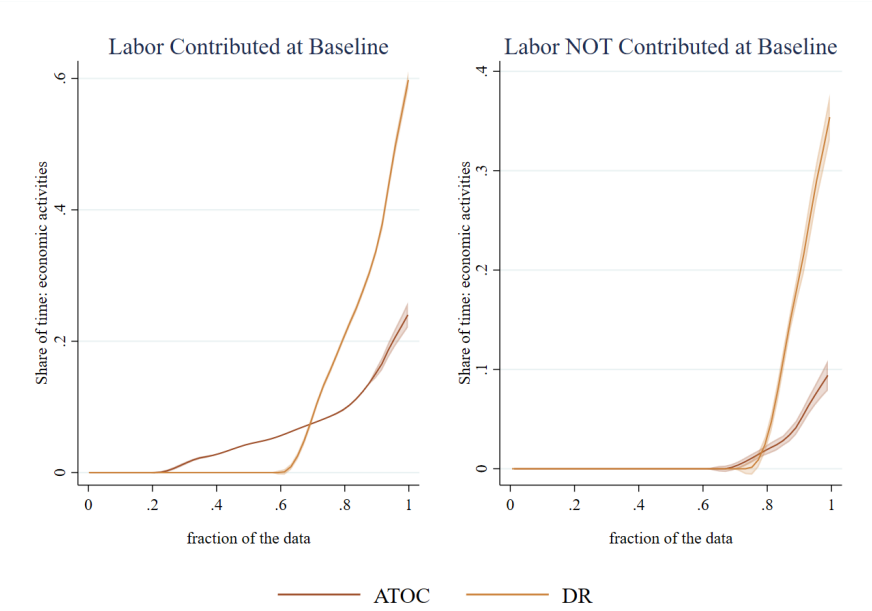
(A) Treatment Effect



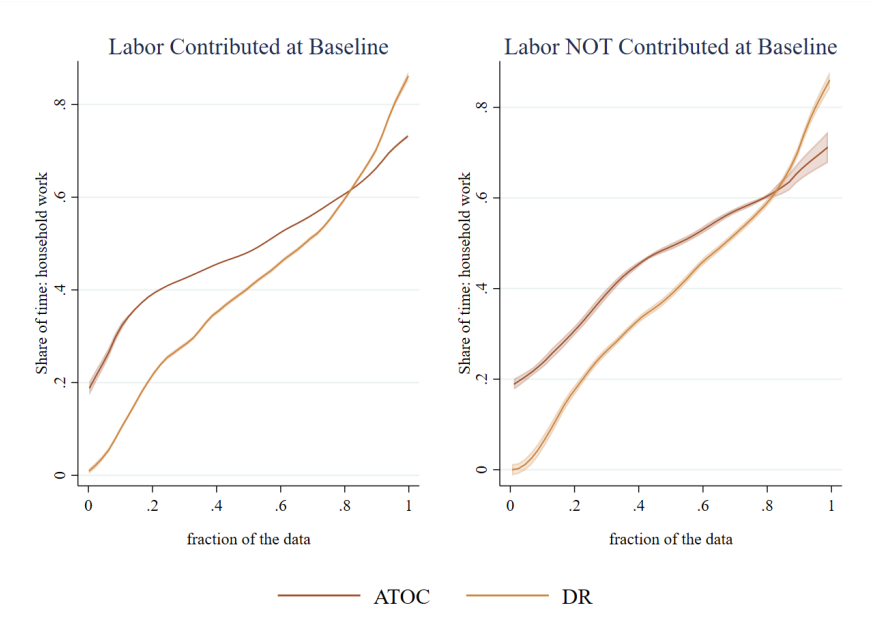
Note. The figure shows the estimated coefficients for Equation 1 when the outcome is a woman's activity share for each category of activity type (economic, social, etc.), defined differently according to treatment group. For the ATOC group, each respondent's activity share for a given category is constructed as the ratio of the number of calls she answered and reported performing an activity in that category to the total number of calls that the woman answered across the validation period. The call center operated from 7 am to 6pm each day. For the DR group, the outcome for that same period for a given category is the ratio of time blocks for which the respondent reported an activity in that category to the total number of time blocks between 7am and 6pm. Positive values indicate that the average ATOC activity share for a given category is larger than the DR share, so a $\beta < 0$ suggests that respondents in the DR group over-estimate time spent on that activity (see **??** for supporting details on this interpretation). VESA (village) fixed effects are included and standard errors are clustered at the VESA level. No other controls are included in this specification, our most conservative estimate.

Figure 6: Activity Shares Conditional on Participation in Economic Activities at Baseline

(A) Economic Activity Shares by Treatment Group and Baseline Work Participation

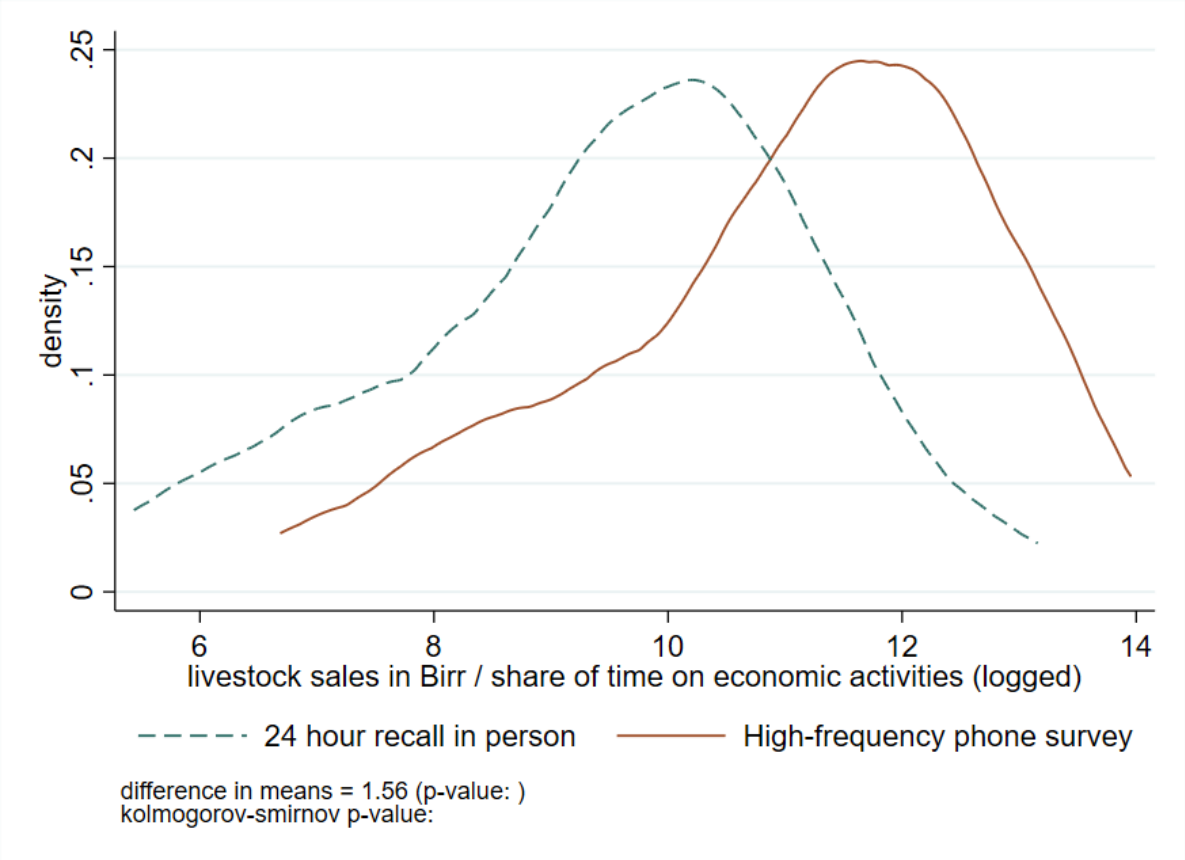


(B) Household Work Shares by Treatment Group and Baseline Work Participation



Note. Respondents are defined as having been participating at baseline if they reported any time spent in the previous seven days on an activity that is classified by the WEAI as "productive" or economically active.

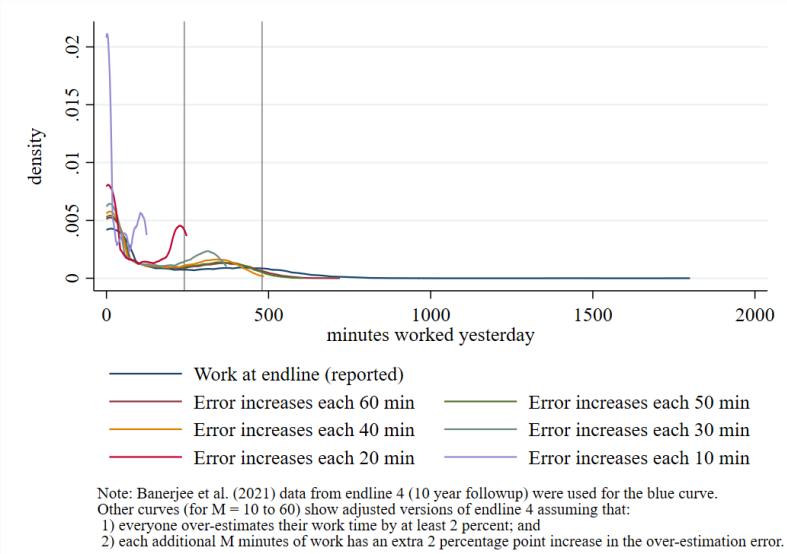
Figure 7: Implied Labor Productivity



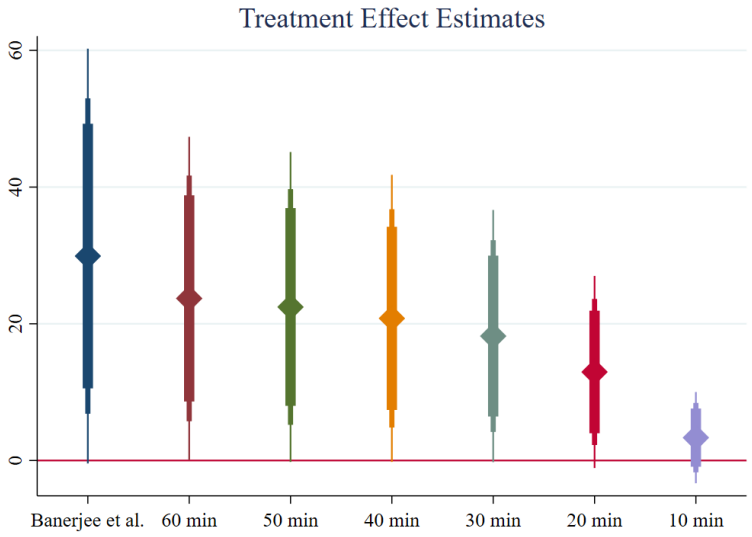
Note. Here we demonstrate the implications of survey modality for measuring labor productivity. We construct "implied labor productivity" as follows: 1) the numerator is the change in value of all livestock owned by the household over the year prior to the baseline interview (respondents reported first the number of units of each livestock type and then estimated the value of each type in Birr). 2) the denominator is the woman's activity share over all economic activities, as measured via her randomly assigned survey modality. Thus the ratio of the change in value of all livestock holdings to the respondent's activity share can be a proxy for average labor productivity. We elected to use livestock value and all economic activity to maximize the number of households for whom we could construct such a measure, since there are many missing values for self-employment income and/or economic activity shares in specific sub-categories. As the figure shows, at every point in the distribution, the implied productivity for the ATOC respondents is higher than the implied productivity for the DR respondents, indicating that, by over-estimating labor supply, recall surveys also under-estimate labor productivity.

Figure 8: Simulating Treatment Effects from Banerjee et al. (2021)

(A) Distribution of work minutes (actual and simulated)



(B) Treatment effects at 10 year endline (actual and simulated)



Note. Authors' calculations using replication data from Banerjee et al. (2021). The navy blue curves show the actual distribution (top) and actual treatment effect (bottom) in minutes using their data and estimation strategy. See Section 6.2 for details.

Tables

Table 1: Summary Statistics and Balance for Baseline Characteristics

	Recall surveys (DR)	Phone surveys (ATOC)	Difference (ATOC - DR)	p-value
Age in years	38.072	37.986	-0.086	0.917
Worked in 7 days before baseline	0.786	0.796	0.010	0.763
Attended any school	0.230	0.227	-0.003	0.923
Years of education, if any	4.673	3.896	-0.778	0.092
Is pregnant	0.087	0.084	-0.003	0.909
Is breastfeeding	0.982	0.982	0.000	0.994
Uses a time-keeping device	0.093	0.102	0.009	0.720
Answers correctly: # hours in a day	0.779	0.782	0.003	0.923
Answers correctly: # minutes in an hour	0.228	0.199	-0.029	0.403
Estimation error: duration of baseline interview (%)	10.981	9.366	-1.616	0.680
Female-headed household	0.158	0.176	0.018	0.565
Adults in wage employment (# in HH)	0.328	0.319	-0.008	0.842
Children under 5 (# in HH)	0.688	0.681	-0.008	0.889
Children 5 to 12 (# in HH)	1.070	1.083	0.014	0.849
Cash support only from PSNP	0.073	0.057	-0.017	0.431
Owns a cell phone (HH)	0.437	0.417	-0.021	0.619
Owns any land (HH)	0.984	0.981	-0.002	0.836
Operates any land (HH)	0.995	0.981	-0.014	0.083
Owns any livestock	0.960	0.944	-0.016	0.354
Land owned by HH, if any (ha)	0.711	0.677	-0.034	0.426
Land operated by HH, if any (ha)	0.896	0.808	-0.089	0.075
Livestock value, if any (2021 Birr)	19846.651	18771.759	-1074.892	0.365
Observations	211	216		

Note: The table reports means of baseline characteristics for 1) the placebo DR group—that is, the subgroup that received phone surveys on placebo content unrelated to time use but responded in-person to the day reconstruction survey; and 2) the ATOC group (the treatment) that received phones and responded to high-frequency calls. Column 3 shows the difference between them, and Column 4 reports the p-value for a t-test testing the null hypothesis of no difference.

Table 2: Phone Survey and Endowment Effects

(a) Phone Survey Effect on Activity Shares

	Economic Activity	Social Activity	Household Work	Care of Others	Personal / Travel
Phone Survey Effect Coefficients	-0.0334**	-0.128***	0.0904***	-0.00751	0.0776***
	(0.0113)	(0.0183)	(0.0210)	(0.0106)	(0.0191)
Constant	0.0832***	0.292***	0.397***	0.0619***	0.166***
	(0.00571)	(0.00927)	(0.0106)	(0.00536)	(0.00968)
Observations	427	427	427	427	427

(b) Phone Endowment Effect on Activity Shares

	Economic Activity	Social Activity	Household Work	Care of Others	Personal / Travel
Endowment Effect Coefficients	0.00396	-0.0189	-0.0144	0.00793	0.0214
	(0.0129)	(0.0165)	(0.0228)	(0.0119)	(0.0163)
Constant	0.0801***	0.309***	0.411***	0.0535***	0.146***
	(0.00631)	(0.00810)	(0.0112)	(0.00585)	(0.00798)
Observations	430	430	430	430	430

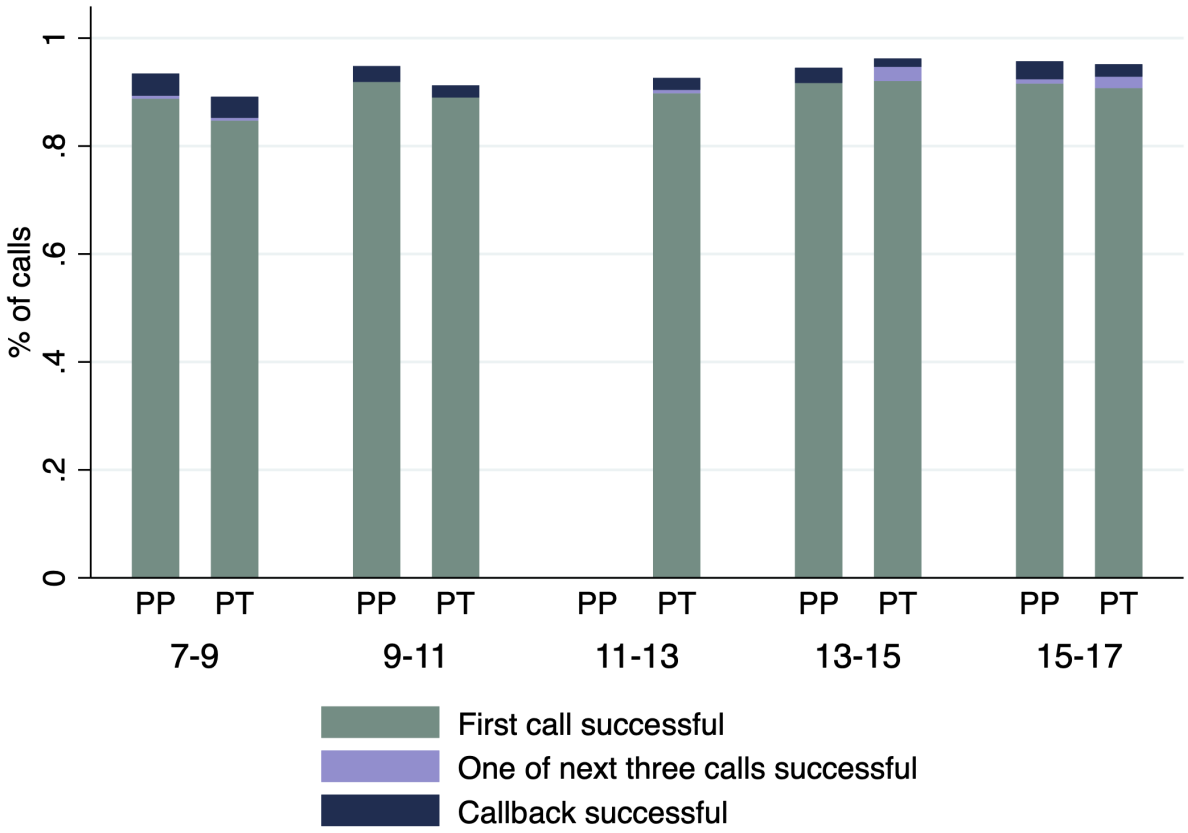
Note. Estimates for Equation 1 are reported for each activity category. The outcome variable is activity shares—see Section 2 or the notes to Figure 3. No controls are included apart from VESA (village) fixed effects. Robust standard errors, clustered at the VESA (village) level, are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Online Appendices

McGavock, McCullough, & Assefa (2023). "Calling for Time: Examining Bias in Time Use Measurement using High-frequency Phone Surveys"

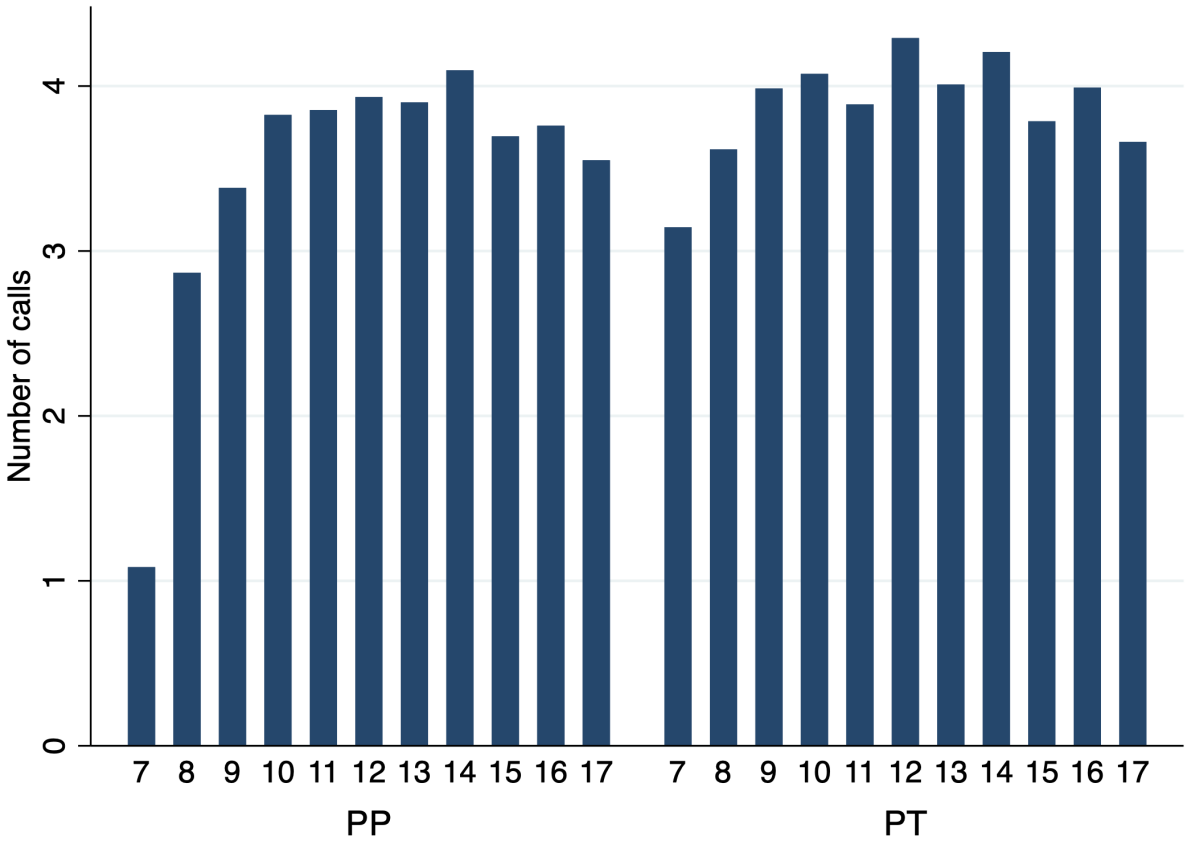
A Supplementary Figures and Tables

Figure A1: Call Success Rates



Note: This figure shows the average call success rate across different day strata for the phone placebo group (DR-PP)—the subgroup of the control group that received the traditional day reconstruction (DR) modules in person but also received phone surveys about their meals, WASH behaviors, and financial transactions—and the phone time use group (ATOC) who responded to questions about their activities at the time of the call, the treatment group. We note the DR-PP group was assigned to receive four calls per day during 4 separate time strata, while the ATOC group was assigned to receive five calls per day during 5 separate time stata. See Appendix B for the hours of the day corresponding with each strata for these two treatment groups.

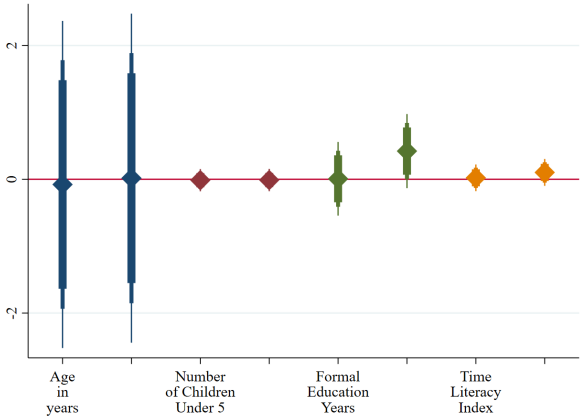
Figure A2: Number of Calls By Hour



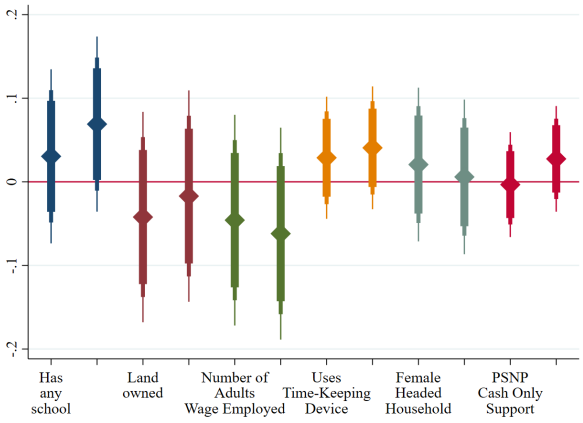
Note: This figure shows the average number of calls respondents received during each hour from 7AM to 6PM in the phone placebo group (PP) and the phone treatment group (PT). See notes to Figure A1 for details on these groups.

Figure A3: Balance Tests

(A) Continuous Variables

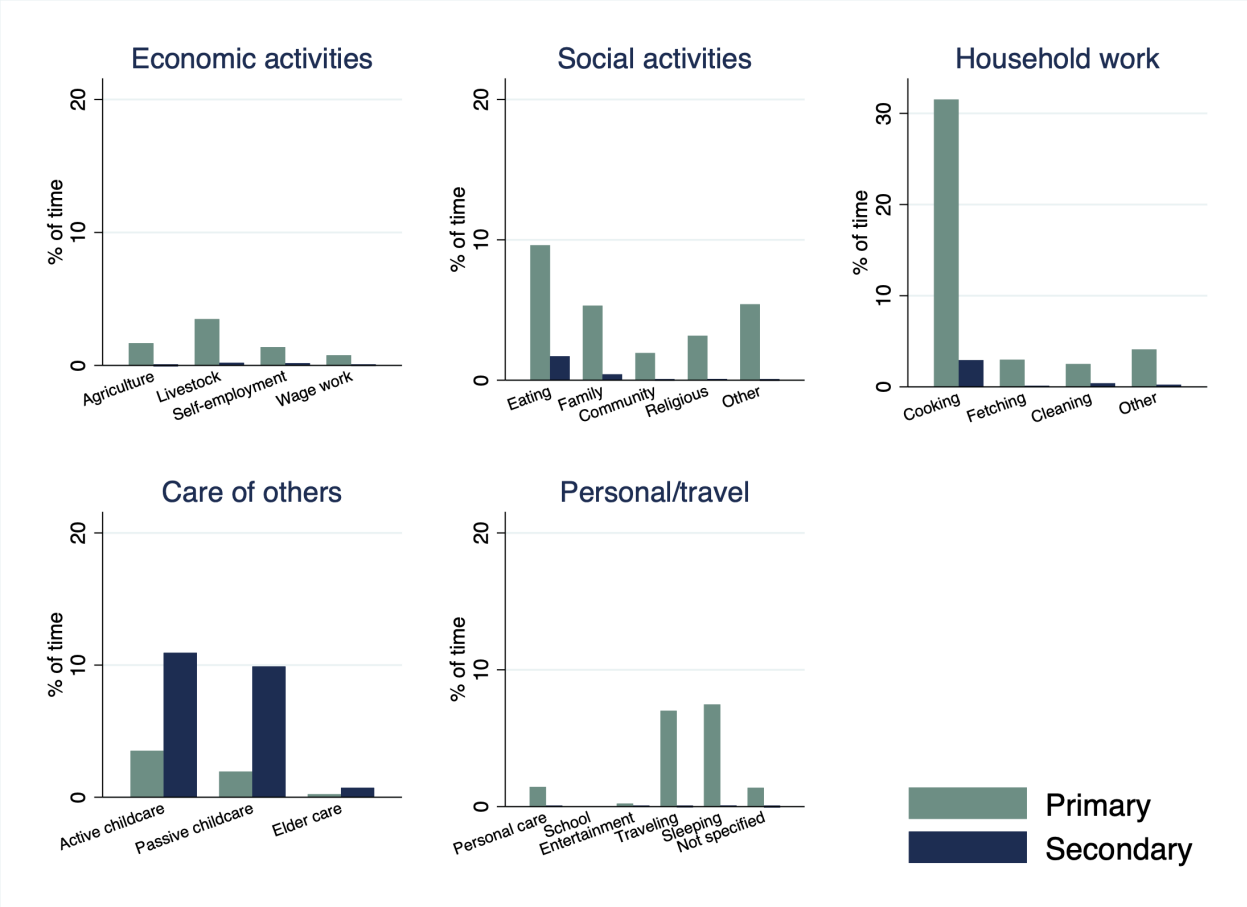


(B) Dummy Variables



Note: To demonstrate that the sample is balanced on observable characteristics from the baseline survey, the figures show coefficients from estimating Equation 1 using baseline characteristics as the outcome variable. For each variable, there is a pair of coefficients, one for the ATOC group and one for the DR phone placebo group. The omitted category is the DR control group.

Figure A4: Primary and Secondary Subcategories - Pooled Data



Note: These figures show the average primary and secondary activity shares for women in the pooled data by sub-category within the five broad activity categories. "Other" social activities is a broad term including all activities not covered by the other four social activity distinctions. "Other" household work activities include weaving, sewing, textile care, washing clothes, shopping, and getting services.

Table A1: Phone vs All IPTU Effect on Activity Shares

	Economic Activity	Social Activity	Household Work	Care of Others	Personal / Travel
Phone Survey Effect Coefficients	-0.0321***	-0.136***	0.0795***	-0.00318	0.0912***
	(0.00869)	(0.0169)	(0.0171)	(0.00815)	(0.0160)
Constant	0.0818***	0.300***	0.404***	0.0576***	0.156***
	(0.00290)	(0.00566)	(0.00572)	(0.00272)	(0.00536)
Observations	646	646	646	646	646

Note: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A5: Number of Children Impacts the Implied Recall Errors —Primary Activity

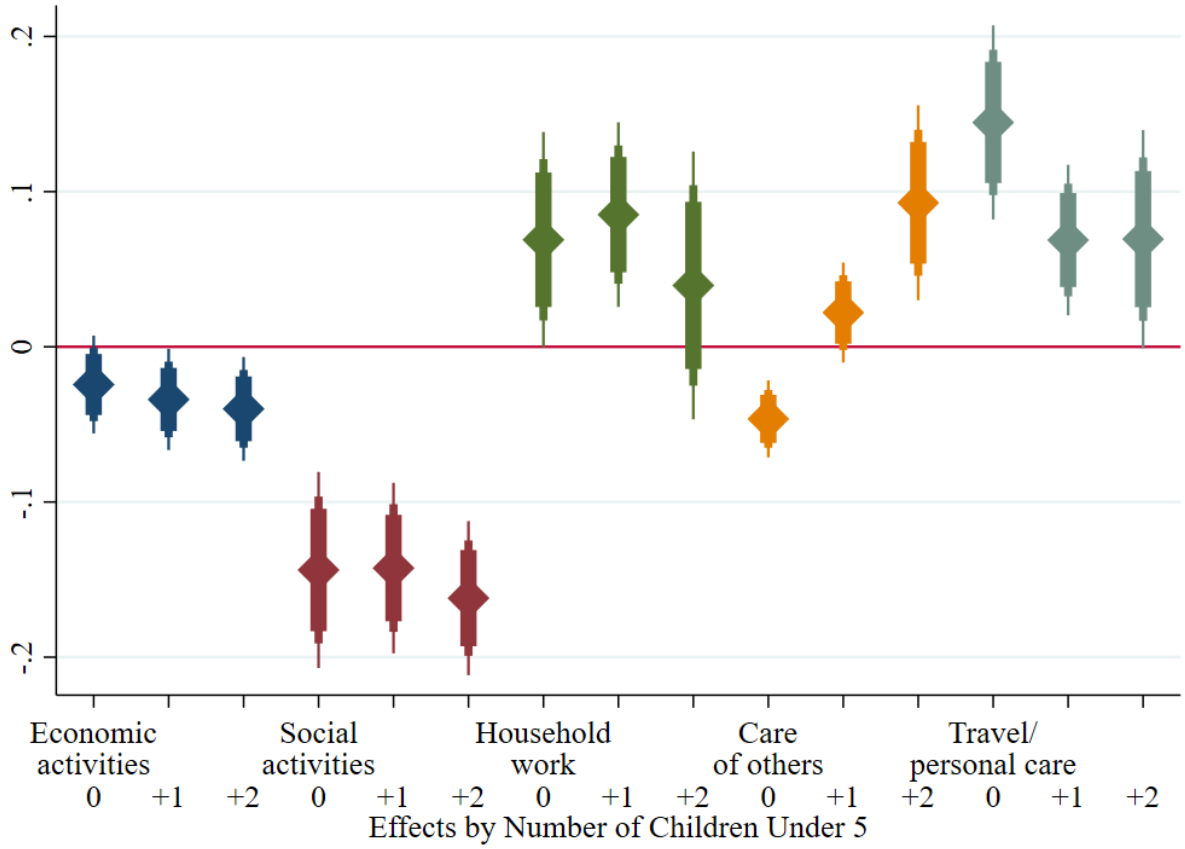


Figure A6: Number of Children Impacts the Implied Recall Errors —Secondary Activities

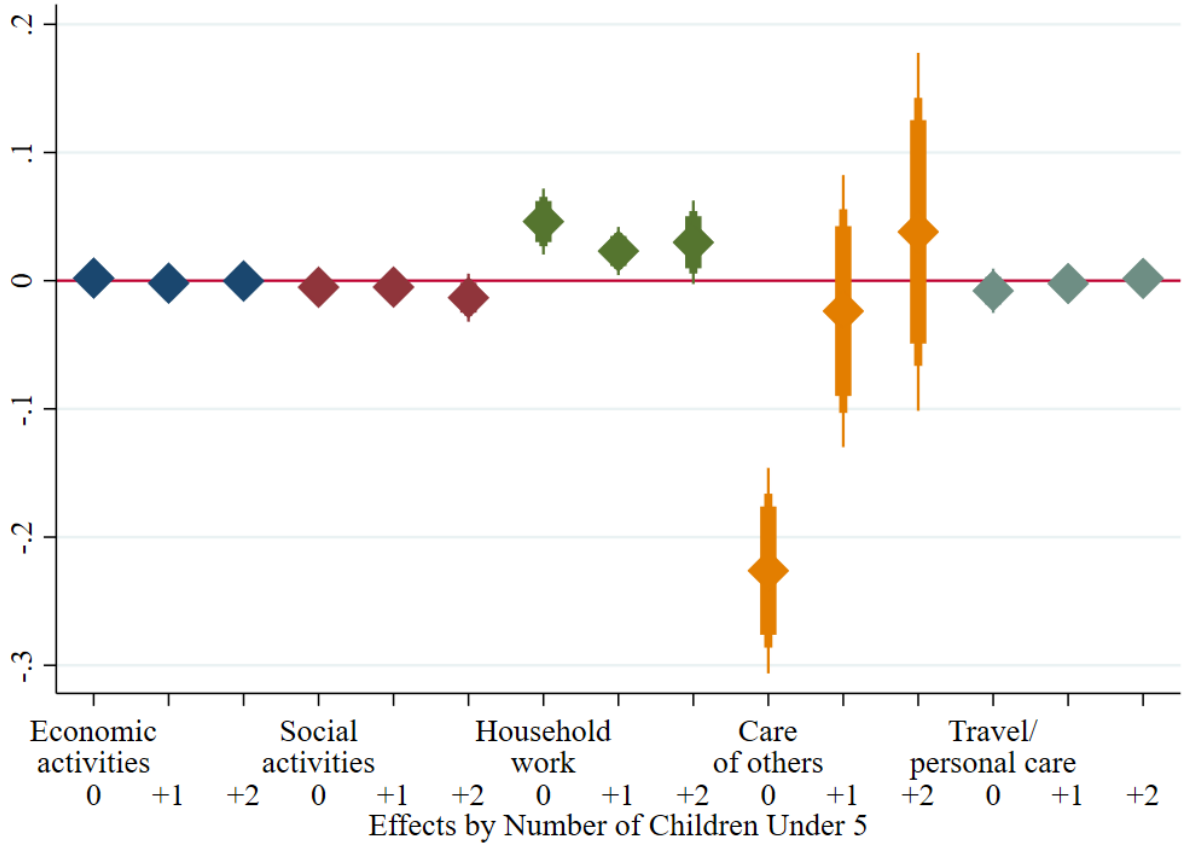


Table A3: Phone Survey Effect on Activity Shares (Wave 2 to 6)

	Economic Activity	Social Activity	Household Work	Care of Others	Personal / Travel
Phone Survey Effect Coefficients	-0.0311*	-0.148***	0.105***	-0.00352	0.0769***
	(0.0128)	(0.0166)	(0.0220)	(0.0107)	(0.0194)
Constant	0.0807***	0.311***	0.381***	0.0585***	0.168***
	(0.00653)	(0.00848)	(0.0112)	(0.00547)	(0.00993)
Observations	366	366	366	366	366

Note: the table reports the coefficient on ATOC for each time category's activity share, and the comparison group is DR-PP (DR-NP is excluded). For the time use category listed in each column header, the table reports the coefficient on ATOC treatment, with the DR-PP group as the comparison group and DR-NP group excluded from the regression. These regressions exclude Wave 1 of survey data. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

B Supplementary Details on Methods and Experimental Design

This appendix provides details on the survey design, including sampling, survey timing, and randomized assignments.

B.1 Sample and Sampling Frame

All households were beneficiaries of the Productive Safety Net Program (PSNP). Some labor-scarce households (e.g. female-headed households) receive direct cash support from PSNP, but the majority of beneficiaries receive support in exchange for their labor contributions to public works projects during the agricultural off-season. Pregnant women are exempt from PSNP public works, but another household member typically provides labor in her place. All households were also beneficiaries either of CARE Ethiopia's Livelihoods for Resilience (LfR) *or* World Vision Ethiopia's Strengthen PSNP4 Institutions and Resilience (SPIR). Both programs were implemented through the Organization for Rehabilitation and Development in Amhara (ORDA), a local NGO.

We focused on the North Wollo zone (Meket and Wadla *woredas*) because that is where both NGOs, who jointly funded our research activities, overlapped in their implementation. A list of VESAs comprised our sampling frame, as VESAs were formed well in advance as part of the NGO programming. VESA members represented about 15 unique households on average, with some variation between five and thirty. Household heads and their spouses, if applicable, were encouraged to join VESAs together. Given the VESA rosters, we performed sample size calculations for each pre-specified hypothesis to confirm that the minimum detectable effect sizes were reasonable. These calculations appear in our pre-analysis plan, which was posted in June 2020 as a registered trial with the American Economic Association (AEARCTR-0005961). From the complete roster of about 700 eligible VESAs in the study regions, we randomly selected 60 to recruit into our study, the largest sample for which we could fund data collection.

VESAs met regularly (two to four times per month) to contribute to their savings and to receive training in livelihoods activities (e.g. how to fatten goats and sheep), beneficial sanitation and hygiene behaviors (WASH), women's empowerment issues, and nutrition. All households who qualify for PSNP were invited to join the VESAs, and CARE and World Vision staff aimed to recruit 100% of those eligible into VESAs. Nonetheless, not every household participates, and membership can fluctuate. Thus, our sample may not be completely representative of the community of poor households in the area. We explore the implications of this for generalizability of our findings in Section 7.

B.2 Survey Waves: Randomizing the timing surveys in VESAs

We assigned sampled VESAs to eight waves, and used stratified randomization to ensure balance across waves in the average distance of VESAs from market towns and in VESA size. This staggered design served several purposes: 1) we were interested in observing seasonality in time allocation; 2) we were able to maintain continuous employment of a small group of high-quality enumerators by smoothing the workload over the course of a year; and 3) we could start and stop data collection if necessary without jeopardizing the research design. This turned out to be a saving grace, as data collection began in January 2020 and, following a series of interruptions due to COVID-19, a contentious election, and uncertainty over civil conflict in Tigray, concluded in August 2021. Due to the spread of violence from Tigray region to our study area in August 2021, we discontinued data collection again out of concern for the safety of our staff and our respondents. When travel was again possible in the region, we decided that resuming the study could pose significant risks to respondents and survey staff and would generate data with potentially limited comparability to pre-conflict data, so we decided to end the study after collecting data from 75% of the intended sample. In total, we were able to complete six waves of data collection, covering 46 VESAs with 646 female respondents from unique households.

B.3 Field team logistics

B.3.1 Baseline interview procedures

Within each wave, field enumerators visited VESAs in a pre-specified random order and followed a highly detailed protocol for data collection that covered a nine-to-eleven-day period in each VESA (see Figure 1A). Specifically, VESAs were assigned an order of visits such that 1) limited the need for the field team to move their home base (a small town with accommodations for enumerators and access to the VESAs) and 2) balanced travel distances within home-base towns. First, all available field enumerators performed baseline interviews with respondents over two-to-four-days, depending on team capacity and the number of households in each VESA. The order of baseline interviews for households in each VESA was also randomly assigned. Each baseline interview lasted approximately two hours and covered a complete household roster, the respondent's different types of economic activities (e.g. wage labor, self-employed business activities), household asset ownership, and a condensed version of the project-level Women's Empowerment in Agriculture (PRO-WEIA) survey instrument (Malapit et al., 2019).

B.3.2 Phone provision

At the end of the baseline interview, enumerators provided a new but basic cell phone (*not* a smartphone) and SIM card to respondents who were randomly assigned to receive phone surveys over the 7-10 day period following the baseline interview. We distributed phones to all respondents in our study to ensure that low phone ownership rates among our target population did not generate selection bias in the sample relative to the target population. The phones were among the most basic and inexpensive models available on the local market, selected for their durability and long-lasting batteries. We also selected a model with few additional features so that respondents would not use them much for other purposes during our 8-10 day survey period.

Enumerators followed a checklist to ensure that the respondent consented to receiving calls, knew how to answer calls (including a practice call), and understood that the phone would be theirs to keep after the survey period, but that for the following 7-10 days, the female respondent would need to carry the phone everywhere and conserve the battery. Each night, enumerators communicated the household roster information for these households to the staff in our call center in Addis Ababa. Phone calls to respondents, which are described in the following sections, began the day after their baseline interview.

B.3.3 Informed consent and preventing misreporting

During baseline interviews and the informed consent process at the beginning of a survey respondent's first phone interview, we took care to inform respondents about the nature of our work and how their data would be used. We communicated that they would not benefit from participation in our study in any way directly (apart from receiving the phone), and ensured that they understood our purposes. We explained that we were collecting information about their lives in order to understand them better, so that we could assist in designing programs that would benefit them. Finally, while we made it clear that participants would not stand to benefit from responding to our surveys in any particular way, respondents knew that our team was connected with the organization that had been working in their villages for some time.

B.3.4 Validation period procedure

Once all baseline interviews were completed in a given VESA, all but one enumerator moved to the next VESA to conduct their baseline interviews. One enumerator remained behind to complete a second in-person interview with each respondent, which occurred on a pre-specified randomly assigned day during the VESA's 7-day "validation period." The content of this in-person "validation interview" differed by treatment group, as described in the next sub-sections. Households also continued to receive multiple phone calls per day during the validation period, according to their treatment assignment.

Respondents who had not received a phone at the end of their baseline interview received one at the end of their validation interview. The complete control group also received a phone so that we would not introduce conflict within VESAs, a request from our sponsor organizations. At the end of the phone survey period, we asked respondents how they had been using the phones and how they planned to use them in the

future. Once the survey period was complete, we reminded respondents that the phones belonged to them and they could use them as they wished.

B.4 Time use surveys

The content of the interviews was the same as the ATUS day reconstruction module and was copied from the WEAI survey instrument to ensure that we produced data that would be comparable to the status quo for time use studies. There are some small ways that our interviews differ from these other time use surveys. Relative to the ATUS, which conducts these interviews by phone with respondents, our module differs in that we require activities to fit into 30-minute blocks of time, whereas ATUS allows shorter or longer intervals that are not rounded to the nearest 30 minutes.

Since some activities may be substantially shorter than 30 minutes, we allowed secondary activities of any type to be recorded, whereas the only secondary activity the ATUS allows to be recorded is childcare. In these respects, however, our survey is consistent with other time use surveys conducted in the developing world, although some use 15-minute blocks, like the WEAI, and others use 60-minute blocks (choices that affect how long the questionnaire takes and are likely related to how closely connected time use outcomes are to the primary purpose for the survey). Unlike those survey modules in the WEAI surveys, we asked respondents about their location and companions for each time block, which is consistent with the ATUS methodology. We use these additional pieces of information to supplement measures of activity measures as we consider mechanisms in Section 5.5.

We actually covered only 17 hours of the previous day (between 5am and 10pm) to speed the data entry process, a common choice for time use modules in rural parts of the developing world where electrification rates are low. We separately asked about sleep hours and interruptions that occurred during the night. This worked in practice for our study setting because few of the households in our study sample had electricity, so many people report going to sleep and rising with the sun. Further, sunrise and sunset occur at about the same time every day throughout the year in the study locations due to their proximity to the Equator (9° latitude). We discuss the data we obtained on women's activities during the late evening and early morning hours and their sleep interruptions in Section 4. Only 31 respondents (7%) reported an activity other than sleep for the first 30 minute period of the day (5-5:30am), with the majority of them listing cooking as their primary activity during that time. For the final 30 minute period of the day (9:30 to 10pm), 108 respondents (25%) reported an activity other than sleep, typically "sitting with family or eating."

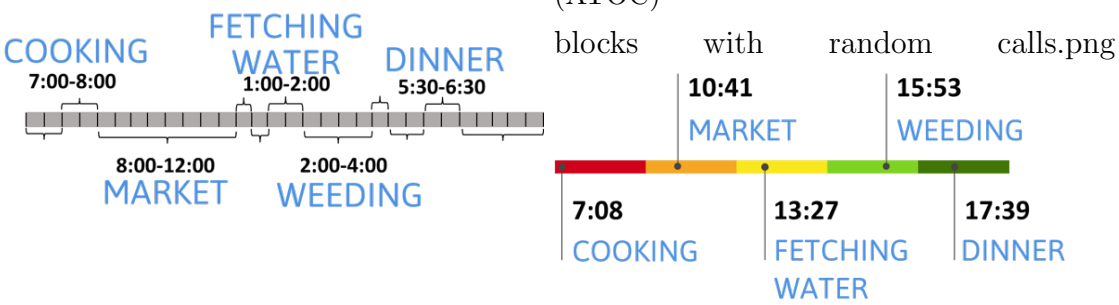
Each station's call schedule was randomly generated such that each respondent received her first call of the day before any respondent received her second call of the day, each received her second before any third calls were placed, etc.

Women who responded to time use surveys by phone also reported on their activities from the short, "bounded recall" period between phone calls (up to two hours). Specifically, we asked "Since the last time we spoke with you, is there anything else you have been doing?" and if there was, we recorded her self-reported duration of each activity in minutes.

Figure B1: Survey Methods

(A) Day Reconstruction (DR)

(B) Activities at the Time of the Call (ATOC)



blocks with random calls.png

Note: These figures show the structure of the survey methods. Panel A shows that the 24-hour recall surveys recorded primary (and secondary) activities for 30-minute blocks of time. Panel B shows the randomly timed, high-frequency phone surveys during which respondents reported on their activities at that moment.

Figure B2: Treatment Groups

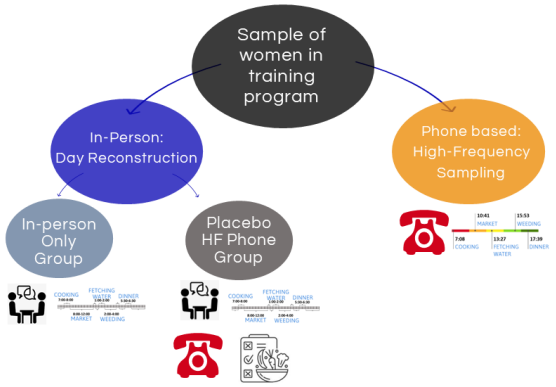
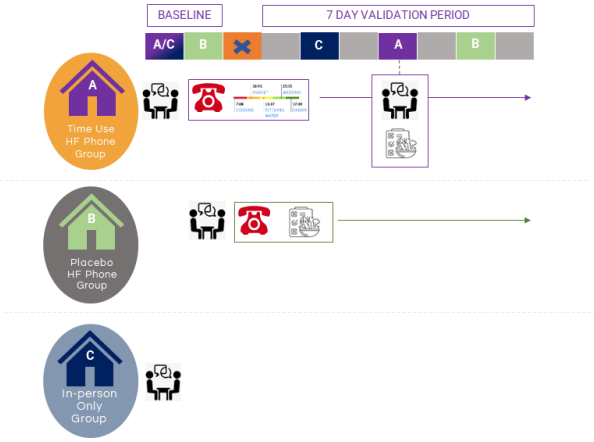


Figure B3: Survey Activities in VESA



B.5 Enumerator training and logistics

We first met with call center and field enumerators together to provide an introduction to the research hypotheses, so that they understood the importance of accurate measurement as a core research priority. We provided very detailed training and referenced materials on how to categorize activities so that there was no ambiguity, including detailed reminders to enumerators within the survey instrument program on their CAPI devices. For example, some women in our context are engaged in "cottage industries," selling some of what they cook for the household to others (e.g. *tella*, a local fermented drink, or *injera*, the staple eaten with most meals). Enumerators were instructed to record these types of activities as economic activities in the event that respondents planned to sell *any* portion of their production. This means that our economic activities activity shares are a conservative upper bound of time spent on livelihoods. We also emphasized extensive mock interviewing during training, with numerous possible scenarios, and adapted survey procedures in response to problems identified during mock interviews.

B.5.1 Enumerator shifts, compensation, and attrition

On most days, there was only one call center enumerator making calls to phone time-use respondents at a time. However, because we made phone calls each day over 11 hours, enumerators worked in shifts. Across different days, enumerators alternated between phone time-use and phone placebo survey stations so that their conversations varied and maintained their interest.

Enumerators were well compensated. During periods of extended pauses in data collection due to COVID-19 and security concerns around civil conflict and uncertainty in Ethiopia, we paid maintenance salaries so that enumerators did not seek employment elsewhere.

We were fortunate that we did not experience any attrition among call center enumerators over the 20 months of data collection. In total, phone time-use respondents were exposed to five enumerators, four primary enumerators and a supervisor who placed calls only on very busy days in the call center. Importantly, enumerators did not have specific shifts assigned to them week after week, so there is no systematic correlation between the day of week or time of day and the identity of the enumerator.

B.5.2 Gender of enumerators

To ensure both high compliance rates with answering phone calls and the safety of all respondents and enumerators, the field enumerators were male and the call center enumerators were female. Out of concern for the safety of our enumerators, we could not employ female enumerators to conduct field interviews. We exclusively employed female enumerators in the call center to avoid causing suspicion or anger from other household members—particularly our respondents' spouses—about the respondents receiving multiple calls per day from unknown men, which could affect attrition. This also generalizes well as call centers in Ethiopia are nearly always staffed by women. Thus, they were also the most experienced call center enumerators we could hire. We discuss possible issues related to enumerator effects in Section 5.5, but we are not concerned that the gender identity of the enumerators accounts for the results that we estimate.

We anticipated that respondents would build a rapport with enumerators that would ensure a level of trust required for honest responses. As Figure **??** showed, there are *no* systematic changes across a respondent's days-in-survey in their activity shares.

B.6 Construction of outcome variables: activity shares

Thus we generate a set of dummies—one for each broad activity category—indicating whether or not the respondent reported performing an activity in that category at the time of the phone call. Taking the mean of these dummies across a respondent's full set of calls gives the share of calls during which the respondent reported performing that activity.

For respondents receiving in-person recall surveys using the day reconstruction method (24HR, the control group) we have a pseudo-panel in that each respondent has one observation for each 30-minute time period in the survey module, from 5am to 10pm. Thus we also generate dummies for each activity category. After limiting the observations for each respondent to those hours that overlap with the call center operating hours, taking the mean of these activities by respondent generates an activity share analogous to the shares for the ATOC group.

B.7 Call center procedures

B.7.1 Callbacks: unanswered, missed, or interrupted call procedures

In case of unanswered calls, we designed our phone based questionnaires and call center protocols to recover as much information as possible about activities at the time of randomly assigned interview times. Interrupted calls and missed calls would lead to callbacks during which the respondent would be asked to identify the reason for the interrupted/missed call.

If a call is missed three times, the enumerator gives up on calling the respondent for that call and a callback is assigned. The time appointed for a callback is determined by consulting the call center schedule for the next available call slot in the enumerator's schedule. Since calls to respondents are randomly assigned

and grouped by call number of the day, there are call slots during the day that are not assigned in advance to households. These can be used for breaks for enumerators and callbacks. When an intended call is interrupted or unanswered, the enumerator assigns the earliest available callback slot to the respondent (provided that it is at least 24 minutes later).

If a callback is not answered or is interrupted, responses to the questionnaire for the original intended call will be marked as missing and no further effort will be made to recover snapshots of information about activities at the time of the intended call or the time of the callback; however, during the next successful call, the respondent will be asked to recall all activities that took place since the previous successful call, which will mechanically cover activities that would have been recorded had the intervening call been successful. During a callback, information about the respondent's current activity and their activities during bounded recall periods (i.e. what happened since we last spoke) are recorded. If possible, information about the respondent's activities *at the time of the original, intended call* is also recorded. During a callback based on an interrupted call, the respondent will be asked to identify, if possible, the reason for the interruption (e.g. battery died, lost connection, could not speak and hung up). If the reason for an unanswered intended call is that the respondent's line was busy at that time, during the callback the respondent will be asked to identify, if possible, the reason and partner from that phone conversation.

Both field and call center enumerators helped to troubleshoot network issues at respondents' homes (e.g., instructing them to shift location slightly). While we had instructed respondents that it was their responsibility to charge the phone, some respondents lived far from centers where they could do so. So in some cases field enumerators who were still conducting in-person interviews during the same period visited respondents whose batteries had died to exchange them for a full one. This should be considered when planning and budgeting for future surveys (which we discuss in Section 7).

B.7.2 Special cases: bounded recall for the evening before, the morning of the call, and callbacks

There are two important deviations from the aforementioned bounded recall protocol. First, during the first call to a respondent on each day (between 7am and 9am), enumerators asked about two separate bounded recall periods of endogenous length. One period covered the evening before, starting with the last call of the previous day and ending with the time the respondent reported having gone to bed the previous evening. The second period covered that morning, starting with the respondent's self-reported wake-up time and ending with the time of the first call of the day. In total, each 24-hour period consisted of 6 bounded recall periods: the last few hours of the previous evening, the early morning, and the four periods between the five calls placed that day. Because the phone calls were placed at randomly assigned times within each day's five strata, the duration of each bounded recall period varies randomly across a respondent's calls.

Unanswered calls generated a second special case for bounded recall periods. On some occasions, a callback was triggered when a survey call was interrupted (e.g. the battery died mid-call or the respondent said she was too busy to talk). More often, a callback was initiated because the respondent did not pick up the phone after the four dials. If a respondent answered the phone at the callback time, we asked her about whether she would be able to tell us what she was doing at the time of the missed call. We recorded this information, but also flagged these observations to confirm that our constructed time shares are not sensitive to whether or not the activity at the time of the missed call is included. We discuss these outcomes in Section 3.3.

C Results Supplement

C.1 Measurement error mechanisms

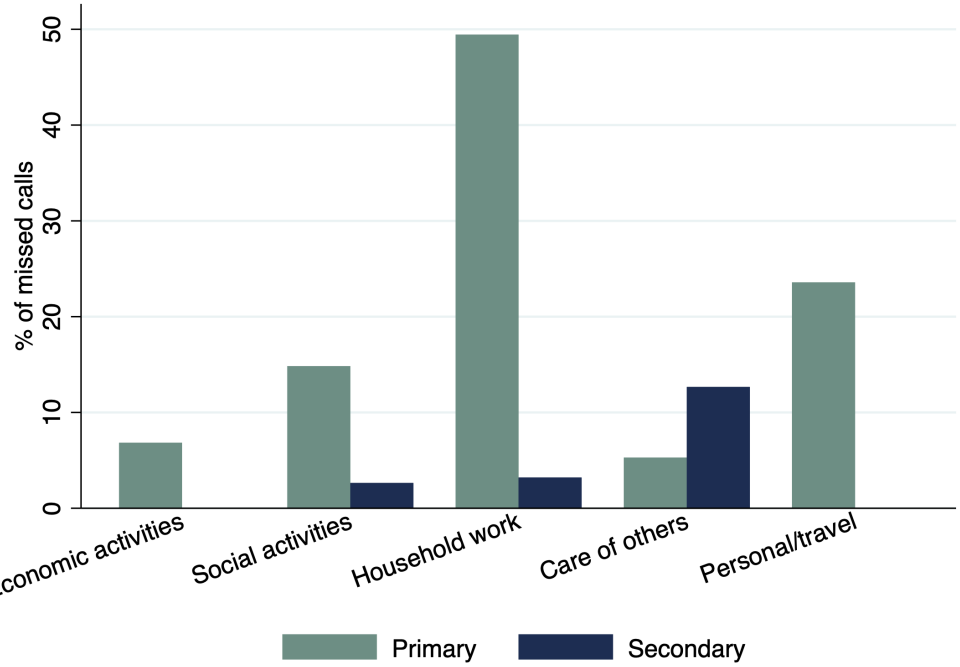
C.1.1 Sampling Errors

Frequency of and reasons for missed phone calls For calls that go unanswered, one might be concerned about within-person selection biases in observed activity shares if respondents are systematically less (or more) likely to answer the phone while doing certain activities. Such selection could arise if the nature of the activity renders the respondent unable to speak on the phone. For example, a respondent could be napping and not hear her phone, she could be carrying a load of firewood with both hands and unable to reach her phone, or she could be visiting a location with poor network quality (such as her fields) or noise pollution (such as a busy marketplace) and fail to answer questions at the time of the call. Selection could also occur if a respondent intentionally ignores phone calls while performing activities she would be ashamed to report, such as taking a break from work, or that she would rather not interrupt, such as watching television or speaking with a customer. We can rule out that the differences in activity shares across survey methods are attributable to systematically different probabilities of observing certain activities at the time of any missed calls.

We address this concern several ways. First, we characterize the extent of missed calls and the nature of these calls using data elicited from respondents during callbacks about their activities at the time of any missed calls. As shown in Figure A1, over 80% of calls are answered by the respondents at the first time they are placed. The also shows that respondents are uniformly likely to answer calls across the call center hours, suggesting that survey compliance is not related to time of day.

An additional 1-2% of calls are answered during three callbacks placed immediately after any missed call. If a respondent did not answer any of these calls, we initiated a protocol to schedule an additional interim follow call during the respondent's next free time slot (see Appendix B). We recovered an additional 3-5% of observations using this callback procedure. As shown in Figure C1, the recovered activity shares are remarkably consistent with the activity shares for successful calls. When recovering information about a missed call, we asked enumerators to record an explanation for why the call was missed. These responses suggest that respondents missed calls for reasons unrelated to their activities at the time of the missed call. Poor network coverage accounted for 50% of the instances when a callback was required, and phone battery issues generated another 10% of callbacks.

Figure C1: Characterizing activities at the time of missed calls



Note. The figure shows recovered data (from a callback) about respondents' activities at the time of a missed call and compares this with the time shares reported by respondents who answered calls at the intended time.

Bounding treatment effects by recoding missed phone calls Callbacks were not always successful before a respondent's next regularly scheduled call, however, and respondents could misremember or otherwise misreport when reporting on the time of a missed call, generating a recall bias. We address remaining bias that could arise from missed calls using a pre-specified bounding exercise to demonstrate that our ATOC activity shares, and the ATOC treatment effect, are not driven by systematic selection effects arising from failing to observe the activity at the time of a missed call (Heckman, 1979). We define several alternative activity shares for each ATOC respondent. First, we bound the activity shares with very conservative set of assumptions — that respondents are *always* (*never*) engaged in each given activity for all missed calls. These form the highest and lowest possible bounds around each activity share. This bounding is very conservative — if respondents are doing economic activities on average 7% of the time during answered calls, it is unlikely they are doing economic activities 100% of the time during non-answered calls. We also bound the activity shares using a slightly less conservative approach. Here, we use the respondent's average activity shares for all calls missed due to network reasons, and then we reassign the unexplained missed calls as always (never) engaged in the activity. These bounds form the *highest realistic* and *lowest realistic* activity shares and are preferred given that missed calls are mostly unrelated to the respondents' current activities.

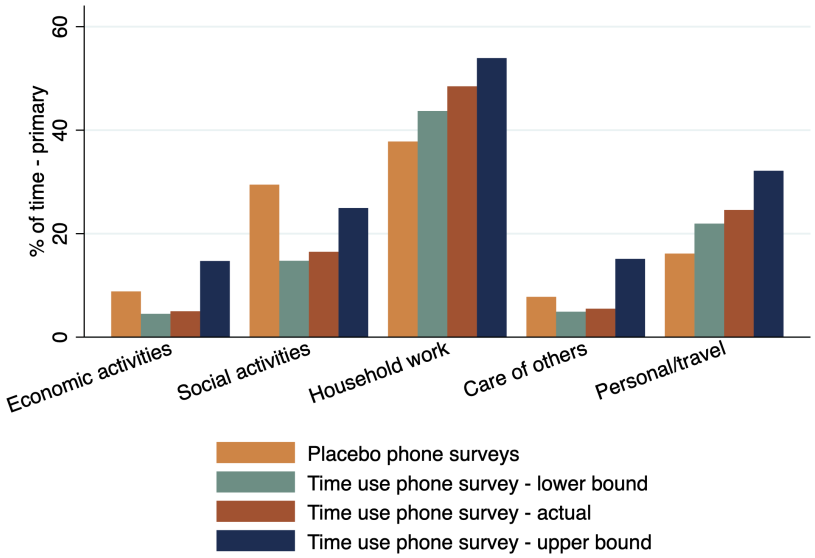
Figure C2A shows highest and lowest possible activity shares calculated using this bounding strategy. Figure C2B shows the estimated treatment effects (using each bound as the outcome variable in Equation 1), in which both the upper and lower bounds on the effects for social activities, household work, and personal/travel keep the same sign as in our main estimates and remain statistically significantly different from zero. The highest possible bounds in Figure C2B for economic activities and care of others are higher than the DR phone placebo group's share, while the lower bounds are lower.

We reiterate that the highest possible and lowest possible bounds are likely much too conservative for economic activities and care of workers, especially, given how infrequently they are reported as a primary activity in the sample in general.

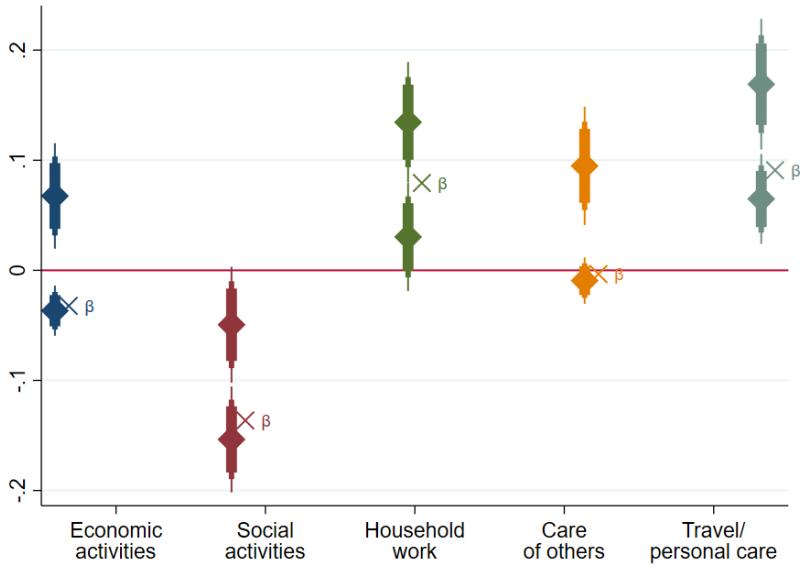
Given these bounding results and the low general prevalence of missed calls, we are not concerned that activity shares from the ATOC method are different from true activity shares. However, we caution that with lower compliance rates and/or a different sample population, missed calls could threaten the precision of activity share measures.

Figure C2: Bounding activity shares by reassigning missed calls.

(A) Summarizing highest and lowest possible shares for each activity



(B) Bounding the treatment effects by reassigning missed calls.

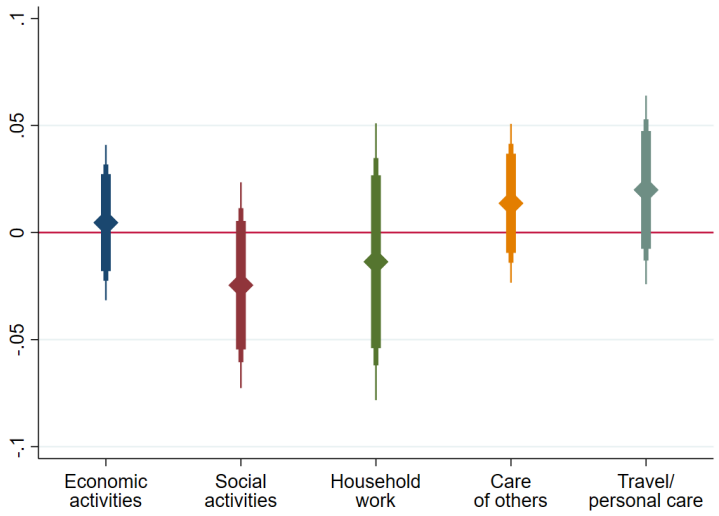


Note. The figure shows upper and lower bounds on the treatment effect on each time category's activity share. The upper bound assumes that the respondent was *always* engaged in that activity when a call was missed and represents the largest possible activity share. The lower bound assumes the respondent was *never* engagged in that activity when a call was missed and represents the smallest possible activity share. The main coefficient reported in Figure 3 is represented by "X β".

C.1.2 Survey-induced changes in true outcomes

Endowment Effect To address concerns that the ATOC treatment could have induced a change in time use through the provisioning of a mobile device to ATOC respondents (or providing small compliance incentives based on the number of calls answered), we compare DR-reported time use between the two DR subgroups — DR-PP and DR-NP. Just like the ATOC group, the DR-PP group was given a phone and participated in multiple phone interviews per day during the data collection period. Figure C3 and Table C1 compare activity shares between the DR-PP and DR-NP groups. Both reported time use through the DR method, and the only difference is that the DR-PP group also was given a phone and participated in phone interviews. As the figure shows, we fail to reject the null hypothesis of different activity shares. This analysis suggests that phone provision did not alter the respondents' time use.

Figure C3: Phone endowment effect



Note: The figure depicts the coefficient on DR-PP treatment for each time category's activity share, with the DR-NP group as the comparison group and the ATOC group excluded from the regression. The confidence intervals reflect p-values of p<0.1, p<0.05, and p<0.01, respectively.

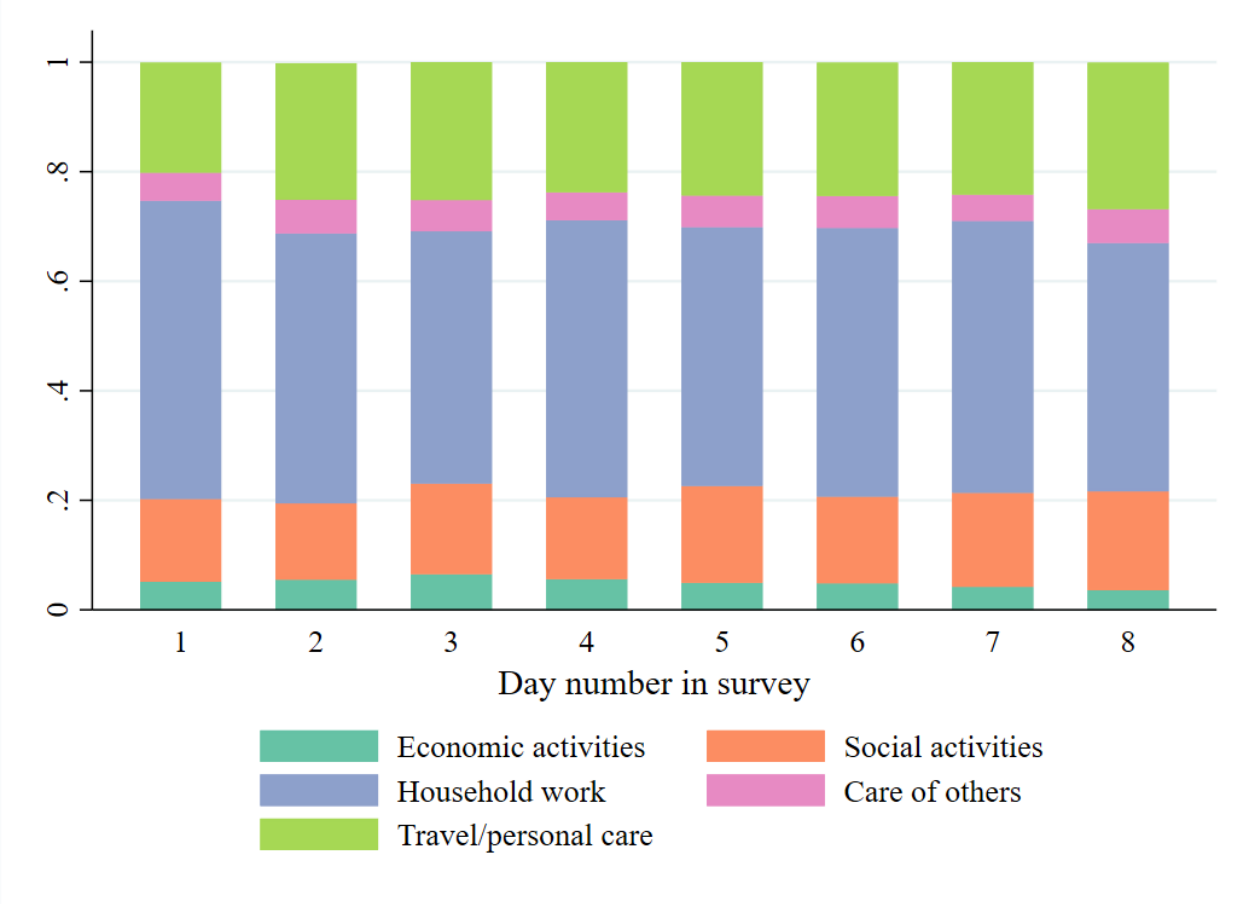
Table C2: Phone Endowment Effect on Activity Shares (Wave 2 to 6)

	Economic Activity	Social Activity	Household Work	Care of Others	Personal / Travel
Endowment Effect Coefficients	0.00334	-0.00420	-0.0220	-0.00100	0.0239
	(0.0146)	(0.0172)	(0.0236)	(0.0127)	(0.0184)
Constant	0.0788***	0.314***	0.402***	0.0586***	0.146***
	(0.00712)	(0.00843)	(0.0116)	(0.00621)	(0.00900)
Observations	366	366	366	366	366

Note: For the time use category listed in each column header, the table reports the coefficient on the DR-PP treatment, with the DR-NP group as the comparison group and the ATOC group excluded from the regression. These regressions exclude Wave 1 of survey data. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Hawthorne Effect We test for Hawthorne effect by looking for changes over time in time use reports (or difference between the first call and subsequent calls) within the ATOC treatment group. Figure C4 summarizes activity shares by day in survey for the ATOC group. We do not see a change in activity shares between the first day and subsequent days, suggesting that repeat questioning about activities does not appear to change the mix of respondents' activities over a one-week period at least.

Figure C4: Activity shares over survey period.



Note: For the ATOC category, the figure summarizes activity shares by day in survey, with day 1 corresponding with the first day of ATOC phone calls.

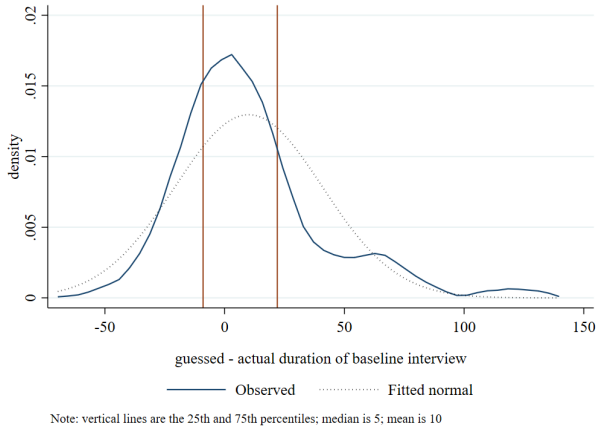
C.1.3 Recall Error Mechanisms

The ATOC effect is larger for women with low time literacy We find a relationship between women's time literacy and the experimental effects. We measured time literacy during the in-person baseline interview by asking each respondent, at the end of the interview, to estimate how much time had passed during the interview. We then compared their estimates to the time-stamped interview duration from the start of the interview to that point. We also asked respondents at baseline how they kept the time and whether they used any timekeeping device.

Respondents were more likely to over-estimate the length of the interview than under-estimate it, and many guessed within 15 minutes of the actual survey length. Figure C5A plots the distribution of guesses, and Figure C5B depicts respondents methods for keeping the time, as reported at baseline.

Figure C5: Timekeeping literacy

(A) Respondents' errors in estimating duration of baseline interview



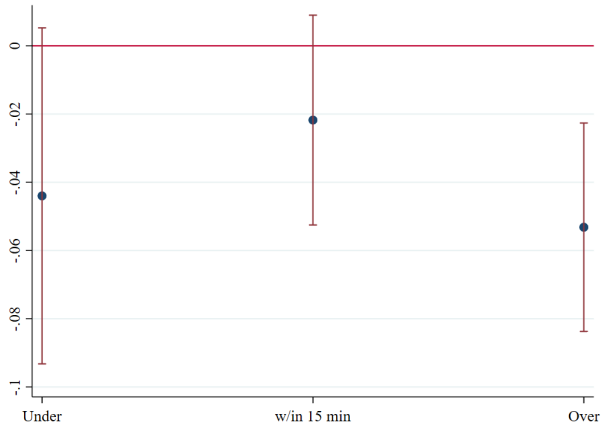
(B) Respondents' methods for telling the time

Note. Panel A reports the probability density function of respondents' errors in estimating the time that elapsed during the baseline interview. Panel B depicts the share of respondents who relied on each different strategy of timekeeping, as elicited at baseline.

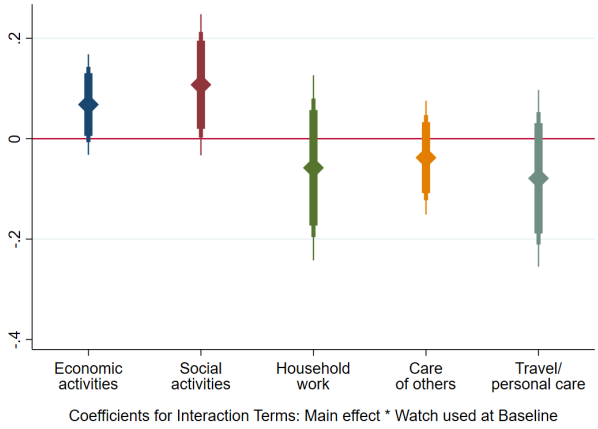
We next re-estimate the treatment effects on the economic activity share outcome for the three sub-groups of respondents based on their time duration estimation accuracy (over-, correct-, and under- estimators) in Figure C6A. We find that those who over-estimate the duration of time by more than 15 minutes are more likely to under-estimate their time spent on economic activities, and this is the only group with a statistically significant coefficient.[30] Figure C6B repeats the main experimental effect of ATOC on reported economic activities for subgroups of respondents according to their strategy for timekeeping. We find that, for four out of five activity categories, the interaction term partially offsets (lessens) the original treatment effect, suggesting that these women report time use with less error than those who use traditional time keeping strategies.

Figure C6: Treatment effect heterogeneity of economic activity share by time estimation accuracy and by timekeeping method.

(A) By accuracy in estimating time duration at baseline



(B) By timekeeping method at baseline

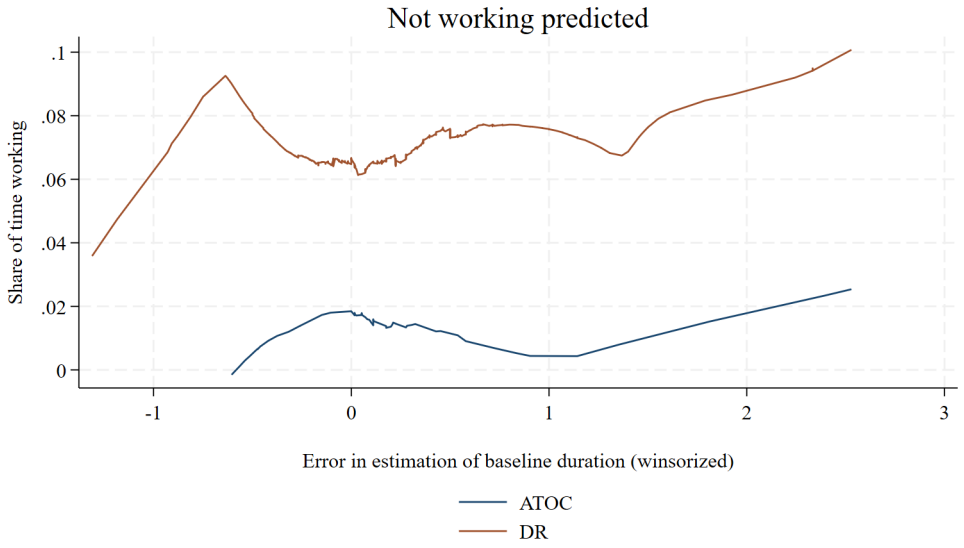


Note. Panel A reports the experimental effect of ATOC on economic activity shares for three sub-samples of respondents according to how accurately they estimated the duration of the in-person baseline interview. Panel B reports the experimental effect of ATOC on economic activity shares for sub-samples of respondents according to their method for timekeeping as measured at baseline.

[30] Again, we may be under-powered a bit here because of our smaller-than-planned sample.

We also show that, within the DR group (but not the ATOC group), the total reported time share for economic activities increases monotonically with the size of the error in estimating the duration of the baseline interview. Figure C7 plots the average economic time share conditional on error in estimating the baseline interview duration separately for respondents in the DR and ATOC groups. The average economic time share is related to respondents' ability to estimate the duration of the baseline interview for respondents in the DR group, suggesting time recall errors are partly responsible for the respondent's estimated economic activity shares. For ATOC respondents, we do not see such a relationship, as we would expect with a reporting method that does not require respondents to accurately estimate time duration.

Figure C7: Economic time share versus accuracy in estimating time duration at baseline for the ATOC and DR groups



Note: By treatment group (ATOC versus DR), the figure shows the average economic time share conditioned on the respondent's error in estimating the duration of the baseline in person interview.

The ATOC effect is larger for women with larger care responsibilities More broadly, the validity of standard 24-hour DR surveys requires that respondents remember their activities and the duration. Some respondents may neglect to report activities that they do not consider noteworthy to the researcher (e.g., passive childcare). Using the data on companions in a similar way, we can identify the extent to which respondents are plausibly performing childcare but do not mention it as an activity in their reports, and whether this differs across the two survey methods. Results on this are forthcoming.

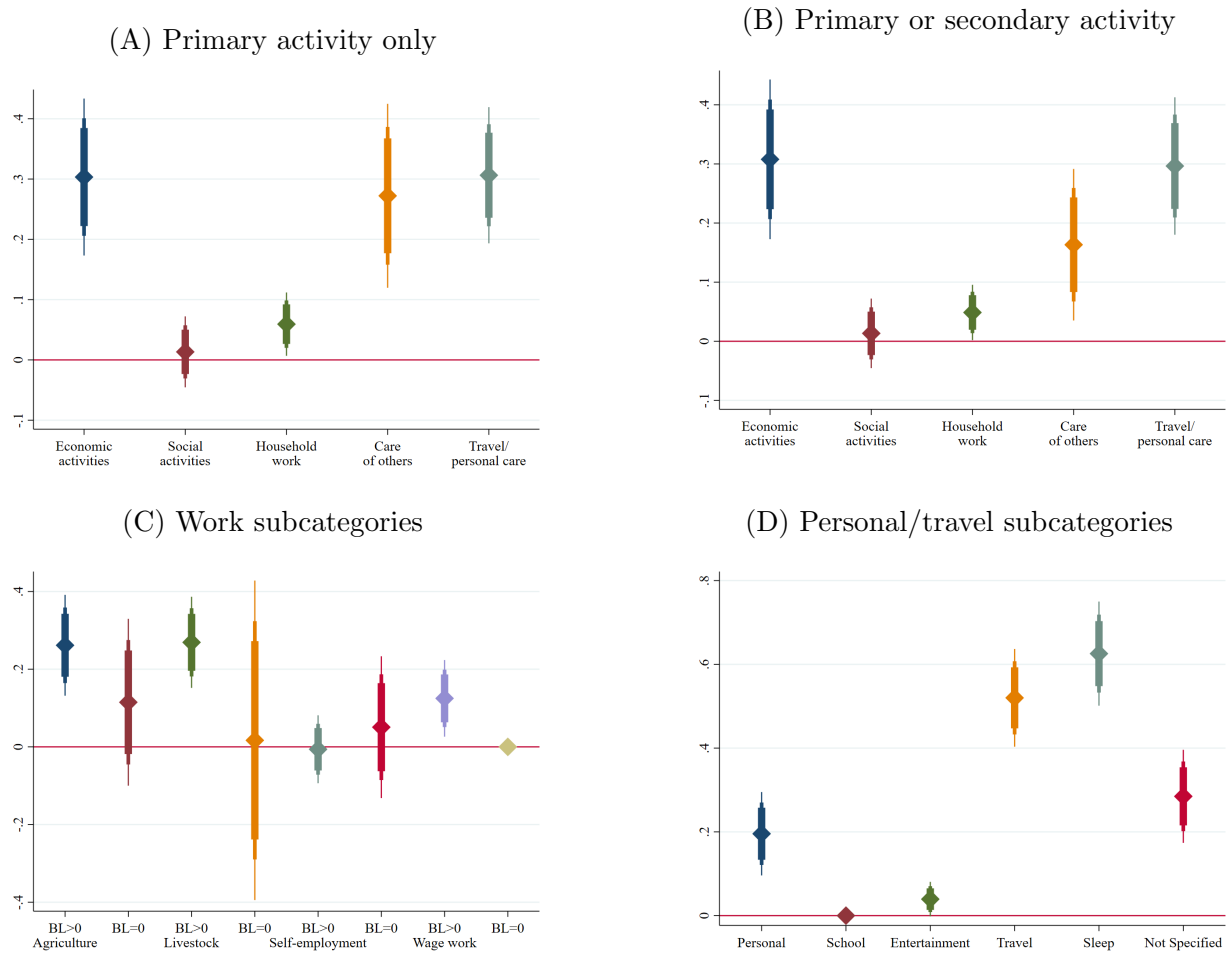
We also test whether the main survey mode treatment effects (Figure 5A) are statistically significantly different for women with different numbers of children. Figure A5 shows the survey mode treatment effects for each of the five activities for women with no children under 5 (the main effect) or women with one or two children under 5.

Respondents are less likely to report brief-duration activities using the DR method Short activities may be omitted from reports during day reconstruction interviews. This can happen for two reasons: 1) respondents do not think to report on them, a respondent-driven effect; or 2) shorter activities are mechanically excluded because respondents and/or enumerators round a longer activity up to the full discrete time block (30 minutes). They might report the shorter activity as a secondary activity in this case, or they might not.

To identify whether DR surveys under-state the occurrence of activities, we repeat our analysis but compare reports from ATOC and DR respondents at the extensive margin. That is, we define a dummy variable for each activity category that takes the value one if the respondent has a non-zero activity share for that category and zero otherwise. If respondents in the ATOC group are more likely to have non-zero activity shares for an activity type that is likely to occur for very short periods, we could consider this evidence that the DR method fails to capture participation by some respondents in certain activities due to their brevity.

Figure C8A shows coefficients from re-estimating Equation 1 using a dummy outcome variable that takes the value one if the respondent's activity share is non-zero (i.e., the respondent reported the activity category at least once). The coefficients for economic activities, care of others, and travel/personal care are positive and of relatively large and similar magnitudes: they all show that respondents in the ATOC group are about 30 percentage points more likely to report any time spent on these activity categories relative to the DR group. If we consider both primary and secondary activities reported, the coefficient on care of others drops by half (Figure C8B) but those for economic activities and travel/personal time remain unchanged. To be clear, it is not possible to tell from this analysis whether the difference in observing an activity at all comes from failing to capture a brief activity in the raw data because of an enumerator's choice or because the respondent fails to report the activity entirely.

Figure C8: Extensive Margin Coefficient Plots



Note. These figures show coefficients for estimating Equation 1 using a dummy variable outcome for whether the respondent reports any time spent on that activity.

For a few activity types, we are able to identify whether we expect to see a given respondent reporting the occurrence of a given activity. First, in the case of economic activities, we have information from the baseline interview on the respondent's participation in livestock rearing, self-employment in small enterprise (e.g., making and selling local beer), agriculture, or wage work during the previous week. We would expect that this would be highly correlated with whether the participant performs economic activities during the survey week (day) as well. Figure C8C shows that the positive coefficient on the extensive margin for economic activities in Figure C8A appears only for the 80% of women who reported participation in one or more economic activities at baseline. The estimate is smaller in magnitude and noisier for the women who did not report participation at baseline, though there may be some some heterogeneity in treatment effects for this subgroup.

Further, Panel B shows that these effects are not limited to a particular type of economic activity, though they are largest for agriculture and caring for livestock (the most commonly reported economic activities at baseline). There is a slightly smaller effect for wage work. Interestingly, the coefficient for participation in agriculture is relatively large: women in the DR group are 12 percentage points less likely to report performing agricultural work in their DR survey than women in the ATOC group.[31] Together, these results

[31]This coefficient has a p-value just under 0.10. We may be under-powered here, however, (recall that civil conflict in our study area forced us to discontinue operations after only 75% of the sample was collected).

provide suggestive evidence that DR methods may understate women's participation in agriculture during the 24-hour period prior to the interview.[32]

Second, Figure C8D shows statistically significant treatment coefficients for activities related to personal care and travel, disaggregated by subcategory. The effects are largest for two activities: travel and sleep.

Finally, we note that our survey procedure itself introduced an activity to each woman's day that was short in duration—any respondent in the DR group who was receiving placebo phone calls was spending about 30 minutes per day on the phone with our survey team but in chunks of 5 to 10 minutes at a time. However, there are *no* instances of DR respondents listing phone survey participation as an activity for a 30-minute period for a primary or even secondary activity. However, the ATOC phone surveys did interrupt in-person surveys on occasion, so that participation in in-person interviews appears as a respondent's primary activity in a few cases. This provides direct proof of concept that short activities adding up to substantial chunks of time across an entire day are frequently omitted in traditional DR surveys. In cases like these, the phone surveys more accurately represented activities.

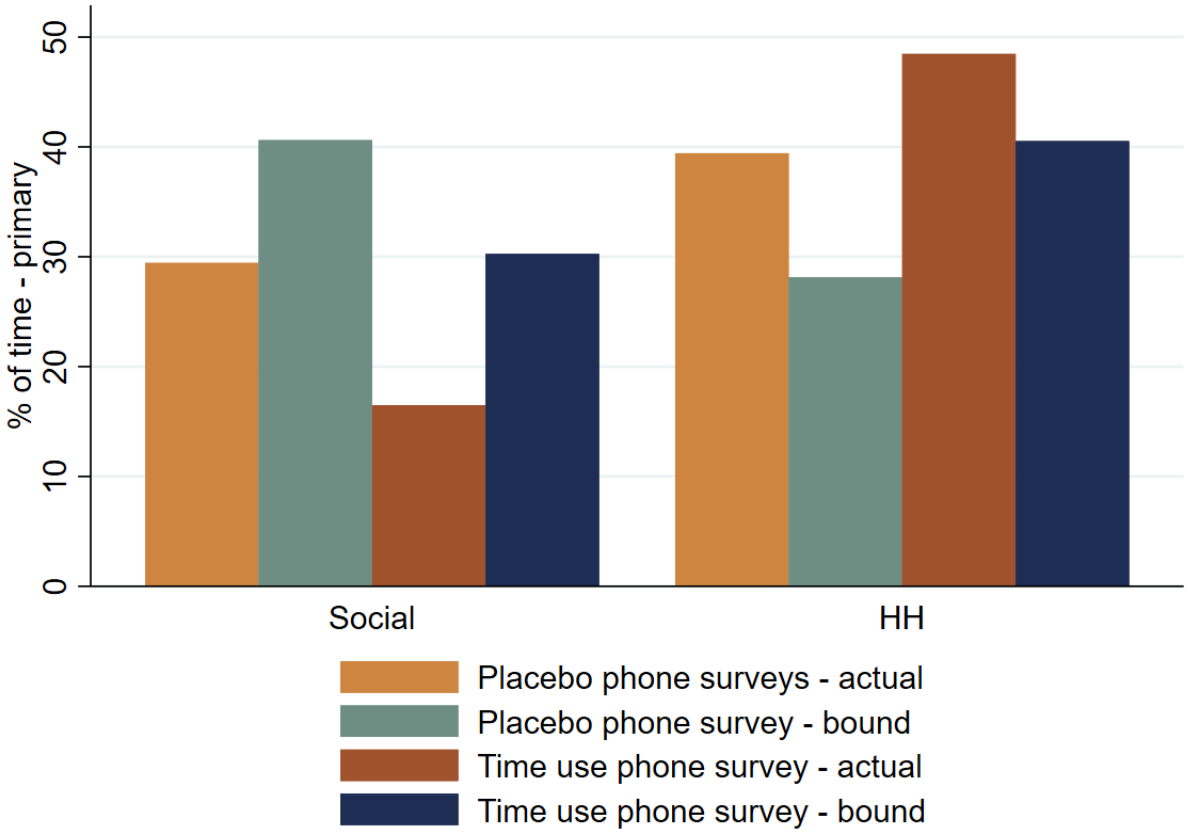
Respondents report simultaneously occurring activities differently using the DR method. This section addresses whether respondents characterize time use differently according to survey method whether they are in the presence of others. According to our main estimates, activity shares for the ATOC respondents are lower for social activities and higher for household chores relative to the DR respondents. Since most respondents live in extended-family households and socialize often with their neighbors, many perform household chores in the company of others. For example, neighbors visit for coffee, a major part of communal life in Ethiopia that occurs in the afternoon before dinner. If women are simultaneously drinking coffee with neighbors and actively preparing a hot meal, they might report the activity differently if asked about it in the moment (by phone) compared to later (in a DR interview).

To test this, we utilize the data we collected about companions (at the time of each call for the ATOC group or during each recall block for the DR group) to construct alternative versions of activity shares for social activities and household work. For respondent i we re-code her activity as a social activity if she was performing household chores but was also in the presence of other adults who were *not* members of her household. This gives an upper bound for the share of time spent on social activities and a lower bound for the share of time spent on household chores.

Figure C9 shows these alternative, hypothetical activity shares separately for the ATOC and DR groups along with the activity shares for these groups as measured using their actual reports. Applying this adjustment changes activity shares for respondents from both groups, the ATOC and DR, and thus cannot explain the difference in activity shares across the two survey types. For example, the gap between the actual shares of time for social activities across the two groups is 13 percentage points—respondents in the placebo phone survey group (far left, orange) reported social activities for 29% of their recall periods while time use phone respondents reported social activities during only 16% of calls. After re-coding chores as "social activities" if outside adults are present, the ATOC group's share increases by 15 percentage points to 31%, and the DR group's share increases by 12 percentage points, to 41%. The treatment effect is robust to this reclassification exercise, suggesting that respondents are not reporting simultaneous social activities differently across methods.

[32]To ensure that the effect on the extensive margin is not entirely driven by activities that occur only on specific days, we re-estimate the effects limiting the ATOC group's responses to only the day that corresponds to the day prior to their validation interview.

Figure C9: Bounding Time Spent on Social Activities



Note: The graph shows the actual and alternative activity shares for social activities and household work, for both placebo and time use phone survey respondents. The alternative activity shares for the social category were constructed by augmenting a respondent's self-reports of primary activities in the social category with information about whether the respondent was with other people apart from household members. For example, the average respondent in the phone placebo group indicated social activities as her primary activity in 30% of the 30-minute time periods that overlap with the call center hours (orange bar farthest to the left). The green bar next to it, at 40%, can be interpreted as the upper bound on social activities for that group if every time that a woman reported a different primary activity but was also with non-household members (family, neighbors, or others), her activity were recoded to "social." Comparing these two bars shows an increase of 10 percentage points for the social activity time share for placebo phone survey respondents. Similarly, the time use phone survey respondents have an increase of about 14 percentage points (comparing the darker red bar to the navy blue bar for social activities). The difference-in-differences (4 percentage points) is not sufficient to close the 13 percentage points gap between the social activity shares using the actual responses across the placebo phone and time use phone groups.

D Lessons & Applications (See Section 6)

D.1 Applications to Productivity Estimates

E Structured Ethics Appendix

- **Policy Equipoise**
  Is there policy equipoise? That is, is there uncertainty regarding participants' net benefits from each arm of the study relative to the other arms and to the best possible policy to which participants could have access? If not, ethical randomization requires two conditions related to scarcity: (1) Was there scarcity, i.e., did the inclusion of multiple arms change the expected aggregate value of the programs delivered? (2) Do all ex-ante identifiable participants have equal moral or legal claims to the scarce programs?

  This was a methodological study, so none of the three treatment arms was welfare enhancing relative to the others. Specifically, in the primary treatment arm women were called multiple times a day and asked about their time use (PT), while other women were called multiple times a day and asked about their eating and hygiene habits and financial transactions (PD), and a third group received no phone calls but was only asked about both items in an in-person interview (CC). The total time that respondents participated in the study was under two weeks. All the women were asked questions about all of the same content, and all of them received phones as part of their participation in the study. While women in the two phone groups (PT and PD) were provided with phones up to 9 days earlier than women in the CC group, we are not concerned about this time lag being an ethical issue in this context, since there was no local emergency during the data collection period for which a phone would be especially valuable. Thus we believe we have achieved policy equipoise across the three arms of our study.

- **Role of Researchers in Implementation**
  Are researchers "active" researchers, i.e. did the researchers have direct decision making power over whether and how to implement the program? If yes, what was the disclosure to participants and informed consent process for participation in the program? Providing IRB approval details may be sufficient, but further clarification of any important issues should be discussed here. If no, i.e., implementation was separate, explain the separation.

  The researchers were active in the design phases of the project, training of enumerators, and piloting, but did not engage directly with participants during the core research period. We worked with EconInsight Centre for Development Economics, whose staff we trained on our detailed protocols for the methodological experiment. We were available to troubleshoot logistical and technical issues during the survey periods.

- **Potential Harms to Participants or Nonparticipants from the Interventions or Policies**
  Does the intervention, policy or product being studied pose potential harm to participants or non-participants? Related, are participants or likely affected non-participants particularly vulnerable? Also related, are participants' access to future services or policies changed because of participation in the study? If yes to any of the above, what is being done to mitigate such risks?

  The participants in the study came from low-income households that were participating in graduation-from-poverty programs in Amhara region, Ethiopia. Participants' socio-economics status makes them a vulnerable population, but our project and the randomized "intervention" (answering survey questionnaires via two different modalities) effectively posed no potential harm. Our informed consent statements made it clear to potential respondents that their participation in our study would not ever impact the household's benefits from nor ability to participate in poverty reduction programs, whether through CARE Ethiopia, World Vision Ethiopia, the PSNP 4 program, or otherwise.

- **Potential Harms to Research Participants or Research Staff from Data Collection**
  Are data collection and/or research procedures adherent to privacy, confidentiality, risk-management, and informed consent protocols with regard to human subjects? Are they respectful of community norms, e.g., community consent not merely individual consent, when appropriate? Are there potential harms to research staff from conducting the data collection that are beyond "normal" risks?

  This study was declared exempt under regulations 2.iii and 3C by the University of Georgia (UGA) Human Research Protection Program on June 20, 2019 (study PROJECT00000661), and Grinnell College acknowledged this without additional review. There was no relevant review board in Ethiopia at the time of the study. Shortly after we began data collection, the COVID-19 pandemic broke out and

we halted research activities when quarantines began across the globe in March 2020. Once activities began resuming, we assessed the potential risks of resuming our activities for research participants and staff as well and reviewed local, national, and NGO-based protocols and regulations. We submitted a modified proposal to the UGA IRB that included additional precautions to accommodate research during the pandemic—physical distancing, mask-wearing, frequent disinfecting of hands, surfaces, and supplies, and outdoor interviewing. Our revised submission (VERSION00000216) was accepted as exempt on November 16, 2020. The research team were all expected to self-monitor for coronavirus symptoms, and they self-isolated immediately if symptoms arose or if they had an exposure (suspected or confirmed).

(need to add parts about civil conflict)

- **Financial and Reputational Conflicts of Interest**
  Do any of the researchers have financial conflicts of interest with regard to the results of the research? Do any of the researchers have potential reputational conflicts of interest?
  The authors do not have any financial or reputational conflicts of interest related to this research.
- **Intellectual Freedom**
  Do any of the researchers have financial conflicts of interest with regard to the results of the research? Do any of the researchers have potential reputational conflicts of interest?
  The researchers did not have any constraints on their intellectual freedom to publish their findings.
- **Feedback to Participants or Communities**
  Is there a plan for providing feedback on research results to participants or communities? If yes, what is the plan? If not, why not?
  The authors have given presentations of their results to program staff at CARE Ethiopia and World Vision Ethiopia. There is no plan to provide feedback to respondents because the intervention was purely methodological, meant to inform stakeholders and researchers.
- **Foreseeable Misuse of Research Results**
  Is there a foreseeable and plausible risk that the results of the research will be misused and/or deliberately misinterpreted by interested parties to the detriment of other interested parties? If yes, please explain any efforts to mitigate such risk.
  We do not foresee any misuse of our results.
- **Other Ethics Issues to Discuss**
  Are there are any other issues to discuss?
  No.